LSAC RESEARCH REPORT SERIES

Robustness of Person-Fit Decisions in Computerized Adaptive Testing

Rob R. Meijer
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 04-06
November 2005

A Publication of the Law School Admission Council

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

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

Table of Contents

Executive Summary

Several statistics have been recommended by researchers for identifying test taker responses to test items that are different from what would be expected, given what is known about the characteristics of the items and the estimated ability level of the test taker. Several of these statistics, often called person-fit statistics, are used for evaluating test taker responses to an entire string of items simultaneously. These statistics allow us to conclude that a particular item response theory (IRT) model either does or does not fit a person's set of responses to items. (Note that IRT is a mathematical model used to analyze test data.) In this sense, these statistics are for use in a global method that only allows us to identify misfitting responses; that is, they do not help us to identify the type of behavior that caused the misfit.

Fortunately, we also have statistics, termed local statistics, that allow us to diagnose the misfit. Such methods may allow us to evaluate if the misfit was caused by violations of one of the assumptions made in applying IRT to the analysis of test data. One such assumption is that of unidimensionality, which requires that a test measure only one ability. This paper focuses on person-fit statistics developed for checking the unidimensionality assumption. Because test data may not be unidimensional, it is worth investigating the effect of unidimensionality violations on the ability of person-fit statistics to identify violations of this assumption.

We applied both global and local person-fit statistics to multidimensional test data from adaptive testing. As may have been anticipated, the results show that some statistics are more robust to unidimensionality violations than others. The context in which certain methods are more useful than others is indicated.

Abstract

Person-fit statistics test whether or not the likelihood of a respondent's complete vector of item scores on a test is low given the hypothesized item response theory (IRT) model. This binary information may be insufficient for diagnosing the cause of a misfitting item-score vector. This paper applies different types of person-fit analysis in a computer adaptive testing context and investigates the robustness of several methods to multidimensional test data. Both global person-fit statistics to make the binary decision about fit or misfit of a person's item-score vector and local checks are applied. Results showed that there are differences between the methods with respect to the robustness in a multidimensional context and that some methods are more useful than other methods.

Introduction

Person-fit researchers have suggested several statistics for identifying misfitting vectors of item scores on the J items from the test; see for a comprehensive review (Meijer & Sijtsma, 2001). These person-fit statistics all assume a particular item response theory (IRT) model (e.g., the three-parameter logistic model) to fit the test data.

By evaluating the whole vector of J item scores simultaneously, person-fit statistics allow the conclusion that a particular IRT model either does or does not fit a respondent's item-score vector. In this sense, most person-fit methods are global methods that identify misfit but do not help to identify the type of behavior that caused the misfit. An exception is due to Klauer (1991; also, see Meijer, 2003), who proposed a method that identifies person misfit caused by either violations of unidimensional measurement, item discrimination, or local independence under the model. Furthermore, person-fit statistics are developed for unidimensional test data. Because many test data may not be perfectly unidimensional, and sometimes are multidimensional, it is interesting to investigate what the effect of violations of unidimensionality is on the power of person-fit statistics. In this study, several person-fit statistics are applied in a computerized adaptive testing (CAT) environment that are sensitive to different types of aberrant response behavior and investigate the robustness of these statistics to multidimensionality.

Another concern in person-fit analysis is that for each respondent an item-score vector of only J observations is available. The number J typically ranges from, say, 10 to 60. This small sample size makes person-fit hazardous from a statistical point of view. In particular, low power may render misfitting item-score vectors difficult to detect, resulting in detection rates that are too low. Due to limited testing time for each ability to be tested, the lengthening of tests to well over, say, a hundred items, is not a realistic option.

An alternative to both the limited value of a binary outcome (that provides little information for individual diagnosis) and the small sample size (that provides little power, implying modest detection rates) may be to seek various other sources of information about an item-score vector's misfit. The combination of these sources may lead to a more accurate decision about misfit or fit, and also provide insight into the cause of an item-score vector's misfit. This study discusses different person-fit analysis that uses various sources of person-fit information as discussed in Emons, Sijtsma, and Meijer (2005). Global person-fit statistic $U3$ is applied (Meijer & Sijtsma, 2001). This is a method that uses kernel smoothing to estimate the person

response function (PRF) and a local person-fit statistic that evaluates unexpected trends in the PRE. Furthermore, two methods especially developed in a CAT context are applied. A simulated example shows how these different methods can be used in practical person-fit analysis.

Different Types of Person-Fit Analysis

The technical details of the methods used at each stage are discussed below. First, a global analysis person-fit statistic $U3$ and a cumulative sum (CUSUM) procedure (Page, 1954) are used to identify fitting and misfitting item-score vectors. Second, a graphical analysis is conducted. Kernel smoothing is used to estimate the person response functions (PRFs) for the misfitting item-score vectors. The PRF gives the probability of a correct response (scored 1) as a function of the difficulty of the items. This function is nonincreasing when the J item response functions (IRFs) in a test do not intersect. For each misfitting item-score vector, the graph of the PRF is inspected for local increases. Third, a local analysis investigates the deviations from the monotone nonincreasing trend in the PRFs using a statistical test proposed by Rosa, Swygert, Nelson, and Thissen (2001) and improved by Meijer (2002).

The combination of global testing, graphical inspection of the PRF for misfitting item score vectors, and local testing of increases found in the PRF together help to better diagnose the misfit.

Item Response Theory

The statistics used in this study are defined in the context of item response theory (IRT). In IRT, the probability of obtaining a correct answer on item i is a function of the latent trait θ and characteristics of the item. This conditional probability $P_i(\theta)$ is the IRF. Item characteristics that are often taken into account are the item discrimination (a), the item location (b), and the pseudo-chance level parameter (c). In parametric IRT, $P_i(\theta)$ often is specified using the 1-, 2-, or 3-parameter logistic model (1-, 2-, 3PLM). The 3PLM (Lord & Novick, 1968, chaps. 17–20) is defined as

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

The 2PLM can be obtained by setting $c_i = 0$ for all items; and the Rasch (1960) model can be obtained by additionally setting $a_i = 1$ for all items.

Some statistics used in this paper are proposed in the context of nonparametric IRT (NIRT). NIRT models assume order restrictions on the IRFs. Let $X_i (i = 1, ..., I)$ denote the binary random variable for the item responses, with realization $x_i = 1$ for a correct or coded response, and $x_i = 0$ otherwise. Let $X_+ = \sum_{i=1}^{I} X_i$ denote the unweighted sum score; let $\pi_i (i = 1, ..., I)$ denote the population proportion of persons with a 1 score on item i; and let $\hat{\pi}_i = N_i/N$ (N is the sample size and N_i the frequency of 1s on item i) be the sample estimate of π_i. We assume that the I items in the test are ordered and numbered from easy to difficult: $\pi_1 \geq \pi_2 \geq ... \geq \pi_I$. The probability of obtaining a 1 score is related to the latent trait θ by the IRF: $P_i(\theta) = P(X_i = 1 \mid \theta)$.

We assume a scalar θ (unidimensionality assumption; UD). Given UD we assume that item scores are locally independent (assumption LI). A typical NIRT assumption is that the IRFs are monotone nondecreasing in the latent trait (assumption M); that is, for two arbitrary fixed values θ_a and θ_b,

$$P_i(\theta_a) \leq P_i(\theta_b) \text{ for } \theta_a < \theta_b; i = 1, ..., I. \quad (2)$$

NIRT models that satisfy the assumptions of UD, LI, and M imply that the total score X_+ stochastically orders θ. Stochastic ordering justifies the use of X_+ for ordering persons on θ and is a useful ordering property in practice whenever a test is used to order respondents. Mokken's (1997) monotone homogeneity model is defined by the assumptions of UD, LI, and M.

For person-fit analysis it is convenient that the IRFs do not intersect because then the same ordering of items by difficulty applies to each respondent and this facilitates the interpretation of test performance. Nonintersection for two items j and i means that if we know for a fixed value θ_0 that $P_j(\theta_0) > P_i(\theta_0)$, then

$$P_j(\theta) \geq P_i(\theta). \quad (3)$$

This is the assumption of invariant item ordering (IIO). Mokken's model of double monotonicity is defined by the assumptions of UD, LI, M, and IIO. Several methods exist to investigate if the double monotonicity model fits a set of items.

The Person Response Function

The PRF for respondent v is defined as the probability of a correct answer to items measuring θ as a function of their item difficulty. This is formalized by a random variable S_v that takes value 1 for items that were answered correctly by respondent v and 0 otherwise. Let $G(\theta)$ be the cumulative θ distribution. Item difficulty is defined as

$$1 - \pi_i = \int_\theta [1 - P_i(\theta)] dG(\theta),\ i = 1, ..., I, \tag{4}$$

and sample estimates $(1 - \hat{\pi}_i)$ can be used to estimate the ordering of the items. It may be noted that under an IIO, theoretically the item difficulties, $1 - \pi_i\ (i = 1, ..., I)$, have the same ordering as the response probabilities, $P_i(\theta)$, $i = 1, ..., I$. The probability for respondent v to give correct answers as a function of item difficulty, $1 - \pi$, can be written as

$$P_v(1 - \pi) = P(S = 1 \mid 1 - \pi, \theta_v). \tag{5}$$

This conditional probability is defined on the continuous scale $(1 - \pi)$ with domain [0,1]. The PRF, $P_v(1 - \pi)$, is nonincreasing under NIRT models that have an IIO. However, Meijer & Sijtsma (2001) showed that person-fit decisions were quite robust against violations of IIO. Kernel smoothing was used to obtain a (quasi-) continuous estimate of the PRF. This estimate was convenient for the localization and the interpretation of misfit.

Kernel Smoothed Estimates of the PRF

Kernel smoothing is a nonparametric regression technique (Ramsay, 1991, 2000). It takes a focal observation, here an item difficulty, for example, $1 - \pi_{i(0)}$, and several of its neighbor item difficulties, and then estimates $P_v(1 - \pi_{i(0)})$ as the weighted mean of the item scores $x_{vi(0)}$ and the x_{vi}'s of the neighbor items. Weights are assigned by the kernel function, $K(.)$. A subset of observations that is used for estimating one function value is called a window. Each observation $1 - \pi_i\ (i = 1, ..., I)$ is focal point once, and windows move from left to right. Windows for items at or near the endpoints of the item ordering contain less data. Special precautions take care of the resulting inaccuracy in estimation.

The bandwidth determines the number of observations used in the estimation of the function values. A broader bandwidth means that adjacent estimated function values are more alike because the windows used for estimation are almost identical. Thus, the PRF is estimated relatively accurately (little variance) but interesting details may get lost (much bias). A narrower bandwidth has the opposite effect: Function values are different because subsequent windows contain few observations as they quickly enter and exit the window as it moves along. Particular jags in the PRF are visible (little bias) at the expense of statistical accuracy (much variance). Thus, for a particular application the choice of the bandwidth involves finding the balance between bias and inaccuracy. This will be explained in more detail shortly.

Let $z_i = [(1 - \pi_i) - (1 - \pi_{i(0)})]/h = (\pi_{i(0)} - \pi_i)/h$, where h is the bandwidth to be defined shortly, and let $K(z_j)$ be the kernel function. The nonparametric regression function we use is defined as

$$\hat{P}_v(1 - \pi_{i(0)}) = \frac{\sum_{i=1}^{I} K(z_i) x_{vi}}{\sum_{i=1}^{I} K(z_i)}. \tag{6}$$

For the kernel function we use the standard normal density,

$$K(z_i) = \frac{1}{\sqrt{2\pi}} \exp^{-z_i^2/2}, \tag{7}$$

which is a common choice. Using the standard normal kernel function, each window in fact uses all J observations, but observations further away from the focal observation receive small weights and truncation eliminates the influence of distant observations.

The Person Response Function and Local Person Fit

Discrete PRF Estimate

For local person-fit testing, a discrete estimate of the PRF was used as in Emons, Sijtsma, and Meijer (2005). Notice in the following equation that this discrete estimate may be seen as an extreme version of kernel smoothing, with uniform kernels that do not overlap. First, the J items are ordered by increasing $(1 - \pi)$ values. Then, they are divided into K ordered disjoint subsets, denoted A_k with $k = 1, ..., K$. For simplicity's sake (but not by necessity), each subset contains m items, such that $A_1 = \{X_1, ..., X_m\}$, $A_2 = \{X_{m+1}, ..., X_{2m}\}$, ..., $A_K = \{X_{I-m+1}, ..., X_I\}$. For respondent v, the expected proportion of correct answers to the items in A_k equals $\tau_{vk} = m^{-1}\sum_{j \in A_k} P_i(\theta_v)$. Given an IIO, an ordering of the items according to the $(1 - \pi_i)$s implies that for each respondent v,

$$m^{-1} \sum_{i \in A_k} P_i(\theta_v) \geq m^{-1} \sum_{i \in A_k + 1} P_i(\theta_v). \tag{8}$$

For the K item subsets it follows that

$$\tau_{v1} \geq \tau_{v2} \geq \ldots \geq \tau_{vK}. \tag{9}$$

Let X_{vj} denote the score of person v on item j. The ordering in Equation 9 is estimated using sample fractions

$$\hat{\tau}_{vk} = m^{-1} \sum_{i \in A_k} X_{vi}. \tag{10}$$

Furthermore, we use a person-fit statistic that quantifies the result that in any item subset the correct answers are most likely to be given to the relatively easy items. Define any item vector $\mathbf{Y}$ (e.g., combine subsets A_k and A_{k+1} into one set) in which items are ordered by ascending difficulty. Then, count the number of item pairs in $\mathbf{Y}$ in which the easiest item is answered incorrectly while the more difficult item is answered correctly. This is the number of Guttman errors. For respondent v the number of (0, 1) patterns on all possible item pairs (including pairs that contain the same item twice) equals

$$G_v = \sum_{j=1}^{I_Y} \sum_{i=1}^{I} (1 - Y_{vj}) Y_{vi}. \tag{11}$$

Next, we show that the function $f(\mathbf{Y}) = G$ is increasing in transposition (IT), to be explained shortly. The IT property of $f(\mathbf{Y})$ is needed for deriving an approximate Type I error probability for the number of Guttman errors given that the items in $\mathbf{Y}$ have an IIO. In general, a function $f(\mathbf{Y})$ is IT if interchanging a 0 and a 1 score in a realization $\mathbf{y}$, such that the 1 score is positioned further to the right, has the effect of increasing $f(\mathbf{Y})$. For example, for $\mathbf{y}_1 = (110010)$, function $f(\mathbf{y}_1) = 2$. Interchanging the second 0 and the second 1 yields $\mathbf{y}_2 = (100110)$ and $f(\mathbf{y}_2) = 4$.

Person-misfit in $\mathbf{Y}$ is revealed by an exceptionally high G value given the expected G value under the postulated NIRT model. For sum score $Y_+ = \sum Y_i$ and realization y_+, and the number of items I_Y, we evaluate the probability $P(G \geq g \mid y_+, I_Y)$ using a theorem proven by Rosenbaum (1987). This theorem compares the expectation of an IT function like $f(\mathbf{Y}) = G$ given that the IRFs have an IIO, with the expectation of $f(\mathbf{Y}) = G$ given that $\mathbf{Y}$ follows the exchangeable distribution; that is, given that all possible item score vectors $\mathbf{Y}$ containing y_+ 1s have equal probability. Under an NIRT model, $\mathbf{Y}$ follows the exchangeable distribution if and only if the response probabilities, $P_i(\theta)(i = 1, ..., I)$, are equal for all items. This means that the IRFs are flat and coincide completely. The theorem says, essentially, that given that the IRFs have an IIO (Equation 1), the number of Guttman errors cannot exceed the corresponding number expected under the exchangeable distribution. Because under an NIRT model we cannot evaluate $P(G \geq g \mid y_+, I_Y)$ directly, we compare it to the corresponding probability under the exchangeable distribution. The latter is at least as great as the former, and thus provides an upper bound.

How is statistic G distributed under the exchangeable distribution? Molenaar and Hoijtink (1990) showed that G is a linear function of the sum of ranks. Thus, under the exchangeable distribution, $P(G \geq g \mid y_+, I_Y)$ can be obtained from the Wilcoxon rank-sum distribution. This probability provides an upperbound for $P(G \geq g \mid y_+, I_Y)$ under IIO. For item subsets containing fewer than 20 items, tables may be used to obtain probabilities of exceedance. For item subsets containing at least 20 items, G is approximately

normally distributed. Meijer and Sijtsma (2001) concluded from a simulation study that for many tests the Type I error rate of G often ranged from 0.02 to 0.03 (nominal alpha = 0.05), with slightly better results for higher θs. This was found for both item sets with and without an IIO.

Methods Especially Developed for CAT

As an alternative to both methods discussed above, Bradlow, Weiss, and Cho (1998) and van Krimpen-Stoop and Meijer (2001) proposed person-fit statistics based on the cumulative sum (CUSUM) procedure (Page, 1954). Note that in CAT a model-fitting item-score pattern consists of an alternation of correct and incorrect responses, especially at the end of the test when $\hat{\theta}$ converges on θ. A string of consecutive correct or incorrect answers could indicate misfit or a bad bank. Sums of consecutive negative or positive residuals $[x_i - p_i(\theta)]$ can be investigated using a CUSUM. For each item i in the test, a statistic T_i can be calculated that equals (a weighted version of) $[x_i - p_i(\theta)]$. A simple statistic is

$$T = 1/k[x_i - p_i(\theta)]. \tag{12}$$

Then, the sum of these T_is is accumulated as follows

$$C_i^+ = \max[0, T_i + C_{i-1}^+], \tag{13}$$

$$C_i^- = \min[0, T_i + C_{i-1}^-], \text{ and} \tag{14}$$

$$C_0^+ = C_0^- = 0, \tag{15}$$

where C^+ and C^- reflect the sum of consecutive positive and negative residuals, respectively. Let UB and LB be some appropriate upper and lower bounds. 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.

Another approach was discussed by Rosa et al. (2001) and strongly related to the PRF approach. For the sake of simplicity we change our notation a little bit. Let, again, the score on item i be denoted by X_i, let the item score vector be denoted by $\mathbf{x}$, and the sum score for a set of items be denoted by x. The likelihood for any summed score is

$$L_x(\theta) = \sum_{(U_i)=x} L(x \mid \theta), \tag{16}$$

where the summation is over all response patterns that contain x correct responses. That is, given θ the likelihood of a summed score is obtained as the sum of the likelihoods of all response patterns that have that summed score. The probability of each score x is then

$$P_x = \int L_x(\theta)\phi(\theta)d\theta \tag{17}$$

where $\phi(\theta)$ is the population density. An algorithm to compute $L_x(\theta)$ was proposed by Lord and Wingersky (1984) and discussed in Thissen, Pommerich, Billeaud, and Williams (1995). This algorithm assumes that the individual $P_i(\theta)$ are estimated under a specified IRT model.

To investigate unexpected sum scores on subtests of items, a generalization of Equation 5 can be used. Assume that there are two subtests x and x'. The likelihood of a combination of sum scores can be calculated by

$$L_{xx'}(\theta) = L_x(\theta)L_{x'}(\theta) \tag{18}$$

and the probability of the response pattern of the summed scores $\{x, x'\}$ equals

$$P_{xx'} = \int L_{xx'}(\theta)\phi(\theta)d\theta. \tag{19}$$

If a score combination is very unlikely, values of $P_{xx'}$ can be calculated for each score combination x and x' and plotted in a diagram (Rosa et al., 2001) and $P_{xx'}$ can then be used to construct a $(1 - \alpha)$ 100% "highest density region" (HDR) for the response combinations. In Table 1 an example of such a diagram is given. Before discussing this diagram, it is important to note that the values of $P_{xx'}$ cannot be interpreted as reflecting likely or unlikely events in any absolute sense because the magnitude of the individual $P_{xx'}$ depends on the number of row and column score points. To construct the HDR, first the cells should be ordered from largest to smallest $P_{xx'}$. The 95% HDR can then be determined by considering all cells that contribute to the first 95% of the cumulative total of $P_{xx'}$. According to the model, 95% of the examinees should obtain score combinations in that list of cells. Cells that are outside this region represent score combinations that are thus unlikely given the model. No particular parametric form for the item response function is assumed in the formulation of the recursive algorithm that is used to calculate Equation 5. All that is needed is a probability under an item response theory model.

TABLE 1
An HDR table for a division of a test into two subtests of the four easiest items and the five most difficult items and 99% HDR. The unshaded area in the table contains the 99% HDR; the score combinations in the shaded area are thus unlikely at the 1% level

Sum score Subtest 2/1	0	1	2	3	4
0	.013138	.041424	.068642	.068949	.035001
1	.005951	.029073	.072723	.108833	.081653
2	.001485	.010954	.040824	.090294	.100122
3	.000248	.002727	.015018	.049094	.081105
4	.000027	.000445	.003621	.017634	.043975
5	.000002	.000038	.000463	.003400	.013138

A Simulation Study

For each CAT we used a similar simulation study as in Robin (2002). One thousand examinees at low, medium, or high ability level were generated. Normal scores were generated according to the normal response model using the three-parameter logistic item parameters. Spuriously low aberrant responses were simulated with a probability of correct answer choice of 0.20. This mimics response behavior on five-choice items that may occur when an examinee has trouble concentrating on the task at the beginning of the test, is unmotivated, or runs out of time. In the case of CATs, tests tend to be short with item difficulty ranging from medium difficulty and around examinee's true ability level. We are interested in understanding and evaluating what may happen in typical or idealized conditions without content and security constraints before trying to address a specific one in follow-up studies. The tests contained 40 items with *a*- parameters set at .90, *b*-parameters equally spaced from –2.0 to 0.0, –1.0 to 1.0, and 0.0 to 2.0 for examinees at –1.5, 0.0, and 1.5 ability levels, respectively, and *c*-parameters set to .15. Now taking into account the general effect of aberrant responses on tests delivered to high and medium ability examinees, examples of aberrant CATs were simulated by shifting the item difficulties of the normal CATs down by 1.0 in the early aberrant case and by 0.5 in the medium and late aberrant case.

For d traits the number of correlations between pairs is $d(d - 1)/2$; thus many relationships could be studied. Here the simplest case was investigated. For $d = 2$ case, the joint distribution of the latent traits θ_1 and θ_2 was standard normal. Four correlations were investigated: $\rho = .5, .6, .7$, and .8. These values represent values often found in multidimensional datasets with a hierarchical structure where a common factor is measured by means of different subtests. Forty items were used. Items 1 through 20 measured θ_1 and items 21 through 40 measured θ_2. For all statistics a significance level of 5% was used.

Measures

Because the scales had fairly high discrimination, global person-fit was analyzed using *U*3 as a descriptive statistic. Extreme *U*3 values appeared in the right tail of the distribution like in the Emons, Meijer, and Sijtsma (2005) study. We classified the score vectors into three X_+ levels, denoted low, medium, and high. For each selected item-score vector, kernel-smoothing was used to estimate a (quasi-) continuous PRF. We used a bandwidth $h = 0.09$. Most misfitting PRFs increased at relatively easy items, but not at the easiest items. The PRFs at high X_+ levels typically showed a brief increase at medium to high item difficulty. The PRFs at high X_+ levels rarely showed misfit at the easiest items. At high X_+ levels, incorrect answers were rare but sometimes scattered throughout the test. Also, note that when an easy item was failed but several more difficult items were succeeded, this failure received much weight and produced a high *U*3, but did not affect the shape of the PRF. Local increases of the PRFs were tested for significance using the number of Guttman errors, *G*. We divided the items into $K = 2$ disjoint subsets, each containing $m = 20$ items.

Results

Table 2 gives the results of the person-fit tests. To illustrate this consider an item score vector **Y** with J_Y = 20 and Y_+ = 13. We considered these subsets as one vector, **Y**, and counted the number of Guttman errors, *G*. The upper bound for the significance probability was obtained from the Wilcoxon rank-sum distribution. For Case 1, *G* = 75, which was significant at the .01 level. The value of *G* may be compared to the maximum number of Guttman errors (G_{max}), given the number of items (J_Y) and the number of correct answers (Y_+). For an item-score vector **Y** with J_Y = 20 and Y_+ = 13, we have $G_{max} = Y_+(J - Y_+) = 13 \times (20 - 13) = 91$. The normed number of Guttman errors (denoted by G^*) is defined as $G^* = \frac{G}{G_{max}}$, which in this example equals .82. A practical rule of thumb for interpreting G^* for a small number of items (e.g., $10 \leq J \leq 20$) is that values of 0.82 and higher indicate serious misfit.

TABLE 2
Detection rate of several person-fit statistics

				Multidimensional											
	Unidimensional			ρ = .5			ρ = .6			ρ = .7			ρ = .8		
Statistic	θ = –1.5	θ = 0	θ = 1.5	θ = –1.5	θ = 0	θ = 1.5	θ = –1.5	θ = 0	θ = 1.5	θ = –1.5	θ = 0	θ = 1.5	θ = –1.5	θ = 0	θ = 1.5
*U*3	.34	.39	.41	.25	.28	.32	.28	.28	.34	.28	.32	.38	.33	.41	.41
PRF	.36	.41	.42	.27	.27	.31	.29	.29	.33	.30	.31	.31	.35	.39	.40
G_v	.32	.31	.38	.24	.26	.35	.26	.27	.35	.28	.28	.32	.27	.29	.37
$P_{xx'}$	.42	.45	.49	.40	.39	.41	.40	.41	.42	.38	.40	.44	.42	.43	.50
CUSUM	.45	.49	.49	.32	.35	.31	.34	.34	.31	.36	.42	.51	.45	.47	.51

Comparing the detection rates of the different statistics it can be concluded that the power of $P_{xx'}$ is somewhat higher than the power of *U*3, *G*, and the CUSUM procedure, across most conditions. Furthermore, the influence of multidimensionality seems negligible for ρ = .8. For ρ = .8 through ρ = .5 there is descending trend for the *U*3 and the CUSUM statistics in the sense that the power decreases, probably due to multidimensionality. However, for $P_{xx'}$ the power is the same across different ρ levels. This is probably due to the fact that this method capitalizes on the different types of subsets. On the item level the lower correlation between the subsets results in more unexpected item scores.

Discussion

The mismatch between the examinee's performance on one subset of items and another may mean that the examinee cheated or that something else went wrong. The context of high-stakes testing, which is expensive in time and money for the examinee may add further considerations. Several possible courses of action include online extension of the test, either switching from a CAT system to a linear form. Other possible actions include cancellation and retesting.

In this study we showed that $P_{xx'}$ had higher power than the other statistics and that multidimensionality may have an effect on the power of a person-fit statistic. In future research, person-fit statistics that are sensitive to multidimensionality are important tools for further exploring the validity of item-score patterns.

References

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93* (443), 910–919.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person response functions. *Psychological Methods, 10* (1), 101–119.

Klauer, K. C. (1991). An exact and optimal standardized person fit test for assessing consistency with the Rasch model. Psychometrika, 56, 213–228.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 453–461.

Meijer, R. R. (2002). Outlier detection in high stakes certification testing. *Journal of Educational Measurement (39)*, 219–233.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using IRT based person-fit statistics. *Psychological Methods, 8* (1), 72–87.

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

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York: Springer Verlag.

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

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

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.

Ramsay, J. O. (2000). *TestGraf: A program for the graphical analysis of multiple-choice tests and questionnaire data* [Computer program and unpublished manuscript]. Montreal, Canada: McGill University.

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

Robin, F. (2002, April). *Investigating the relationships between test response behavior, measurement and person-fit*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response item-scale scores for patterns of summed scores (pp. 253–292). In D. Thissen & H. Wainer (Eds), *Test scoring*. Majwah, NJ: Lawrence Erlbaum.

Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika, 52*, 217–233.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39–49.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). Cusum-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199-217.