

Influence of Test and Person Characteristics on Nonparametric Appropriateness Measurement

Rob R. Meijer, University of Twente

Ivo W. Molenaar, University of Groningen

Klaas Sijtsma, Utrecht University

Appropriateness measurement in nonparametric item response theory modeling is affected by the reliability of the items, the test length, the type of aberrant response behavior, and the percentage of aberrant persons in the group. The percentage of simulees defined a priori as aberrant responders that were detected increased when the mean item reliability, the test length, and the ratio of aberrant to nonaberrant simulees in the group increased. Also, simulees "cheating" on the most difficult items in a test were more easily detected than those "guessing" on all items. Results were less stable across replications as item reliability or test length decreased. Results suggest that relatively short tests of at least 17 items can be used for person-fit analysis if the items are sufficiently reliable. *Index terms: aberrance detection, appropriateness measurement, nonparametric item response theory, person-fit, person-fit statistic U3.*

Person-fit or appropriateness measurement is concerned with identifying persons whose item score patterns on a test are unusual (aberrant) given what is expected based on an item response theory (IRT) model or the score patterns of the other (nonaberrant) persons in the group. For people who respond aberrantly to a test, it is questionable whether the test score is an appropriate measure of the trait that is being measured. Hence, additional information is required to reach a sound conclusion about such persons.

Aberrant patterns may, for example, provide information about cheating and guessing on examinations (Levine & Rubin, 1979), membership in a subgroup that was initially not identified as relevant for the investigation (e.g., a subgroup suffering from

a language deficiency; van der Flier, 1982), scoring and other clerical errors (Hulin, Drasgow, & Parsons, 1983), and deficiency in some of the traits required to solve the items from a subdomain in ability and achievement tests (Tatsuoka, 1985).

Many methods for the detection of aberrant response patterns have been proposed (e.g., Drasgow, Levine, Williams, McLaughlin, & Candell, 1989; Klauer & Rettig, 1990; Levine & Rubin, 1979; Molenaar & Hoijtink, 1990; Trabin & Weiss, 1983; van der Flier, 1982; Wright & Stone, 1979). All methods are particularly sensitive to item score patterns that have correct or keyed responses (scored as 1s) for relatively difficult items and incorrect or not keyed responses (0s) for relatively easy items. This generally accepted approach to aberrance is in accordance with Guttman's (1950) scalogram model that excludes all item score patterns in which at least one item pair has a 0 score for the easier item and a 1 for the more difficult item in the pair.

The present study extended the work of Reise & Due (1991). They investigated the influence of three test characteristics on decisions about the fit of item score patterns. Using simulated data, they studied the influence of test length, the spread of the item difficulties, and the degree of aberrance on the power of the standardized log-likelihood statistic (I_2) (Drasgow, Levine, & Williams, 1985) for person-fit evaluation in the context of the three-parameter logistic model (3PLM; e.g., Lord, 1980, p. 12). Model-fitting response vectors (FRV) were generated according to the 3PLM. Nonfitting response vectors (NRV) were generated under the 3PLM using items that had a weaker discrimination than

APPLIED PSYCHOLOGICAL MEASUREMENT

Vol. 18, No. 2, June 1994, pp. 111-120

© Copyright 1994 Applied Psychological Measurement Inc.

0146-6216/94/0146-6216/94/020097-14\$1.95

the items used for generating the FRVs. Using I_2 , Reise & Due (1991) concluded that test length, the spread of the item difficulties, and the degree of nonfit (defined by the discrepancy between the discrimination parameters for the two groups) influenced the detection rate for NRVs. Holding constant all other factors, an increase in either the test length or the spread of the item difficulties yielded, in general, an increase in the power of I_2 to detect NRVs. Furthermore, as the difference between the discrimination parameters of the items used to generate item scores for the NRVs and the discrimination parameters of the items used to generate item scores for the FRVs increased, the power of I_2 to detect NRVs increased.

There are at least four differences between the present study and the Reise & Due (1991) study:

1. The present study used nonparametric IRT models, whereas the Reise and Due study used the parametric 3PLM. Consequently, different characteristics that influence the power to detect nonfit were studied here.
2. In the present study, not only the influence of test characteristics but also the influence of person, item, and group characteristics on the power of a person-fit statistic were studied.
3. The present study defined a response vector as nonfitting if it had low probability conditional on the total score given that a person belonged to the population of interest, whereas in the Reise and Due study a pattern was nonfitting if it provided less information for estimating the latent trait (Lord, 1980, p. 12) than predicted by a specific IRT model.
4. In addition to the power of a person-fit statistic to detect NRVs, the extent to which person-fit results can be replicated was investigated here.

Most person-fit research using simulated data has been concerned with the percentage of a priori defined NRVs that are detected by a particular statistic (e.g., Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979; van der Flier, 1980). Because such detection rates may vary across different samples, the question arises as to how stable the obtained person-fit results are across different samples. In this study, both the rate of detection and the stability of

person-fit statistics were examined.

The objective of this study was to systematically investigate the power of the nonparametric person-fit statistic, U_3 , as a function of item characteristics, test characteristics, person characteristics, and the group to which examinees belonged. The parametric person-fit literature (e.g., Moleenaar & Hoijtink, 1990; Reise & Due, 1991) has shown that test length and the spread of item difficulties influence the rate of detection of nonfit. Furthermore, the type of NRV (e.g., resulting from cheating, guessing, or poor motivation) may influence the power of a person-fit statistic (e.g., Drasgow et al., 1985; Kogut, 1987). Because most person-fit studies have not systematically varied item, test, individual, and group characteristics that influence person-fit statistics in one design, the exact manner in which these characteristics influence the power of a person-fit statistic is unclear.

Nonparametric Person-Fit Research

This study used Mokken's (1971) model of double monotonicity, a nonparametric IRT approach (Meijer, Sijtsma, & Smid, 1990; Mokken & Lewis, 1982). This approach assumes unidimensionality, local stochastic independence, and monotonically non-decreasing item response functions (IRFs). The latter assumption allows an item-independent ordering of respondents on the unidimensional IRT scale by means of the estimated true score from classical test theory. In addition, and most important in the present context, the IRFs do not intersect. They may touch locally, however, or even coincide. Except for ties, this property of the IRFs allows an invariant or group-independent ordering of items according to their difficulty. The usefulness of this model lies in its potential to order persons and items; these orderings are item-independent and group-independent, respectively, given a well-defined domain of items and a population of persons for which the model of double monotonicity holds. Similar models also have been discussed by Rosenbaum (1987) and Grayson (1988).

Several nonparametric IRT and non-IRT person-fit statistics have been proposed (see Harnisch & Linn,

1981, for a comparative study of six such statistics.) This study was concerned with the power of the U3 statistic (van der Flier, 1980, 1982) to detect NRVs under varying conditions. This statistic has proven to be useful under varying test conditions in simulation and empirical research (van der Flier, 1980, 1982).

In nonparametric IRT, the IRFs are not parametrically defined (Mokken, 1971, pp. 115–117; Rosenbaum, 1987) and, consequently, the discriminating power of an item cannot be estimated numerically. As a result, the concept of item discrimination is not useful in practical applications of nonparametric IRT. The reliability of an item, however, can be substituted for its discrimination (Meijer, Sijtsma, & Molenaar, 1993) because both item characteristics express the degree to which observed item scores can be repeated independently under similar conditions. In general, keeping all other person and test characteristics fixed, an increase in the item discrimination, and thus in the item reliability, corresponds to a higher degree of repeatability of observed scores on an individual item. As discrimination tends to infinity and item reliability to unity, response performance tends to satisfy the deterministic Guttman (1950) model. In this study, item scores were simulated using the two-parameter logistic model (2PLM; Hambleton & Swaminathan, 1985, p. 36); the item discrimination parameter was varied across cells of the design. Item reliability was used to interpret the simulation results.

The U3 Statistic

Van der Flier (1980, 1982) developed the person-fit statistic U3 in the context of the nonparametric model of double monotonicity. Let P denote a probability, and let X_g denote the binary (0,1) score on item g . $P(\mathbf{X})$ denotes the probability of a specific item score pattern $\mathbf{X} = (X_1, \dots, X_k)$ as estimated from the marginal distribution, and π_g denotes the proportion correct score on item g ($g = 1, \dots, k$) in the population of interest. Let r denote the realization of the total score ($X = r$).

Suppose that k items are ordered such that $\pi_1 \geq \pi_2 \geq \dots \geq \pi_k$. An item score vector with 1s in the first r positions and 0s in the last $k - r$ posi-

tions is called a Guttman vector because it perfectly satisfies the requirements of the Guttman (1950) scalogram model. Such a vector is denoted by $\mathbf{X}^* = (X_1^*, \dots, X_k^*)$. The vector with 0s in the first $k - r$ positions and 1s in the last r positions is called a reversed Guttman vector because given that $X = r$, it is the score pattern with the maximum number of Guttman errors. A reversed Guttman vector is denoted by $\mathbf{X}' = (X_1', \dots, X_k')$. Using this vector notation, U3 is defined as:

$$U3 = \frac{\ln P(\mathbf{X}^*) - \ln P(\mathbf{X})}{\ln P(\mathbf{X}^*) - \ln P(\mathbf{X}')} \tag{1}$$

U3 ranges from 0 to 1, where 0 indicates that the observed response pattern is an ideal Guttman pattern, and 1 indicates that the observed pattern is a reversed Guttman pattern. Increasing values indicate that patterns deviate further from perfect Guttman patterns.

Van der Flier (1982) investigated characteristics of U3 in a simulation study. He concluded that for sets of 17 and 29 items with uniformly or normally distributed π_g values, the U3 distributions within different score groups could be combined into one common distribution. Consequently, U3 values can be compared across different score groups.

Item Reliability

Mokken (1971, p. 143) defined the reliability of item g as

$$\rho_g = \frac{\pi_{gg} - \pi_g^2}{\pi_g(1 - \pi_g)} \tag{2}$$

where π_{gg} denotes the joint proportion correct on item g in two independent replications. The marginal proportion, π_g , can be estimated from empirical data. Because independent replications of items usually are not available, π_{gg} cannot be estimated directly from the data. To obtain an approximation of π_{gg} , one or two items from the same test as item g can be used that have IRFs that are highly similar to the IRF of item g (Mokken, 1971, p. 146). These items are considered to be approximately equivalent to item g .

Let the latent trait value as defined in IRT (Lord, 1980, p. 12) of person i ($i = 1, \dots, n$) be denoted by

θ_i , and assume that the k items are numbered and ordered such that $\pi_1 \geq \pi_2 \geq \dots \geq \pi_k$. Furthermore, let the IRF of item g be denoted by $\pi_g(\theta)$. Given that the IRFs of all k items are nonintersecting, then for items $g-1$, g , and $g+1$

$$\pi_{g-1}(\theta) \geq \pi_g(\theta) \geq \pi_{g+1}(\theta), \text{ for all } \theta. \quad (3)$$

Based on the assumption that the IRFs of the neighbor items $g-1$ and $g+1$ in the item ordering are more similar to the IRF of item g than the other IRFs in the test, Mokken (1971) used the IRF of item $g-1$ or $g+1$, or both, as a predictor for a real replication of item g . Using only one neighbor item, $g-1$ or $g+1$, π_{gg} is approximated by extrapolation using $\hat{\pi}_{g-1}$ (or $\hat{\pi}_{g+1}$), $\hat{\pi}_g$, and the joint proportion $\hat{\pi}_{g-1,g}$ (or $\hat{\pi}_{g,g+1}$) of persons with correct responses on both items $g-1$ and g (or both g and $g+1$). If both neighbor items are used, π_{gg} is approximated by interpolation using $\hat{\pi}_{g-1}$, $\hat{\pi}_g$, $\hat{\pi}_{g+1}$, $\hat{\pi}_{g-1,g}$, and $\hat{\pi}_{g,g+1}$.

Sijtsma & Molenaar (1987) extended and refined the methods proposed by Mokken in the context of total score reliability. In the context of individual item reliability, Meijer et al. (1993) investigated the statistical properties of the two methods proposed by Mokken and a new method (Sijtsma & Molenaar, 1987) based on Mokken's methods. Assuming nonintersecting IRFs, they found that Sijtsma and Molenaar's method estimated item reliability almost without bias for almost all items in a test. Some bias existed for the extremely easy and difficult items when the item difficulties were widely spaced. The sampling distribution of Sijtsma and Molenaar's estimator was symmetrical with a peakedness comparable to that of the normal distribution.

Method

Design

Data matrices of order n (persons) \times k (items) were generated (for the simulation procedure, see Sijtsma & Molenaar, 1987) using 2PLM IRFs (e.g., Hambleton & Swaminathan, 1985, p. 36) and a standard normal distribution for θ . This procedure was repeated eight times for each cell of a completely crossed design. There were four levels of

uniform discrimination ($a = .5, 1.0, 2.0, \text{ and } 5.0$) for all k items; two levels of test length ($k = 17$ and $k = 33$); two levels of number of NRVs (NNRV = 50 and NNRV = 25); and two types of misfit—"guessing" and "cheating."

The item difficulties (bs) were equidistant between $[-2, 2]$ with a distance equal to .25 for the 17-item test and .125 for the 33-item test. For each test, the median b thus was 0.0. There were 450 simulees in each dataset. Therefore, the two levels of NRV included 11% (NNRV = 50) and 5.5% (NNRV = 25) NRVs, respectively.

Cheaters and Guessers

Cheaters. The group of "cheating" simulees was generated as follows. First, θ values were drawn at random from a standard normal distribution. Negative θ values were used to generate cheating simulees. It was assumed that cheating simulees answered most items on their own (item scores were simulated to fit the IRT model) except for the three most difficult items from the 17-item test and the six most difficult items from the 33-item test. The answers on these most difficult items were changed to simulate copying from more able examinees taking the same test. It was assumed that this cheating always resulted in correct answers; thus, 1s were substituted for these item scores for each cheater. Therefore, depending on the test length, the group of cheating simulees was characterized by correct answers on the three or six most difficult items despite their relatively low θ level.

Note that in the most favorable situation (the situation in which the probability of a correct answer of a cheating simulee was largest, that is if $\theta = 0.0$ and $a = .5$), for the three and six most difficult items, respectively, the probability of correctly answering an item on the basis of the θ was at most .33. For other combinations of parameters this value was always smaller and often would be considerably smaller. Thus, three or six correct responses to the most difficult items clearly did not fit the model.

Guessers. "Guessing" simulees were assumed to answer the items by randomly guessing the correct answer on each of the k items in the test with

a probability of .25. This probability corresponds to the probability of obtaining the correct answer by guessing in a multiple-choice test with four alternatives per item.

Replications

Within a particular cell of the design, the same θ values and the same IRFs were used to independently generate eight data matrices (replications). Sets of θ values and characteristics of IRFs varied across different cells. In a particular cell, the rate of detection of a priori defined NRVs was considered for each of the eight data matrices. In addition, the joint rate of detection of a priori defined NRVs in pairs of datasets was considered for four randomly chosen pairs of datasets per cell. The eight datasets per cell were used to obtain a standard deviation (SD) of the mean percentage of replicable simulated NRVs.

Dependent Variables

Four dependent variables reflected the rate of detection within one dataset. First, within each cell the mean percentage of a priori defined NRVs successfully identified by means of U_3 (valid NRVs, VNRV) was determined across the eight replicated datasets. For one dataset, the percentage of VNRVs was found by ordering all 450 simulees according to increasing U_3 and then by determining the percentage of a priori defined NRVs among the NNRV simulees with the highest U_3 values. Note, in particular, that no cut score was used for U_3 . Rather, in each sample the NNRV simulees with the highest U_3 values were selected and the percentage of true NRVs among them was determined. Using the same procedure, the mean percentage of FRVs that were incorrectly classified as NRVs (false NRVs, FNRVs) and the mean percentage of NRVs that were incorrectly classified as FRV (false FRV, FFRV) were determined across the eight replications. Fourth, the mean percentage of FRVs correctly classified as FRV (valid FRV, VFRV) was determined across the eight replications.

To examine the stability of person-fit results within one cell, the mean value of the following dependent variables was computed across four pairs of replications (each sample belonged to only one

pair; thus all samples were used): (1) the percentage of NRVs that were detected in two replications; (2) the percentage of NRVs that were detected in either one of the replications; (3) the percentage of NRVs detected in neither replication; and (4) the level of agreement between U_3 values in the two replications.

Calibration Samples

To determine the U_3 values and the item reliabilities, π_g values were estimated in separate calibration samples. A calibration sample consisting of 450 FRVs was generated using a new sample of 450 θ s from the standard normal distribution and the same IRFs used for the other data matrices in a particular cell. Calibration samples were used for two reasons. First, the presence of NRVs in a sample may violate the assumption of nonintersecting IRFs for several items. Consequently, the item reliabilities could not be estimated without bias for these items. Second, the calibration samples were used so that the results could not be attributed to the reduced power of U_3 due to bias caused by having NRVs in the sample.

If the π_g s were estimated in a sample that contained both FRVs and NRVs, the π_g s and their ordering might be systematically different from results in a model-fitting sample. For example, due to cheating the item that was the most difficult in a group of FRVs might no longer appear to be the most difficult in a mixed group. The detection of NRVs would, therefore, be more difficult because the $\hat{\pi}_g$ s and their ordering were partly produced by these NRVs. In the context of parametric IRT, Kogut (1987) showed that the power of a person-fit statistic is reduced if the item difficulty is estimated in a calibration sample that includes NRVs. Because the reduced power of a statistic may vary for different types of NRVs (Kogut, 1987), the use of $\hat{\pi}_g$ values estimated in a sample with NRVs may confound the rate of detection for cheating and guessing simulees with the power of the statistic.

Agreement Between Replications

The level of agreement between U_3 values in two replications was determined using Gower's (1971) coefficient. This coefficient is based on the sum of abso-

lute differences between two variables and is normed against the admissible range of these variables. With this coefficient, it is possible to compare the values of U3 within each cell. Let U_i be the U3 value of person i ($i = 1, \dots, n$) in a dataset, $U_i^{(rep)}$ be the U3 value of this person in a replication, and R be the admissible range of U3. Then Gower's coefficient can be written as

$$G = 1 - \frac{\sum_{i=1}^n |U_i - U_i^{(rep)}|}{nR} \quad (4)$$

The maximum value of Gower's coefficient is 1, which means that the U3 values in the two replications are identical. Its minimum value is 0, which reflects maximum differences between the U3 values [see Gower (1971) for details].

Results

Item Reliability

Table 1 shows the means and SDs of ρ_g in the calibration samples for $k = 17$ and $k = 33$ and four levels of a . Because for each combination of test length and a two types of nonfit were combined with two frequencies of NRVs, four calibration samples from different cells were available to estimate the mean item reliability within a particular test. Thus, the ρ_g given in Table 1 is the mean ρ_g in one test averaged across these four replications. Table 1 shows that as a increased the mean ρ_g increased for both $k = 17$ and $k = 33$. This result agrees with results from an analytical study by Meijer et al. (1993).

Note that $a = 5.0$ corresponds to a mean ρ_g of approximately .60 (Table 1). Consequently, even with a very high a the data contained many item score patterns that were not perfectly scalable.

Table 1
 Mean and SD of Item Reliability Across
 Four Replications for Four Levels of a

a	$k = 17$		$k = 33$	
	Mean	SD	Mean	SD
.5	.051	.033	.054	.031
1.0	.167	.036	.155	.039
2.0	.348	.058	.339	.059
5.0	.635	.074	.636	.078

Detection of Cheating and Guessing

The percentages of cheating and guessing simulees detected in one dataset, averaged across eight replications, are shown in Table 2. An increase in a resulted in an increase in the rate of detection for both the cheaters and the guessers. The rate of detection was higher for the cheaters than for the guessers in all except one cell (for $k = 17$, $NNRV = 25$, and $a = 1.0$, the mean for the cheaters was 57 and the mean for the guessers was 59.5). This reflects higher U3 values for cheaters.

For fixed a and a fixed k , the rate of detection for $NNRV = 50$ was almost always higher than for $NNRV = 25$. This result can be explained by considering the detection of NRVs to be a selection procedure, in which NRVs are selected from a group consisting of FRVs and NRVs. The base rate (BR) is the proportion of simulated NRVs in the sample.

Table 2
 Mean and SD of the Percentage of Cheaters and Guessers Classified as NRVs
 Averaged Across Eight Replications

NNRV and a	Cheaters				Guessers			
	$k = 17$		$k = 33$		$k = 17$		$k = 33$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NNRV = 50								
.5	55.8	4.1	74.8	3.6	40.0	4.9	56.0	4.2
1.0	73.0	5.9	88.3	2.5	63.5	3.4	77.5	4.8
2.0	90.0	3.4	97.0	1.7	87.5	2.6	95.8	1.2
5.0	99.3	1.0	100.0	0.0	97.3	1.4	98.8	1.0
NNRV = 25								
.5	40.5	8.1	64.5	5.5	34.0	8.2	40.5	5.5
1.0	57.0	8.8	81.5	8.0	59.5	5.6	62.5	8.6
2.0	88.0	4.9	94.5	2.8	86.0	6.0	93.0	3.1
5.0	100.0	0.0	100.0	0.0	93.0	3.3	99.0	1.7

For 50 NRVs and a total sample of 450 simulees, $BR = 50 / 450 \approx .111$. For 25 simulated NRVs, $BR = 25 / 450 \approx .056$. The selection ratio (SR) is the proportion of simulees (NRV and FRV) selected. Thus, in this study $SR = BR$. The probability, $P(VNRV)$, of detecting a VNRV equals (Wiggins, 1973, p. 247):

$$P(VNRV) = BR \times SR + C \times \{[BR(1 - BR)SR(1 - SR)]\}^{1/2}, \quad (5)$$

where C equals the validity coefficient of a test for the prediction of a particular criterion. Comparing the percentage of VNRVs for $NNRV = 50$ and $NNRV = 25$ given a fixed a and a fixed k , the test characteristics are the same and, consequently, C is a constant. If the base rate equals .111, according to Equation 5,

$$P(VNRV) \approx .012 + C \times .099. \quad (6)$$

If the base rate equals .056,

$$P(VNRV) \approx .003 + C \times .053. \quad (7)$$

Because $P(VNRV)$ in Equation 7 is smaller than $P(VNRV)$ in Equation 6, a decrease of the BR makes it more difficult to classify a simulee as a NRV. From Table 2, it is clear that an increase in test length yielded a higher percentage of VNRVs.

Table 3 presents the average percentage of replicable cheating and guessing simulees across four replications of pairs of datasets. These data show that an increase in a or an increase in k yielded a higher percentage of replicable VNRVs. Furthermore, for all

cells the percentage of replicable cheating simulees was higher than the percentage of replicable guessing simulees.

Because the percentage of replicable VNRVs obviously can never exceed the smallest percentage of VNRVs found in one of two datasets, these trends can partly be explained by the percentage of VNRVs within one dataset. For example, the percentage of replicable VNRVs in two replications can never exceed 50 if this is the smallest detection percentage of VNRVs in one of these replications. In general, the percentage of VNRVs within each replication increased as a increased. Consequently, it was expected that the percentage of replicable VNRVs also would increase as a increased. The lower rate of detection in two replications for $NNRV = 25$ compared with $NNRV = 50$ can be attributed to the lower base rate in one replication for $NNRV = 25$ compared with the base rate for $NNRV = 50$.

The mean percentages of cheating and guessing simulees detected in only one of the two replications are given in Table 4. Thus, this table contains results based on the sum of the percentage of NRVs classified as NRV in Replication 1 and FRV in Replication 2 and the percentage of NRVs classified as FRV in Replication 1 and NRV in Replication 2. In general, these percentages decreased as a increased. Only in the case of $a = 1.0$, $k = 17$, and $NNRV = 25$ was the percentage of NRVs (both guessers and cheaters) detected in one replication larger than in the case of $a = .5$, $k = 17$, and $NNRV = 25$. This result can be explained by the large SD. Given the size of

Table 3
 Mean and SD of the Percentage of Replicable Cheaters and Guessers Classified as NRVs
 Averaged Across Four Pairs of Replications

NNRV and a	Cheaters				Guessers			
	$k = 17$		$k = 33$		$k = 17$		$k = 33$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NNRV = 50								
.5	25.0	8.1	54.5	3.6	17.5	2.2	29.5	4.3
1.0	53.5	6.7	79.0	2.4	43.0	5.0	60.0	2.4
2.0	80.5	2.2	94.0	2.8	76.0	2.2	92.5	.9
5.0	98.5	.9	100.0	0.0	94.5	1.7	98.0	.4
NNRV = 25								
.5	18.0	4.5	42.0	6.0	11.0	1.7	20.0	4.0
1.0	37.0	8.7	68.0	8.5	33.0	2.3	54.0	5.4
2.0	80.0	2.8	89.0	3.3	73.0	7.1	87.0	4.4
5.0	100.0	0.0	100.0	0.0	86.0	6.6	98.0	2.0

Table 4
 Mean and SD of the Percentage of Cheaters and Guessers Classified as NRVs
 in One of Two Replications Averaged Across Four Pairs of Replications

NNRV and <i>a</i>	Cheaters				Guessers			
	<i>k</i> = 17		<i>k</i> = 33		<i>k</i> = 17		<i>k</i> = 33	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NNRV = 50								
.5	57.5	7.5	40.5	5.0	45.0	1.7	53.0	4.1
1.0	39.0	5.9	18.5	3.6	41.0	8.8	35.0	2.2
2.0	19.0	1.7	6.0	2.8	23.0	4.1	6.5	.9
5.0	1.5	.9	0.0	0.0	5.5	1.7	2.0	1.4
NNRV = 25								
.5	45.0	6.2	45.0	5.9	46.0	8.2	41.0	9.4
1.0	54.0	9.5	27.0	5.9	53.0	6.6	37.0	7.2
2.0	16.0	4.0	11.0	3.3	26.0	6.0	13.0	4.4
5.0	0.0	0.0	0.0	0.0	14.0	6.6	2.0	2.0

the SDs per cell, this reversal could well be due to sampling error.

The percentage of simulees detected in neither replication was not tabulated, because it follows logically from the results in Tables 3 and 4. For example, in the case of $a = .5$, $k = 17$, and $NNRV = 25$ the percentage of cheating simulees detected in both replications was 18% (Table 3), whereas the percentage of cheaters detected in just one or the other of the replications was 45% (Table 4). Thus, the percentage of cheaters detected in neither replication was $100\% - 18\% - 45\% = 37\%$.

Agreement of U3 Values

The level of agreement of U3 values between the two replications increased both in the cheating and the guessing condition as a or k increased. In gen-

eral, the level of agreement was higher in the cheating than in the guessing condition. Table 5 shows that across all cells in the design, Gower's coefficient ranged from .833 in the case of $a = .5$, $k = 17$, and $NNRV = 50$ guessing simulees to .988 in the case of $a = 5.0$, $k = 33$, and $NNRV = 50$ cheating simulees. The data in Table 5 show that as either a or test length decreased, or when a NRV involved fewer Guttman errors on the most difficult items, there was less similarity between the values of U3 in two replications for the same simulee.

Discussion

The rate of detection of NRVs increased with increasing item discrimination (and therefore reliability) of the items, test length, and the number of NRVs in the total group. In addition, the rate of detection

Table 5
 Mean and SD of Gower's Coefficient Averaged Across
 Four Pairs of Replications for Cheaters and Guessers

NNRV and <i>a</i>	Cheaters				Guessers			
	<i>k</i> = 17		<i>k</i> = 33		<i>k</i> = 17		<i>k</i> = 33	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NNRV = 50								
.5	.839	.006	.892	.003	.833	.005	.888	.007
1.0	.870	.004	.903	.003	.859	.006	.896	.003
2.0	.928	.006	.945	.003	.908	.003	.933	.001
5.0	.936	.002	.988	.002	.963	.003	.974	.002
NNRV = 25								
.5	.842	.002	.889	.004	.839	.004	.888	.004
1.0	.861	.005	.905	.003	.865	.002	.899	.005
2.0	.920	.002	.935	.005	.918	.001	.937	.003
5.0	.975	.001	.981	.002	.969	.030	.974	.001

was also a function of the type of NRV (i.e., guesser versus cheater).

In general, person-fit analysis will be used as an exploratory technique to find respondents who behave unexpectedly on the basis of an IRT model or with respect to other examinees in the group. In the relatively rare cases in which a particular type of nonfitting behavior is expected to underlie item responses, it is advisable to construct tests in which the most difficult items provoke nonfitting responses. Nonfitting behavior may be difficult to recognize ad hoc if the items used are not specifically selected to elicit this type of behavior.

The present study supports the conclusion of Reise & Due (1991) that long tests with item difficulties along a broad range should be used to identify lack of fit of persons to a specified IRT model. However, an item or test characteristic that is not favorable for effective person-fit analysis may, to a large degree, be compensated for by another item or test characteristic that is more favorable. For example, although it is not desirable to use short tests for person-fit analysis, the use of highly reliable (highly discriminating) items may yield a rate of detection that is approximately the same as for longer tests with weakly discriminating items. Table 2 shows that for mean item reliability of .16 ($a = 1.0$), $k = 33$, and $N_{NRV} = 50$ cheating simulees, the percentage of VNRVs was approximately the same as for mean item reliability of .36 ($a = 2.0$), $k = 17$, and $N_{NRV} = 50$ cheating simulees (88.3 and 90, respectively). Consequently, if it is not possible to select many items with widely spaced difficulties (Reise & Due, 1991), a smaller number of highly reliable items should be selected so as to permit the same rate of detection.

In practice, IRT item discriminations range from 0.0 to 2.0 (Hambleton & Swaminathan, 1985, p. 36). It can be concluded that for "realistic" situations (a standard normal distribution of θ and $a = .5, 1.0$, and 2.0 , which correspond to mean item reliabilities of approximately .05, .16, and .34, respectively): (1) agreement (Gower's coefficient) between person-fit values would range from .833 to .945 (Table 5); and (2) the percentage of NRVs detected in two replications would range from 11

to 94 (Table 3).

Note that the higher values (.945 and 94, respectively) were obtained with mean item reliability of .34 ($a = 2.0$), 33 items, and a standard normal distribution of θ . Therefore, it is possible to obtain high agreement and detection results for 33-item tests and a mean item reliability of .34 ($a = 2.0$). For 17-item tests with mean item reliability of .34 and a standard normal distribution of θ , the percentage of NRVs detected in two replications ranged from 73 to 80.5 (Table 3). Thus, even for 17-item tests, relatively high percentages of NRVs can be detected if the items are sufficiently reliable.

References

- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.
- Drasgow, F., Levine, M. V., Williams, E. A., McLaughlin, M. E., & Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement*, *13*, 285-299.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857-871.
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383-392.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction* (pp. 60-90). Princeton NJ: Princeton University Press.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*, 133-146.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. Homewood IL: Dow Jones-Irwin.
- Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, *43*, 193-206.
- Kogut, J. (1987). *Detecting aberrant response patterns in the Rasch model* (Report 87-3). Enschede: University of Twente, Department of Education.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the ap-

- propriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.
- Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1993). *Reliability estimation for single dichotomous items based on Mokken's IRT model*. Manuscript submitted for publication.
- Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, 14, 283-298.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. New York: De Gruyter.
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6, 417-430.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75-106.
- 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.
- Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika*, 52, 217-233.
- Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, 52, 79-97.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, 10, 55-73.
- Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 83-108). New York: Academic Press.
- van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets & Zeitlinger.
- van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13, 267-298.
- Wiggins, J. S. (1973). *Personality and prediction*. Reading MA: Addison-Wesley.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Author's Address

Send requests for reprints or further information to Rob R. Meijer, Faculteit der Toegepaste Onderwijskunde/OMD, Universiteit Twente, Postbus 217, 7500 AE Enschede, The Netherlands. Internet: meijer@edte.utwente.nl.