The Influence of the Presence of Deviant Item Score Patterns on the Power of a Person-Fit Statistic

Rob R. Meijer

University of Twente

Studies investigating the power of person-fit statistics often assume that the item parameters that are used to calculate the statistics are estimated in a sample without misfitting item score patterns. However, in practical test applications calibration samples likely will contain such patterns. In the present study, the influence of the type and the number of misfitting patterns in the calibration sample on the detection rate of the ZU3 statistic was investigated by means of simulated data. An increase in the number of misfitting simulees resulted in a decrease in the power of ZU3. Furthermore, the type of misfit and the test length influenced the power of ZU3. The use of an iterative procedure to remove the misfitting patterns from the dataset was investigated. Results suggested that this method can be used to improve the power of ZU3.
Index terms: aberrance detection, appropriateness measurement, nonparametric item response theory, person fit, person-fit statistic ZU3.

In applications using item response theory (IRT) models it is often assumed that the data contain item score patterns of persons whose answering behavior does not fit a specified IRT model. These item score patterns should be detected because scores of such persons may not be adequate descriptions of their trait level.

Person-Fit Research

Recently, several person-fit statistics have been proposed to detect anomalous score patterns (e.g., Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979; Meijer, 1994; Molenaar & Hoijtink, 1990; van der Flier, 1982). In a number of studies, person-fit statistics have been applied successfully. For example, Harnisch & Linn (1981) used person-fit statistics to identify schools that had curricula that did not match test content; van der Flier (1982) used person-fit statistics to distinguish students with a different ethnic background on an intelligence and a developmental test; Tatsuoka & Tatsuoka (1983) detected examinees with erroneous rules of operation on an arithmetic test; Miller (1986) identified school classes that had a poor match between test content and instructional coverage; and Schmitt, Cortina, & Whitney (1993) used person-fit statistics to detect unmotivated test takers.

Some person-fit statistics assume a parametric IRT model, some are defined in the context of a nonparametric IRT model, and others do not use IRT. Meijer & Sijtsma (1995) provide a review of these statistics. This study used only person-fit methods defined in a nonparametric IRT model context. An advantage of nonparametric IRT models is that they are often less restrictive with respect to the data than parametric models. However, measurement is restricted to an ordinal level, whereas parametric models allow measurement on an interval or ratio scale. For a discussion that favors ordinal scaling see Cliff & Donoghue (1992).

In person-fit measurement, two steps can be distinguished. First, a model is fit to the data and item parameters are estimated. Second, person-fit statistics are calculated in a new sample using the estimated item parameters from the calibration sample and persons with inflated statistic values are classified as aberrant or nonfitting.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 2, June 1996, pp. 141–154

0146-6216/96/020141-14$1.95

Parametric Person Fit

Several studies have investigated the power of person-fit statistics to detect aberrant patterns (e.g., Drasgow et al., 1985; Reise & Due, 1991). Typically, model-fitting response vectors (FRVs) are generated according to an IRT model, and nonfitting response vectors (NRVs) are generated according to some realistic type of aberrant response behavior. Then, the percent of NRVs is determined that are successfully identified by means of a person-fit statistic [valid NRVs (VNRVs)] given a fixed error rate of FRVs classified as NRVs (Type I error). In general, in these studies it was assumed that the item parameters were known (e.g., from a previous calibration sample) and that the IRT model fit the data (Step 1). In practice, however, a sample may contain an unknown proportion of NRVs. This may affect the power of the person-fit statistics; that is, the percent of NRVs correctly identified as nonfitting given a fixed Type I error rate.

Using parametric IRT modeling, Levine & Drasgow (1983) conducted several studies to investigate the influence of NRVs on the power of a log-likelihood statistic, l, in the context of the three-parameter logistic model (3PLM; e.g., Lord, 1980, p. 12). Let X_g be the binary (0,1) item score on item g, where a "1" denotes a correct or keyed response, and a "0" denotes an incorrect or not keyed response. Let P denote a probability. Furthermore, let θ denote the attribute being measured, let $\hat{\theta}$ be the maximum likelihood estimate of θ, and let $P_g(\hat{\theta})$ denote the probability that a person with measurement value θ provides a correct or keyed response to item g [this probability defines the item response function (IRF)]. Then l can be written as

$$l = \sum_g \left\{ X_g \ln P_g(\hat{\theta}) + (1 - X_g) \ln \left[1 - P_g(\hat{\theta}) \right] \right\}, \tag{1}$$

where the sum is across all items. l will be a relatively large negative value if a person responds incorrectly ($X_g = 0$) to items for which his/her probability of a correct response according to the model is relatively high, or if the person responds correctly ($X_g = 1$) to items for which his/her probability is relatively low. Levine & Drasgow (1983) concluded that the power of l was not seriously affected even with many NRVs in the calibration sample and that the detection rate with empirical test data was comparable to the detection rate with simulated data.

In the context of the Rasch model (e.g., Baker, 1992, pp. 114–170), Kogut (1987) investigated the influence of NRVs on the power of two other parametric person-fit statistics: l_z and M. l_z is a standardized version of l and was proposed by Drasgow et al. (1985) because l was confounded with $\hat{\theta}$. To obtain l_z, the expected value of l and the variance of l across replications are needed. The expected value is given by

$$l = \sum_g \left\{ X_g \ln P_g(\hat{\theta}) + (1 - X_g) \ln \left[1 - P_g(\hat{\theta}) \right] \right\}, \tag{2}$$

and the variance is given by

$$\mathrm{V}(l) = \sum_g P_g(\hat{\theta}) \left[1 - P_g(\hat{\theta}) \right] \left\{ \ln \left[\frac{P_g(\hat{\theta})}{1 - P_g(\hat{\theta})} \right] \right\}^2. \tag{3}$$

Using these results, l_z equals

$$l_z = \frac{l - \mathrm{E}(l)}{\mathrm{V}(l)^{1/2}}. \tag{4}$$

In the context of the Rasch model, Molenaar & Hoijtink (1990) proposed M as a simplified version of l. They showed that for the Rasch model, given a fixed number-correct score, M differed from l only by a constant. Let b denote the item difficulty as defined in IRT (Lord, 1980, p. 12); then M equals

$$M = -\sum_{g} b_g X_g . \quad (5)$$

Using Rasch homogeneous datasets, Kogut (1987) concluded that, as a result of the presence of deviant item score patterns in the sample, the power of l_z and M was seriously reduced. The reduced power was the result of the biased estimation of the bs in samples with both FRVs and NRVs. Bias was defined as the difference between the numerical values of the bs estimated in a sample with FRVs and NRVs and the bs estimated in a sample with only FRVs.

The conflicting findings of the Levine & Drasgow (1983) study and the Kogut (1987) study may be the result of the idiosyncrasies of both studies. In one of the studies conducted by Levine & Drasgow (Study 2, pp. 123–125), 3,000 FRVs were simulated according to the 3PLM using the estimated item parameters from a previous calibration of the Scholastic Aptitude Test (Verbal). NRVs were simulated by modifying 200 FRVs: For each vector, 20% of the item scores were randomly selected and, regardless of the answer to these items, a correct answer was substituted for the item scores with probability .2. This simulation procedure corresponds to persons guessing at random to 20% of the items on a test with items that have five alternatives. The power of l (i.e., the percentage of NRVs detected given a fixed percentage of FRVs misclassified) was compared using the parameters estimated in the sample with FRVs and in the sample with only FRVs and NRVs.

In the Kogut (1987) study, 2,000 FRVs were generated for a 20-item test using the Rasch model. Three types of 500 NRVs were simulated as follows:

1. Item scores were simulated under the Rasch model, with the exception of the item scores on the five most difficult items or the five easiest items, which were simulated with a probability of a correct answer equal to .2, .25, or .5. Thus, for each test 25% of the items were altered;
2. Item scores were simulated according to the 3PLM with item discrimination parameters (a) equal to 1 and guessing parameters (c) equal to .2, .25, or .5;
3. Item scores were simulated using distinct θs on two different subsets of items (the 5 easiest and the 5 most difficult items). Datasets of 2,500 persons were created by merging the 2,000 FRVs with the 500 NRVs. The power (percent of NRVs correctly classified as NRVs given a 5% error rate) of l_z and M was compared using the bs estimated in the sample with only FRVs and the bs estimated in the sample with both FRVs and NRVs.

Comparing the designs of the two studies, possible explanations for the different findings are (1) the type of statistic: l was used in the Levine & Drasgow (1983) study, whereas in the Kogut (1987) study l_z and M were used—some statistics may be more sensitive to the presence of NRVs than others; (2) the number of NRVs in the sample: in the Levine and Drasgow study, the percent of NRVs in the dataset was 6.7, whereas in the Kogut study this percent was 20—the higher percent of NRVs may be responsible for the reduced power in the Kogut study; (3) the type of NRVs: the studies simulated different types of NRVs and it has been shown (e.g., Meijer, Molenaar, & Sijtsma, 1994) that some types of NRVs are easier to detect than others.

In addition to the power study, Kogut (1987) used the following iterative estimation procedure to improve the power of M: (1) item and person parameters were estimated in the datasets containing both FRVs and NRVs; (2) item scores were simulated using the estimated parameters obtained in Step 1; and (3) M values were calculated and response vectors with the 5% highest values (indicating aberrance) were removed from the dataset. Steps 1–3 were repeated until no clear improvement of the power of M was obtained. Kogut (1987) showed that this method was quite successful in removing the NRVs from the dataset; for several cases, the power of M was considerably improved after three iterations.

In a nonparametric IRT context, it is unknown how person-fit statistics will be influenced by the presence of NRVs in the dataset. It is clear that the results obtained by Levine & Drasgow (1983) and Kogut (1987) cannot be easily generalized to nonparametric IRT modeling. Therefore, these studies were ex-

tended here to a nonparametric IRT framework.

Nonparametric Person Fit

Van der Flier (1980, 1982) developed a standardized version of the person-fit statistic U3 in the context of the nonparametric Mokken (1971; Mokken & Lewis, 1982) monotone homogeneity model (MHM). The MHM is based on the assumptions of unidimensionality, local stochastic independence, and monotonicity of θ. The MHM restricts the IRFs to be nondecreasing, but they may intersect. The MHM has been successfully applied to empirical data by Kingma & TenVergert (1985); Meijer, Sijtsma, & Smid (1990); and Sijtsma & Verweij (1992). Let π_g denote the proportion-correct score on item g ($g = 1, \ldots, k$) and let r denote the realization of the number-correct score of a person on the test ($X = r$). Then U3 can be written as

$$U3 = \frac{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \sum_{g=1}^{k} X_g \ln\left(\frac{\pi_g}{1-\pi_g}\right)}{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \sum_{g=k-r+1}^{k} \ln\left(\frac{\pi_g}{1-\pi_g}\right)}. \tag{6}$$

Van der Flier (1980, 1982) showed that a standardized version of U3, denoted ZU3, was approximately standard normally distributed given an invariant ordering of persons according to their θs. To obtain ZU3, the expected value and variance of U3 across replications are needed. Note that for a fixed number-correct score all terms in Equation 6 are constant, except for

$$\sum_{g=1}^{k} X_g \ln\left(\frac{\pi_g}{1-\pi_g}\right). \tag{7}$$

Van der Flier (1980, 1982) derived the mean (η) and variance (β) of Equation 7 as

$$\eta = \sum_{g=1}^{k} \pi_g \ln\left(\frac{\pi_g}{1-\pi_g}\right) + \frac{\sum_{g=1}^{k} \pi_g(1-\pi_g) \ln\left(\frac{\pi_g}{1-\pi_g}\right)}{\sum_{g=1}^{k} \pi_g(1-\pi_g)} \left(r - \sum_{g=1}^{k} \pi_g\right), \tag{8}$$

and

$$\beta = \sum_{g=1}^{k} \pi_g(1-\pi_g) \left[\ln\left(\frac{\pi_g}{1-\pi_g}\right)\right]^2 - \frac{\left[\sum_{g=1}^{k} \pi_g(1-\pi_g) \ln\left(\frac{\pi_g}{1-\pi_g}\right)\right]^2}{\sum_{g=1}^{k} \pi_g(1-\pi_g)}. \tag{9}$$

Then, for a fixed number-correct score U3 should also be normally distributed with

$$E(U3|X=r) = \frac{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \eta}{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \sum_{g=k-r+1}^{k} \ln\left(\frac{\pi_g}{1-\pi_g}\right)}. \tag{10}$$

and

$$V(U3|X=r)=\frac{\beta}{\left[\sum_{g=1}^{r}\ln\left(\frac{\pi_g}{1-\pi_g}\right)-\sum_{g=k-r+1}^{k}\ln\left(\frac{\pi_g}{1-\pi_g}\right)\right]^2}. \quad (11)$$

ZU3 can be written as

$$ZU3=\frac{U3-E(U3)}{V(U3)^{1/2}}. \quad (12)$$

The ZU3 statistic is the only statistic that can be used in a nonparametric IRT context that has theory-based significance levels. These significance levels were found to be highly in agreement with the significance levels found in a sample distribution using tests consisting of 17 and 33 items, a standard normal distribution for θ, and 3PLM IRFs (Meijer, Muijtjens, & van der Vleuten, 1996; Van der Flier, 1982). Furthermore, van der Flier (1982) concluded that for sets of 17 and 29 items with uniformly or normally distributed π_g values both the U3 and ZU3 distributions within different score groups could be combined into one common distribution. Consequently, the values of both statistics can be computed across different score groups. Meijer et al. (1996) investigated the sample distribution for tests that consisted of 17 and 33 items with item scores generated under the 3PLM using item parameters often found in practice and a standard normal distribution for θ. It was found that the 90th, 95th, and 99th percentile values of ZU3 were 1.30, 1.64, and 2.40, respectively. Thus, these empirically observed significance levels were largely in agreement with theoretically-based significance levels.

Purpose

This study was designed (1) to investigate the power of a nonparametric person-fit statistic ZU3 as a function of the number and the type(s) of NRVs present in the calibration sample, and (2) to investigate the usefulness of an adapted version of the iterative estimation method (Kogut, 1987) in a nonparametric IRT context. From Kogut (1987) and Levine & Drasgow (1983), it appeared to be appropriate to vary the number and the type of NRVs in the calibration study. Test length was varied because Meijer et al. (1994) showed that test length had a large effect on the power of a nonparametric person-fit statistic. Note that the parametric studies described above did not systematically vary these characteristics and it was thus not clear how these characteristics would influence the power of a person-fit statistic.

Method

Two studies were conducted. In the first study, the power of ZU3 was investigated as a function of the number and the type of NRVs that were present in a calibration sample. Analogously to the Levine & Drasgow (1983) and the Kogut (1987) study, power was defined as the proportion of NRVs classified as aberrant given a fixed proportion of FRVs classified as aberrant. In the second study, the power of ZU3 was investigated as a function of the number of iterations that was used to delete NRVs from the dataset.

Types of Nonfitting Responses

Many potential causes of aberrant response behavior can be distinguished. If a researcher wants to identify persons whose item score patterns are unexpected given a specified IRT model, aberrant score patterns should be specified. In this study, two often-mentioned and encountered types of aberrant behavior that differ in the way they deviate from the model assumptions were selected: guessing and cheating.

Persons taking a multiple-choice test will benefit from guessing the answers to items they do not know

(although this may depend on the scoring formula that is applied). This type of guessing was simulated in the Levine & Drasgow (1983) study. Another type of guessing was empirically studied by van den Brink (1977). He described persons who took a multiple-choice exam only to familiarize themselves with the questions that would be asked. Because returning an almost completely blank answer sheet may focus a teacher's attention on the ignorance of the examinee, each examinee randomly guessed the correct answers on almost all items of the test.

Cheating behavior was reported by, for example, Frary (1993). He reported on a study of answer copying in an administration of a test in elementary schools. In these schools, salary adjustments depended on the results obtained by the students and it was feared that some teachers might encourage cheating to raise their students' scores on the test. Frary could identify (by means of a statistic that was specially developed to detect answer copying) persons with item scores that were so much alike that answer copying was very likely. A similar type of cheating was recently reported in the Netherlands ("Re-examination," 1994) in which high school students had to take the final exam again after it appeared that the teacher had changed the answers on the examination to raise the grades.

Another type of cheating is test preview. Recently, in the Netherlands in an administration of a nationwide examination, it became known that some students had been aware of some of the questions that would be asked on the exam—a relative of one of the students worked at the institute where the test was constructed. He had given the exam to a student who had sold some of the questions to friends and classmates. (For an analogous case of test preview in the U.S.A., see Hulin, Drasgow, & Parsons, 1983, p. 113.). In this study, aberrant item score patterns that are in agreement with all three types of aberrant behavior discussed above were simulated.

Study 1

Step 1. Datasets of 2,000 FRVs were simulated (for the simulation procedure see Sijtsma & Molenaar, 1987) both for a 17-item test and for a 33-item test using the 3PLM and a standard normal distribution for θ. Item discrimination was drawn from a uniform distribution with $a \sim U[.5,1.5]$; bs ranged from [−2.0,2.0] and were equidistant with distance between the items equal to .25 in the 17-item test and equal to .125 in the 33-item test; the guessing parameter was drawn from a uniform distribution with $c \sim U[0.0,.2]$. Note that these datasets were generated according to parameters often found in practice (e.g., Hambleton & Swaminathan, 1985). Although this study was conducted in a nonparametric IRT context, a parametric IRT model was used to generate the item scores. The 3PLM was selected because this model is the most widely used nonrestrictive parametric IRT model according to which data can be generated that also satisfies the Mokken MHM. The difference between the models is that the 3PLM assumes that the IRFs are logistic, whereas in the MHM the form of the IRFs is left free as long as they are nondecreasing.

Step 2. Four types of datasets containing 2,000 NRVs were simulated: cheating on the most difficult items, cheating on items of medium difficulty, guessing on all items of the test, and guessing on 20% of the items in the test. These types of NRVs were selected because they represented realistic types of NRVs (cheating and guessing) and because they represented different levels of severeness of nonfitting score patterns. As will be explained further below, cheating on items of medium difficulty will result in item score patterns that are less aberrant than cheating on the most difficult items. By selecting these types and levels, the detection rate of ZU3 was studied in various situations.

The first dataset consisted of cheaters who had a negative θ value (sampled from a standard normal distribution) and answered the items according to the 3PLM, except for the three most difficult items on the 17-item test and the six most difficult items on the 33-item. These items were scored as if the examinees had correctly answered them. Possible explanations are that cheaters correctly answered these items by obtaining answers from a more able examinee (assuming that this cheating always resulted in correct

answers) or that cheaters knew the answers to these questions because they had seen the test already and memorized answers to some of the most difficult items (test preview).

The second dataset was generated according to the same procedure as the first dataset except that cheating took place on the three (17-item test) or six (33-item test) items of medium difficulty. Again, a possible explanation for this type of cheating is test preview in which the items of medium difficulty were known in advance.

The third dataset consisted of guessers who answered all items with a probability of a correct answer of .20, which corresponds to answering an item with five alternatives by randomly guessing. These simulated data were in agreement with the guessing behavior studied by van den Brink (1977).

A fourth dataset was generated according to the same procedure as Levine & Drasgow (1983). Thus, FRVs were simulated according to the 3PLM [the same item parameters were used as for the simulation of the FRVs above] and NRVs were simulated by modifying the FRVs. For each vector, 20% of the item scores were randomly selected and regardless of the answer to these items a correct answer was substituted for the item scores with probability .2. This simulation procedure corresponds to persons guessing at random to 20% of the items on a test with five-alternative items.

From a theoretical point of view, these four conditions were interesting because they represented different types of NRVs. For cheating on the most difficult items in the most favorable situation (the situation in which the probability of a correct answer of a cheating person was largest, that is if $\theta = 0.0$, $a = .5$, $b = 1.5$, and $c = .2$), for the three and six most difficult items the probability of correctly answering an item on the basis of θ was at most .46. For other combinations of parameters, this value was always smaller. For cheating on items of medium difficulty, the probability in the most favorable situation ($\theta = 0.0$, $a = .5$, $b = -.25$, and $c = .2$) was .62. Thus, it was expected that cheating on the most difficult items would be easier to detect than cheating on items of medium difficulty because the first type of NRV is clearly more aberrant than the latter. With respect to guessing on all items, guessing on all items is clearly more aberrant than guessing on only 20% of the items; consequently, it was expected that the latter type of NRV would be more difficult to detect than the former type.

Another point of theoretical interest is that the bias of the π_g estimates ($\hat{\pi}_g$) may be influenced differently by the different types of NRVs. If bias is defined as the difference between the $\hat{\pi}_g$s obtained in the sample with FRVs and NRVs minus the $\hat{\pi}_g$s obtained in the sample with FRVs, then it can be expected that cheating on items of medium difficulty and the most difficult items may result in a positive bias of the $\hat{\pi}_g$s of these items (more "1" scores on the difficult items than expected under the model) and guessing on all items may result in a negative bias of the $\hat{\pi}_g$s on the easiest items (fewer "1" scores on the easiest items than expected under the 3PLM). This study investigated whether these different types of bias could be removed by the iterative estimation procedure.

Furthermore, it was expected that as the number of NRVs increased in the calibration samples, the bias of the $\hat{\pi}_g$s would increase, and as a result the power of ZU3 would decrease. However, it was not known to what degree the different types of NRVs and the number of persons in the sample would influence the detection rate.

Step 3. From each dataset of 2,000 NRVs, 100, 200, 300, and 400 vectors were sampled (no overlap between the samples) and substituted for a corresponding number of FRVs in the datasets generated above in Step 1. Consequently, for each of the four types of NRVs, four datasets were created with 5%, 10%, 15%, and 20% NRVs. Thus, 16 datasets were created (4 types of NRVs × 4 levels of number of NRVs). In each dataset, the π_g values were estimated that were needed to calculate ZU3. Thus, these 16 datasets were the calibration samples.

Step 4. 16 datasets with 2,000 NRVs were created according to the same procedure as in Step 2. For each simulee in these datasets, ZU3 values were calculated using the π_g values estimated in Step 3.

Note that datasets with both FRVs and NRVs in Step 4 were not used to investigate the power of ZU3. This

was to avoid the risk that the power of ZU3 was confounded by the unequal base rate (i.e., the unequal proportion of NRVs) in the samples. Meijer et al. (1994) showed that the larger the base rate, the easier it was to classify a guesser or a cheater as a NRV.

Study 2

1. Four datasets of 2,000 simulees were generated according to the same procedure as in Step 3 of Study 1 with (1) 20% cheaters on the most difficult items, (2) 20% cheaters on items of medium difficulty, (3) 20% guessers on all items, and (4) 20% guessers on 20% of the items.
2. π_g values were estimated in the datasets.
3. For each dataset, ZU3 values were calculated and simulees with ZU3 > 1.64 were classified as aberrant and were deleted.
4. Steps 2 and 3 were repeated until no clear improvement of the power of ZU3 was found.

Analysis

The π_g values required to compute ZU3 were estimated in each of the 16 calibration samples containing 5%, 10%, 15%, or 20% NRVs. These statistics were computed as the first step in the analysis. ZU3 then was computed in each of the 16 samples with only NRVs using the $\hat{\pi}_g$s from the calibration samples. Next, the item score patterns were ordered according to increasing ZU3 (indicating nonfit) and in each sample the number of NRVs that were classified by ZU3 as NRVs was determined. The following three critical values were used: ZU3 = 1.30 (one-tailed $\alpha = .1$ error rate); ZU3 = 1.64 (one-tailed $\alpha = .05$ error rate); and ZU3 = 2.40 (one-tailed $\alpha = .01$ error rate).

Results

Study 1

Table 1 shows the percentages of cheaters and guessers correctly classified as NRVs. For a fixed α and a fixed type of NRV, both for $k = 17$ and $k = 33$, Table 1 shows that as the percent of NRVs in the calibration sample increased, the power of ZU3 decreased. For example, consider the situation $\alpha = .05$, $k = 17$, and cheating on the most difficult items. Table 1 shows that when the calibration sample consisted of only FRVs, there were 64% VNRVs; 53% with 10% NRVs; and 30% with 20% NRVs in the calibration sample.

For both cheaters and guessers for a fixed α and a fixed percent of NRVs in the calibration sample, the percent of VNRVs was larger for $k = 33$ than for $k = 17$. For example, for $\alpha = .01$, 5% NRVs, and guessing on all items, there were 53% VNRVs for $k = 17$ and 71% for $k = 33$. This is in agreement with the findings by Meijer et al. (1994) who found that NRVs were easier to detect for longer tests. However, they used calibration samples with only FRVs. Using these samples with FRVs and NRVs the same trend was observed.

Furthermore, for a fixed α, percent of NRVs, and test length (1) cheating on the most difficult items was easier to detect than cheating on items of medium difficulty, and (2) guessing on all items was easier to detect than guessing on 20% of the items. For example, for $\alpha = .05$, 15% NRVs, and $k = 33$, there were 70% VNRVs for cheating on the most difficult items and 54% for cheating on the medium difficulty items; for guessing on all items there were 68% VNRVs and 44% for guessing on 20% of the items. This was in agreement with theoretical expectations. Note that for $\alpha = .05$, the presence of 5% NRVs had only a minor influence on the detection rate of ZU3, whereas increasing the number of NRVs increased the reduction in power. The reduction was most explicit for cheaters on the most difficult items and for guessers on all items (approximately 30% reduction for 20% NRVs in the calibration sample compared with only FRVs in the calibration sample). For example, for $\alpha = .05$, $k = 33$, and cheating on the most difficult items, there was a 31% (89% – 58%) difference in percent of VNRVs.

For cheaters on items of medium difficulty and guessers to 20% of the items, the influence on the

Table 1
Percent of Cheaters and Guessers Classified as NRV (VNRVs) as a Function of the Type I Error Rate (α) and the Percent of NRVs in the Calibration Sample for 17- and 33-Item Tests

α and % NRV in Calibration Sample	Cheaters				Guessers			
	Most Difficult Items		Items of Medium Difficulty		All Items		20% of Items	
	$k = 17$	$k = 33$	$k = 17$	$k = 33$	$k = 17$	$k = 33$	$k = 17$	$k = 33$
$\alpha = .01$								
none	53	73	34	54	55	74	25	36
5%	51	70	32	52	53	71	24	35
10%	45	63	30	47	46	65	21	32
15%	36	56	24	40	38	56	18	29
20%	23	45	18	33	26	47	15	25
$\alpha = .05$								
none	64	89	37	67	62	86	37	54
5%	62	87	35	64	59	83	35	51
10%	53	79	32	60	51	78	32	47
15%	34	70	27	54	45	68	26	44
20%	30	58	22	48	35	59	24	41
$\alpha = .10$								
none	73	92	54	75	70	89	42	62
5%	70	90	52	72	67	87	39	59
10%	65	82	48	69	59	80	36	55
15%	57	74	41	63	52	72	32	52
20%	45	61	35	55	41	60	28	48

detection rate was not as high (approximately 15% reduction). For $\alpha = .01$ and $\alpha = .10$ the same trends (and percentages) were obtained. For example, for $\alpha = .05$, $k = 17$, and cheating on the items of medium difficulty, there were 37% VNRVs with only FRVs in the calibration sample, whereas there were 22% VNRVs with 20% NRVs in the calibration sample. For $\alpha = .01$, these percentages were 34% and 18% and for $\alpha = .10$ they were 54% and 35%, respectively.

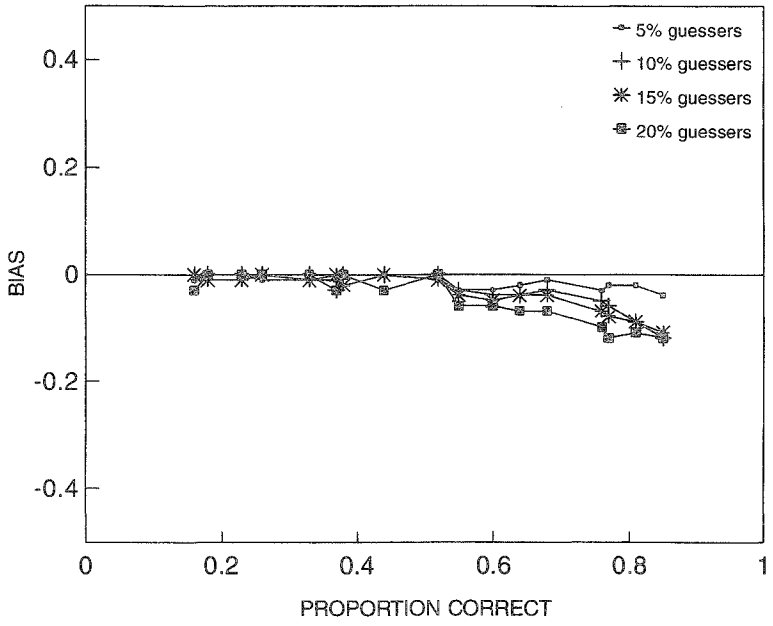
The reduced power of ZU3 as a result of the presence of NRVs in the calibration sample may be due to the biased estimation of $\hat{\pi}_g$. For example, due to cheating, the item that was most difficult in a group with only FRVs might no longer appear to be the most difficult in a group with FRVs and NRVs. The detection of NRVs would, therefore, be more difficult because the $\hat{\pi}_g$s and their ordering were partly produced by these NRVs.

Figure 1a shows the bias of the $\hat{\pi}_g$ values for $k = 17$ with guessing on all items; Figure 1b shows the bias of the $\hat{\pi}_g$ values for $k = 17$ with guessing on 20% of the items. Figure 1a shows that with 5% guessers some negative bias occurred on the easiest items. An increase in the percent of guessers resulted in an increase in the bias on the easiest items, whereas the $\hat{\pi}_g$s of the more difficult items remained almost unbiased. Figure 1b shows the same trends; however, the bias for 5% guessers was almost 0, and the absolute bias as a result of the presence of 20% guessers was not larger than .04. This may explain the smaller decrease in power of ZU3 for guessers to 20% of the items than for guessers to all items.

Figure 2 shows the bias results for $k = 17$ and cheating on the most difficult items (Figure 2a) and on the items of medium difficulty (Figure 2b). Figure 2a shows that the $\hat{\pi}_g$s of the three most difficult items became positively biased, varying from approximately .04 with 5% NRVs in the dataset to approximately .16 with 20% NRVs in the dataset. The bias for items of medium difficulty was somewhat smaller (Figure 2b). Again, the smaller decrease in the power of ZU3 for cheating on items of medium difficulty may be explained by the smaller bias. The same trends (not shown here) were found for $k = 33$ (for both cheating and guessing).

Figure 1
Bias Results Due to Guessing for the 17-Item Test

a. Guessing on all Items



b. Guessing on 20% of the Items

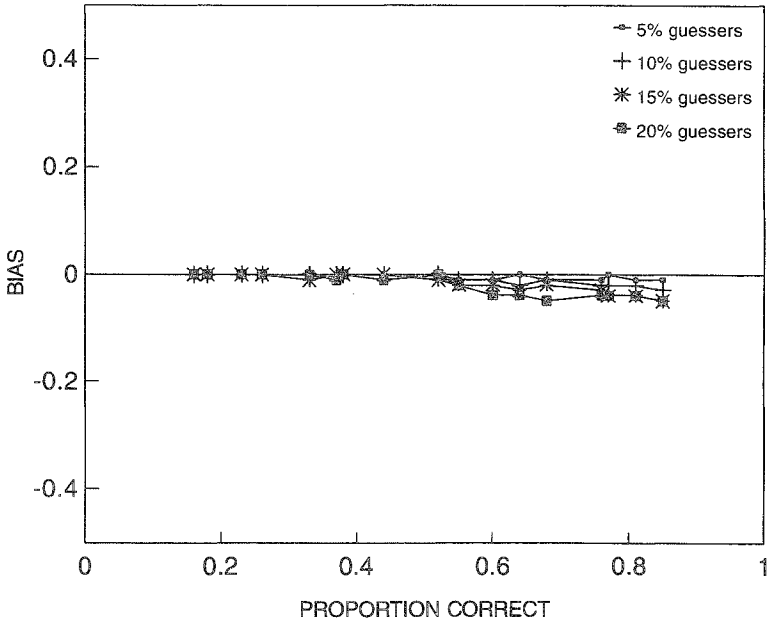
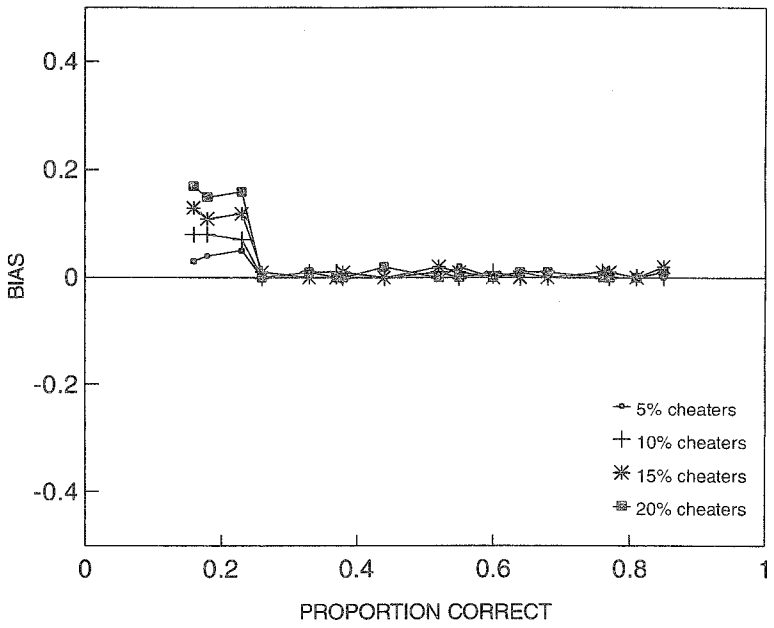
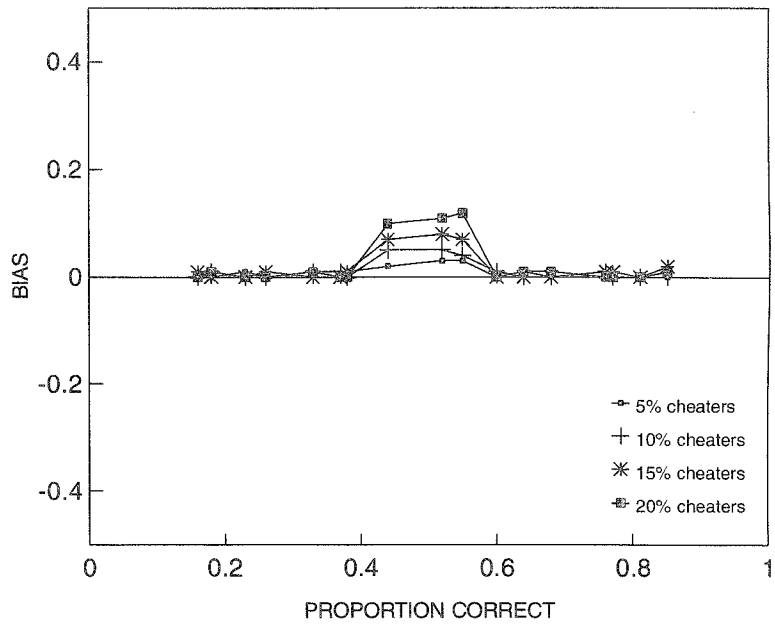


Figure 2
Bias Results Due to Cheating for the 17-Item Test

a. Cheating on the Three Most Difficult Items



b. Cheating on Medium Difficulty Items



Study 2

Table 2 shows the percent of VNRVs after deleting NRVs with ZU3 > 1.64 in four subsequent iterations. For $k = 17$, after one iteration the percent of VNRVs increased by 16% (for cheating on the most difficult items, the percent increased from 27% to 43%), 3% (cheating on items of medium difficulty), 12% (guessing on all items), and 4% (guessing on 20% of the items). For all types of NRVs, the percent of VNRVs only slightly increased after the second, third, and fourth iterations, and the percent of VNRVs that was found using the values estimated in the dataset with only FRVs (π_g FRV) was not reached.

Table 2
Percent of Cheaters and Guessers Classified as NRV Before Iteration, After Iterations 1–4, and Detected Using the Percentage VNRVs Based on π_g Values Estimated in a Sample With FRVs

	Cheaters				Guessers			
	Most Difficult Items		Items of Medium Difficulty		All Items		20% of Items	
Iteration	$k = 17$	$k = 33$	$k = 17$	$k = 33$	$k = 17$	$k = 33$	$k = 17$	$k = 33$
Before	27	56	25	46	34	57	24	43
1	43	72	28	53	46	72	28	48
2	46	80	30	59	53	83	31	49
3	50	82	30	63	55	84	32	50
4	54	86	30	65	55	84	32	52
π_g FRV	63	89	37	67	62	86	37	54

For $k = 33$, Table 2 shows that the first iteration also resulted in the largest increase in VNRVs [e.g., for cheating on the most difficult items the percent increased 16% (56% to 72%), for cheating on items of medium difficulty it increased 7%, for guessing on all items 15%, and for guessing on 20% of the items it increased 5%]. After the fourth iteration, the percentages of VNRVs for cheating on items of medium difficulty and guessing on 20% of the items were only slightly smaller than the percentages found using the values obtained with only FRVs.

The bias reduction of the $\hat{\pi}_g$s followed the same trend as the percent of VNRVs; the largest reduction was found after the first iteration, whereas smaller reductions were found after the other iterations (not shown). In general, both for $k = 17$ and $k = 33$ the Type I error after the second iteration was slightly higher than for the first two iterations. For example, for $\alpha = .05$, $k = 17$, and cheating on the most difficult items, for the second iteration the Type I error was 6%, whereas for the third and fourth iterations it was 6.5% and 7.1%, respectively. For the same conditions and $k = 33$, the Type I error was 6.1% for the second iteration, and 6.5% and 6.7% for the third and fourth iterations, respectively.

Discussion

The power of ZU3 might be seriously reduced due to the presence of NRVs in a dataset. Two factors are important: the number of NRVs in the calibration sample and the type of NRVs. The type of aberrant response behavior influences the degree to which the $\hat{\pi}_g$s become biased. The results showed that cheating on the most difficult items and guessing on all items resulted in a larger bias of the $\hat{\pi}_g$s and a lower detection rate than cheating on items of medium difficulty and guessing on 20% of the items. Although for 5% NRVs in the calibration sample the differences were small, for 10% to 20% NRVs the differences were large (see Table 1). For example, for $\alpha = .10$, $k = 17$, and guessing on all items, the percent of VNRVs decreased 3% with 5% NRVs, whereas for 10%, 15%, and 20% NRVs the percentage of NRVs decreased 8%, 7%, and 11%, respectively. Furthermore, the number of NRVs was important. For all types of NRVs, the detection rate was seriously reduced when the number of NRVs increased from 5% to 20%. For 5% NRVs, the reduction in

power was small (compared to only FRVs in the calibration sample), whereas the power was seriously reduced as the number of NRVs increased. This may explain the difference between the Kogut (1987) and the Levine & Drasgow (1983) study. In the Kogut study, there were 20% NRVs, whereas in the Levine and Drasgow study there were 6.7%.

Although ZU3 was used here, using other person-fit statistics—such as l_z, M, or the number of Guttman errors (Meijer, 1994)—would probably yield the same results. Earlier person-fit literature showed that detection rates of these statistics were similar (e.g., Kogut, 1987; Meijer, 1994). An interesting extension of this study would be to compare the detection rates of a person-fit statistic with a statistic especially developed to detect cheating behavior (Frary, 1993). These latter statistics, however, are constructed in such a way that they use information from the response option that is selected by the examinee in a multiple-choice test. Most person-fit statistics are developed for dichotomous item scores and are therefore difficult to compare with these statistics.

The use of an iterative procedure to delete NRVs gave mixed results. The type of NRVs was important, as was the test length. For relatively short tests ($k = 17$), even after four iterations the power of ZU3 stayed below the initial level of the case in which no NRVs were present. However, for relatively long tests ($k = 33$) the power was approximately the same after three or four iterations. Because at each iteration the number of FRVs that were classified as NRVs increased, the number of iterations should be kept to a minimum.

These results suggest that it is possible to remove NRVs iteratively using ZU3. Note that this procedure is a technical one in the sense that it is performed to obtain item parameters that are better suited for person-fit analysis. If a test is, for example, used to measure the trait level of a person, item score patterns should not be blindly removed on the basis of a person-fit value. However, if a dataset is available to calibrate the item difficulties, the iterative estimation procedure proposed by Kogut (1987) can also be used in a nonparametric IRT context.

Finally, users of person-fit techniques will have to decide for themselves whether the hit rates presented in this study are sufficiently high for practical applications. This will depend on the type of application envisaged. In selection situations, the number of aberrant persons not identified may be too high; in a mastery testing situation action may be justified.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Cliff, N., & Donoghue, J. R. (1992). Ordinal test fidelity estimated by an item sampling model. *Psychometrika, 57,* 217–236.

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.

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education, 6,* 153–165.

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.

Kingma, J., & TenVergert, E. M. (1985). A nonparametric scale analysis of the development of conservation. *Applied Psychological Measurement, 9,* 1–23.

Kogut, J. (1987). *Detecting aberrant response patterns in the Rasch model* (Report No. 87-3). Enschede: University of Twente, Department of Education.

Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109–131). New York: Academic Press.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Jour-*

nal 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. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311–314.
Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111–120.
Meijer, R. R., Muijtjens, A. M. M., & van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9,* 77–89.
Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8,* 261–272.
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.
Miller, M. D. (1986). Time allocation and patterns of item response. *Journal of Educational Measurement, 23,* 147–156.
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.
Re-examination after fraud teacher (Herexamen na fraude leraar). (1994, June, 13). *NRC Handelsblad,* p. 8.
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.
Schmitt, N. S., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17,* 143–150.
Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52,* 79–97.
Sijtsma, K., & Verweij, A. C. (1992). Mokken scale analysis: Theoretical considerations and an application to transitivity tasks. *Applied Measurement in Education, 5,* 355–373.
Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 7,* 215–231.
van den Brink, W. P. (1977). Het verken-effect [The scouting effect]. *Tijdschrift voor Onderwijsresearch, 2,* 253–261.
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.

Acknowledgment

The author thanks two anonymous reviewers for their comments and suggestions on an earlier draft of this paper.

Author's Address

Send requests for reprints or further information to Rob R. Meijer, Department of Educational Measurement and Data Analysis, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: meijer@edte.utwente.nl.