DOCUMENT RESUME

ED 389 758 TM 024 391

AUTHOR Meijer, Rob R.
TITLE The Influence of the Presence of Deviant Item Score Patterns on the Power of a Person-Fit Statistic. Research Report 94-1.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE Nov 94
NOTE 33p.
AVAILABLE FROM Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Evaluation Methods; Foreign Countries; Identification; *Research Methodology; *Scores; Simulation; *Test Items
IDENTIFIERS Aberrance; Calibration; Item Parameters; *Item Score Patterns; *Person Fit Measures; Power (Statistics)

ABSTRACT

In studies investigating the power of person-fit statistics it is often assumed that the item parameters that are used to calculate the statistics can be estimated in a sample without aberrant persons. However, in practical test applications calibration samples most likely will contain aberrant persons. In the present study, the influence of the type and the number of aberrant persons in the calibration sample on the detection rate of the ZU3 statistic was investigated by means of simulated data. The ZU3 is a standardized version of the person-fit U3 statistic developed by H. Van der Flier (1980). An increase in the number of aberrant simulees resulted in a decrease in the power of ZU3. Furthermore, the type of aberrant behavior influenced the power of ZU3. The use of an iterative procedure to remove the aberrant persons from the dataset was investigated. Results suggested that this method can be used to improve the power of ZU3. (Contains 4 tables, 6 figures, and 17 references.) (Author/SLD)

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

Research Report 94-1

Rob R. Meijer

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

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

Rob R. Meijer

The influence of the presence of deviant item score patterns , Rob R. Meijer - Enschede : University of Twente, Faculty of Educational Sciences and Technology, November, 1994. - 27 pages

Abstract

In studies investigating the power of person-fit statistics it is often assumed that the item parameters that are used to calculate the statistics can be estimated in a sample without aberrant persons. However, in practical test applications calibration samples most likely will contain aberrant persons. In the present study, the influence of the type and the number of aberrant persons 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 aberrant simulees resulted in a decrease in the power of *ZU3*. Furthermore, the type of aberrant behavior influenced the power of *ZU3*. The use of an iterative procedure to remove the aberrant persons 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*.

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

In applications of item response theory (IRT) models it is often assumed that the data contain item score patterns of persons whose answering behavior is aberrant or deviant on the basis of what could be expected given the model. These item score patterns should be detected because it is questionable whether the total score gives an adequate description of a person's ability or achievement level.

Recently, several person-fit statistics have been proposed to detect anomalous score patterns (e.g., Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979; Molenaar & Hoijtink, 1990; Van der Flier, 1982). Some statistics take as given a parametric IRT model, whereas other statistics were defined in the context of a nonparametric IRT model, or outside the IRT context. For a review of these statistics refer to Meijer and Sijtsma (in press). In this study we restrict ourselves to person-fit methods defined in a nonparametric IRT model context. Nonparametric IRT models have the advantage 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 a interval or ratio scale. For a discussion that favor ordinal scaling refer to Cliff and Donoghue (1992).

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

In several studies the power of person-fit statistics to detect aberrant patterns have been investigated. Typically, model-fitting response vectors (FRV) were generated according to an IRT model, and nonfitting response vectors (NRV) were generated according to some realistic type of aberrant behavior 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 and that the IRT model perfectly satisfied 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.

Using parametric IRT modeling, Levine and 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) and Kogut (1987) used a standardized version of this statistic, l_z, and the Molenaar and Hoijtink (1990) statistic, M, in the context of the Rasch model (e.g., Baker, 1992, pp. 114-170). In the first study it was concluded that the power of l was not seriously affected even with relatively many NRVs in the calibration sample. Besides, the rate of detection with empirical test data was comparable with the rate of detection with simulated data. However, in the latter study it was concluded that as a consequence of the presence of NRVs the power of l_z and M was seriously reduced. These conflicting findings may be the result of the idiosyncrasies of both studies.

In one of the studies conducted by Levine and Drasgow (1983, Study 2, pp. 123-125), 3000 FRVs were simulated according to the 3PLM using the estimated item parameters from a previous fitting of the SAT-V. NRVs were simulated by modifying 200 FRVs: for each vector 20% of the item scores was randomly chosen and irrespective 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 with five alternatives. The power of l was compared using the parameters estimated in the sample with FRVs and in the sample with FRVs and NRVs.

In the Kogut (1987) study, 2000 FRVs were generated on a 20-item test under 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 sco es on (a) the five most difficult items or (b) 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 was altered; (2) item scores were simulated according to the 3PLM with item discrimination parameters equal to 1 and guessing parameters equal to .2, .25, or .5; (3) item scores were simulated using distinct abilities on two different subsets of items (5 easiest, 5 most difficult items). Datasets of 2500 persons were created by merging the 2000 FRVs with the 500 NRVs. The power of l_z and M was compared using the item difficulties estimated in the sample with only FRVs and the item difficulties 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 kind of statistic; in the Levine and Drasgow (1983) study l was used, 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) in the Levine and Drasgow (1983) study the percentage of NRVs in the dataset was 6.7, whereas in the Kogut study this percentage was 20; the higher percentage 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, in press) that some types of NRVs are easier

to detect than others.

Furthermore, 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 (1); (3) M values were calculated and response vectors with the 5% highest values (indicating aberrance) were removed from the dataset. Step (1) through (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.

Besides the presence of a single type of NRVs in the calibration sample, another factor that may affect the power of person-fit statistics is the presence of different types of NRVs in one dataset. In studies using simulated data it is almost always assumed that there is only one type of NRVs present. In practical test situations, however, it is most likely that several types are present. For example, some examinees may cheat on the most difficult items, whereas others may guess to complete the test; or some examinees may answer the items according to the 3PLM and some may guess to all items, whereas the majority of persons answers the items according to the Rasch model. How will the presence of different types of answering behavior influence the power of a person-fit statistic ?

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 and Drasgow (1983) and Kogut (1987) cannot be easily generalized to nonparametric IRT modeling. Therefore, in this study we will (1) examine the rate of detection of a nonparametric person-fit statistic as a

function of the number and the type(s) of NRVs present in the calibration sample and (2) investigate the usefulness of an adapted version of the iterative estimation method (Kogut, 1987) in a nonparametric IRT context.

Nonparametric person-fit research

Van der Flier (1980, 1982) developed the person-fit statistic *U3* in the context of the nonparametric Mokken (1971; Mokken & Lewis, 1982) model of double monotonicity. This model is based on the assumptions of unidimensionality, local stochastic independence, monotonicity in the latent attribute (Θ), and nonintersecting item response functions (IRF).

Let P denote a probability, and let $\mathbf{X}_i$ be the item score vector of a person i (i, ..., n) with dichotomous item scores where 1 indicates a correct or keyed response and 0 indicates an incorrect or not keyed response. Furthermore, let $\mathbf{X}^*_i$ be the vector that satisfies the Guttman (1950) model (i.e., given an item ordering according to increasing item difficulty all 1s are to the left of all 0s) and let $\mathbf{X}_i'$ be the reversed Guttman vector (all 1s to the right of all 0s). *U3* can be written as

$$U3_i = \frac{\ln P(X^*_i) - \ln P(X_i)}{\ln P(X^*_i) - \ln P(X'_i)} . \qquad (1)$$

Van der Flier (1980) showed that a standardized version of *U3*, denoted *ZU3*, was approximately standard normally distributed given an invariant ordering of the persons according to their ability level. The advantage of *ZU3* to *U3* is that it

has theory-based significance levels and that it is based on less restrictive assumptions than *U3* (i.e., it is not assumed that the item ordering is invariant across Θ).

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 (Van der Flier, 1982; Meijer, 1994). Therefore, we will use *ZU3* in this study.

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 the dataset. 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.

Study 1

1. Datasets of 2000 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 $a \sim U[.5,1.5]$; item difficulties (b) range 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 $c \sim U[0,.2]$. .

2. Two datasets of 2000 NRVs were simulated. The same types of NRVs were used as in Meijer et al. (in press). 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. Item scores on these items were scored as correctly answered. It was assumed that cheaters had correctly answered these items by looking at a more able examinee. The second dataset consisted of guessers who answered all items with a probability of a correct answer of .25 which corresponds to answering an item with four alternatives by randomly guessing. In practical test situations these persons may make the exam just to get familiar with the type of items, without being properly prepared for it.

3. From each dataset of 2000 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 in (1). Consequently, for both guessers and cheaters four datasets were created with 5%, 10%, 15%, and 20% NRVs. In a similar way, four datasets were created with two types of NRVs by replacing half of the 5% (and 10%, 15%, 20%, respectively) FRVs by cheaters and half of them by guessers. The proportion correct score on each item g, π_g, ($g = 1, ..., k$) was estimated in the eight datasets containing FRVs and NRVs and in a dataset containing only FRVs.

4. Two datasets with 2000 NRVs were created according to the same procedure as in (2). For each simulee in these datasets, *ZU3* values were calculated using the π_g values estimated in (3). Item score patterns with *ZU3*>1.96 were classified as NRVs.

Note that we do not use datasets with both FRVs and NRVs in (4) to investigate the power of *ZU3*. This was done to avoid the risk that the power of *ZU3* was confounded by the unequal base rate (i.e., the unequal proportion NRVs) in the samples. Meijer et al. (in press) showed that the larger the base rate, the easier it is to classify a guesser or a cheater as a NRV.

Study 2

1. One dataset of 2000 simulees was generated with 20% cheaters and one dataset was generated with 20% guessers according to the same procedure as in (3) in Study 1. Furthermore, a dataset of 2000 simulees was generated with 10% cheaters and 10% guessers.

2. π_g values were estimated in the datasets.

3. For each dataset, *ZU3* values were calculated and simulees with *ZU3*>1.96 were classified as aberrant and were deleted.

4. Step (2) and (3) were repeated until no clear improvement of the power of *ZU3* was found.

Results

Study 1

Table 1 shows the percentages of cheaters and guessers correctly classified as NRV (VNRVs). Both for k=17 and k=33 it can be seen that if the percentage NRVs in the sample increased the power of *ZU3* decreased. Furthermore, for both cheaters and guessers for a fixed percentage of NRVs, the percentage of VNRVs was larger for k=33 than for k=17. This is in agreement

with the findings by Meijer et al. (in press) who found that NRVs were easier to detect for longer tests. Note that the presence of 5% NRVs already reduced the percentages VNRVs with 10% (k=17, guessing) to 15% (k=17, cheating). With 20% NRVs, the reduction was between 36% (k=17, guessing; k=17, cheating) and 49% (k=33, guessing).

Insert Table 1 about here

If there were no NRVs in the calibration sample, both for k=17 and k=33, cheaters were easier to detect than guessers. However, for k=17 and 5%, 10%, and 15% NRVs and k=33 and 10% NRVs the percentage of VNRVs was higher for guessers than for cheaters.

The reduced power in the datasets that contained both FRVs and NRVs may be due to biased estimation of π_g. For example, due to cheating the item that was most difficult in a group of FRVs might be 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.

The Figures 1 and 2 show the bias of the $\hat{\pi}_g$ values for k=17, and cheaters and guessers, respectively. Figure 1 shows that the $\hat{\pi}_{g^S}$ of the three most difficult items on which was cheated become positively biased varying from approximately .04 with 5% NRVs in the dataset to approximately .16 with 20% NRVs in the dataset. Figure 2 shows that for 5% guessers almost all $\hat{\pi}_{g^S}$ were unbiased (bias $\leq$ .02). Furthermore, an increase in the percentage guessers resulted in an increase in the (negative) bias for the easiest items,

whereas the $\hat{\pi}_{g^S}$ of the more difficult items remain almost unbiased. The Figures 3 and 4 show the bias results for k=33 and cheaters and guessers. The trends are almost the same as for k=17 and will, therefore, not further be discussed.

Insert the Figures 1, 2, 3, and 4 about here

Table 2 shows the percentages of VNRVs for the case of both cheaters and guessers in the calibration sample. In almost all cells the percentages of VNRVs were higher compared with the percentages found for one type of NRVs in the calibration sample (Table 1). Although more $\hat{\pi}_{g^S}$ in the test became biased in comparison with the situation that the dataset contained only one type of NRVs, the bias was less high (cf. the Figures 1 and 2 with Figure 5, and the Figures 3 and 4 with Figure 6) which may explain the higher detection rate for the case with two types of NRVs.

Insert Table 2 and the Figures 5 and 6 about here

Study 2

Table 3 shows the percentages VNRVs after deleting NRVs with $ZU3$>1.96 in four subsequent iterations. For k=17, after one iteration the percentages VNRVs increased with 14% (cheaters) and 13% (guessers). The percentages VNRVs only slightly increased after the iterations 2, 3, and 4 and the percentage of VNRVs that was found using the $\hat{\pi}_g$ values estimated in the

dataset with FRVs in Study 1 (Table 1) was not reached. For k=33, iteration 1 also resulted in the largest increase in VNRVs (20% for cheaters and 22% for guessers). After iteration 3 the percentages VNRVs were only a little smaller than the percentages found using the $\hat{\pi}_g$ values obtained with FRVs (cf. third row of Table 1 with the last row of Table 3).

The bias reduction of the $\hat{\pi}_{g^S}$ followed the same trend as the percentage VNRVs; the largest reduction was found after the first iteration, whereas smaller reduction was found after the other iterations.

In general, both for k=17 and k=33 the Type I error after the second iteration was a little higher than for the first two iterations. For k=17, for iteration 2 it equalled 6%, whereas for the iterations 3 and 4 it equalled 6.5% and 7.1%. For k=33, for iteration 2 it equalled 6.8%, and for the iterations 3 and 4 it equalled 6.5 and 6.7%, respectively.

Insert the Tables 3 and 4 about here

Table 4 shows the percentages VNRVs using the $\hat{\pi}_g$ values estimated in the group with two types of NRVs. These percentages are approximately the same for cheaters and somewhat higher for guessers compared with those in Table 3. Obviously, the higher percentage guessers can be explained by the higher percentages VNRVs for guessing simulees in Table 2. The trends were further the same as in Table 3. The Type I error was of the same magnitude as for one type of NRVs in the dataset.

Discussion

Due to the presence of NRVs in a dataset, the power of *ZU3* may be seriously reduced. In general, the larger the percentage NRVs the smaller the power of *ZU3*. Even with a relatively small percentage of 5% NRVs in the dataset, the percentage VNRVs was 10 to 15% lower than with only FRVs. This is in agreement with the findings of Kogut (1987) who found that the power of *M* was seriously reduced if the *b*s were estimated in a dataset containing NRVs. The findings by Levine and Drasgow (1983) were not confirmed. A possible explanation is that we choose two relatively severe types of NRVs. For example, in their study guessing took place on 20% of the items, whereas in the present study guessing took place on all items.

The use of an iterative procedure to delete NRVs gave mixed results. 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 two or three iterations. Since at each iteration the number of FRVs that were classified as NRVs increased, the number of iterations should be kept to a minimum.

With respect to the presence of two types of aberrant item score patterns in one dataset it can be concluded that the rate of detection of both guessers and cheaters was approximately the same or somewhat higher than in the case with a single type of NRVs. The use of the iterative procedure yielded about the same results as with a single type of NRVs.

These results suggest that it is possible to remove NRVs iteratively by means of *ZU3* to eliminate NRVs from the dataset. It should be realized, however, 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 obtain some insight in the ability 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.

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). Appropiateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

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 (Eds.), *Measurement and prediction* (pp. 60-90). Princeton: Princeton University Press.

Kogut, J. (1987). *Detecting aberrant response patterns in the Rasch model.* (Report 87-3). Enschede: University of Twente, Department of Education.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

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

Levine, M. V., & Drasgow, F. (1983). Appropriateness Measurement: validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 109-131). New York: Academic Press.

Meijer, R. R. (1994). Nonparametric and group-based person-fit statistics: a validity study and an empirical application. *Manuscript submitted for publication.*

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (in press). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement.*

Meijer, R. R., & Sijtsma, K. (in press). Detecting aberrant item score patterns: A review of recent developments. *Applied Measurement in Education.*

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* New York/Berlin: 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.

Sijtsma, K. & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52,* 79-97.

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.

Table 1

Percentage Cheaters and Guessers Classified as NRV

(One Type of NRV in the Calibration Sample)

	cheating		guessing	
%NRV	k=17	k=33	k=17	k=33
-	64	89	62	86
5	49	76	52	73
10	39	60	41	62
15	30	53	33	51
20	28	48	26	37

Table 2

Percentage Cheaters and Guessers Classified as NRV

(Two Types of NRVs in the Calibration Sample)

	cheating		guessing	
%NRV	k=17	k=33	k=17	k=33
5	53	81	56	78
10	45	70	49	71
15	35	60	41	59
20	26	49	32	49

Table 3

Percentage Cheaters and Guessers Classified as NRV as a Function of the Number of Iterations (One Type of NRV in the Calibration Sample)

	cheating		guessing	
iteration	k=17	k=33	k=17	k=33
- [a]	27	48	26	36
1	42	68	39	58
2	46	80	44	74
3	50	82	46	82
4	54	82	47	83

[a] percentage of VNRVs before the first iteration

Table 4

Percentage Cheaters and Guessers Classified
as NRV as a Function of the Number of Iterations
(Two Types of NRVs in the Calibration Sample)

	cheating		guessing	
iteration	k=17	k=33	k=17	k=33
- [a]	27	50	32	49
1	40	69	45	69
2	44	81	49	72
3	49	83	50	78
4	51	84	53	80

[a] percentage of VNRVs before the first iteration

Figure Captions

Figure 1. Bias Results for the 17-Item Test and Cheaters

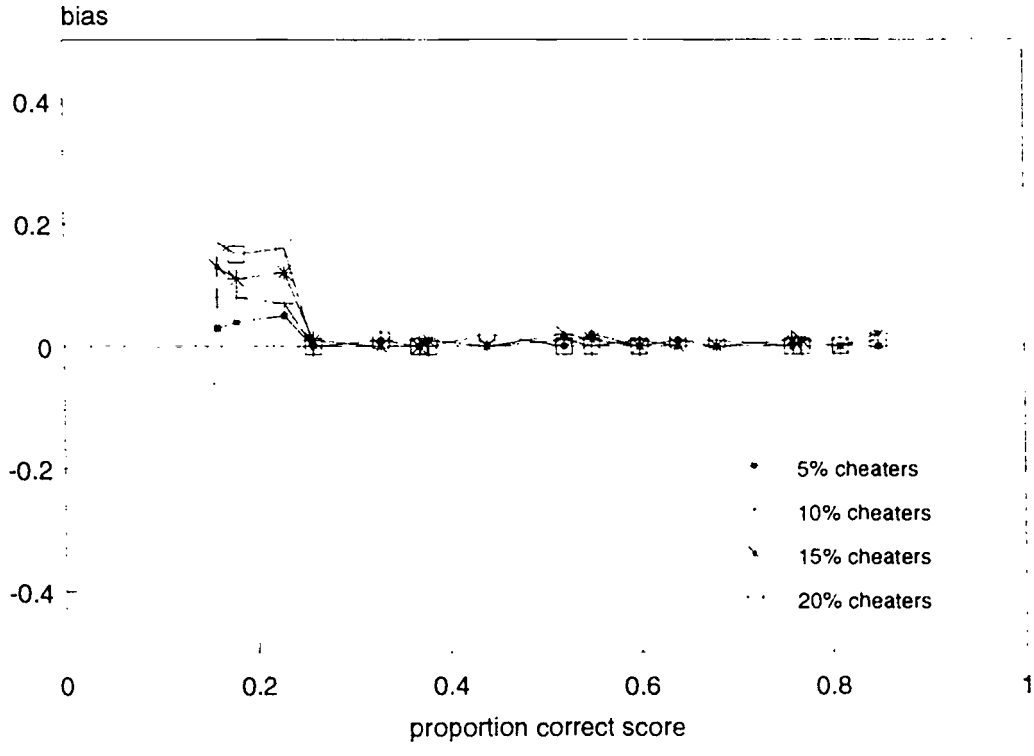
Figure 2. Bias Results for the 17-Item Test and Guessers

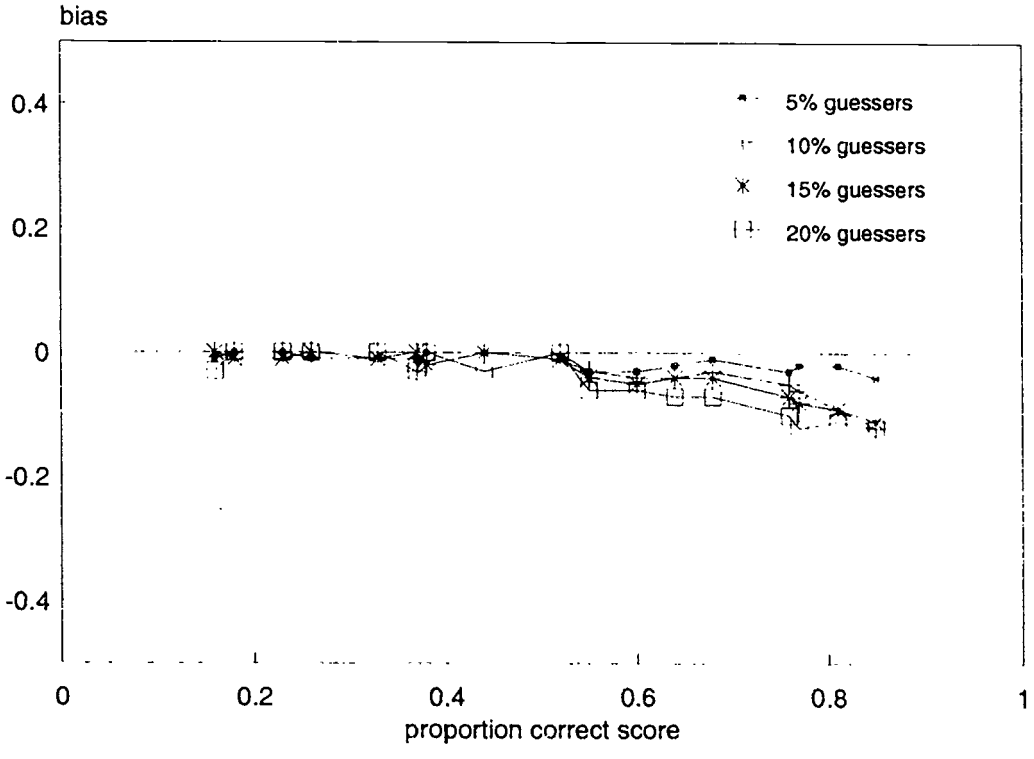
Figure 3. Bias Results for the 33-Item Test and Cheaters

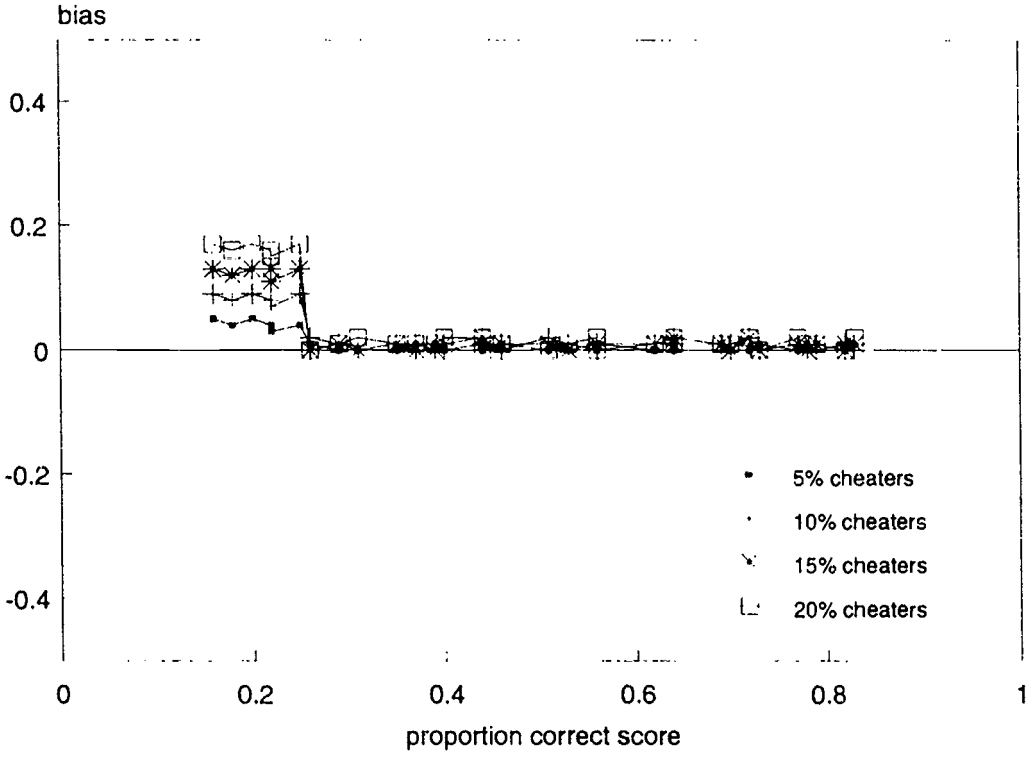
Figure 4. Bias Results for the 33-Item Test and Guessers

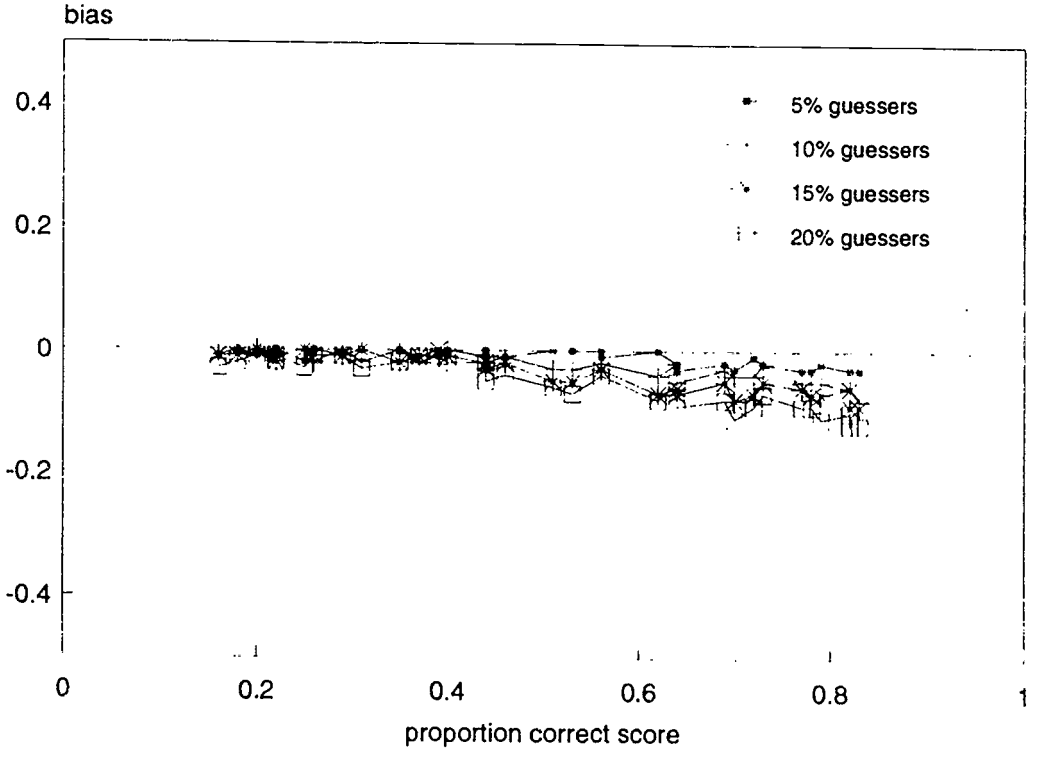
Figure 5. Bias Results for the 17-Item Test and Cheaters and Guessers

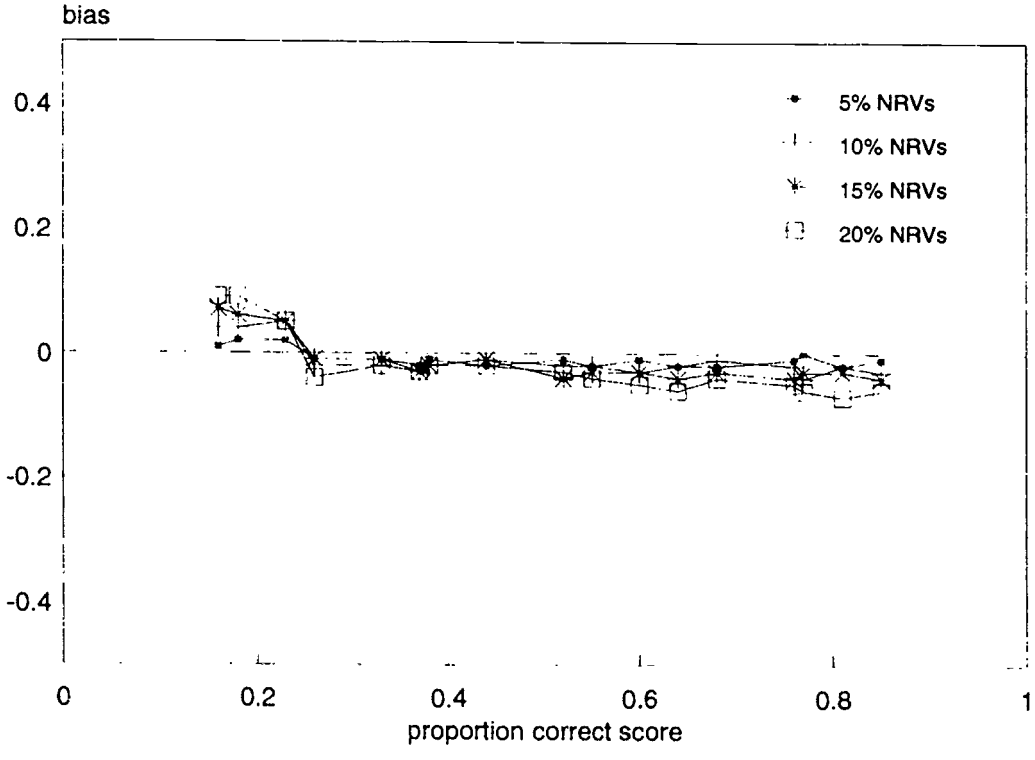
Figure 6. Bias Results for the 33-Item Test and Cheaters and Guessers

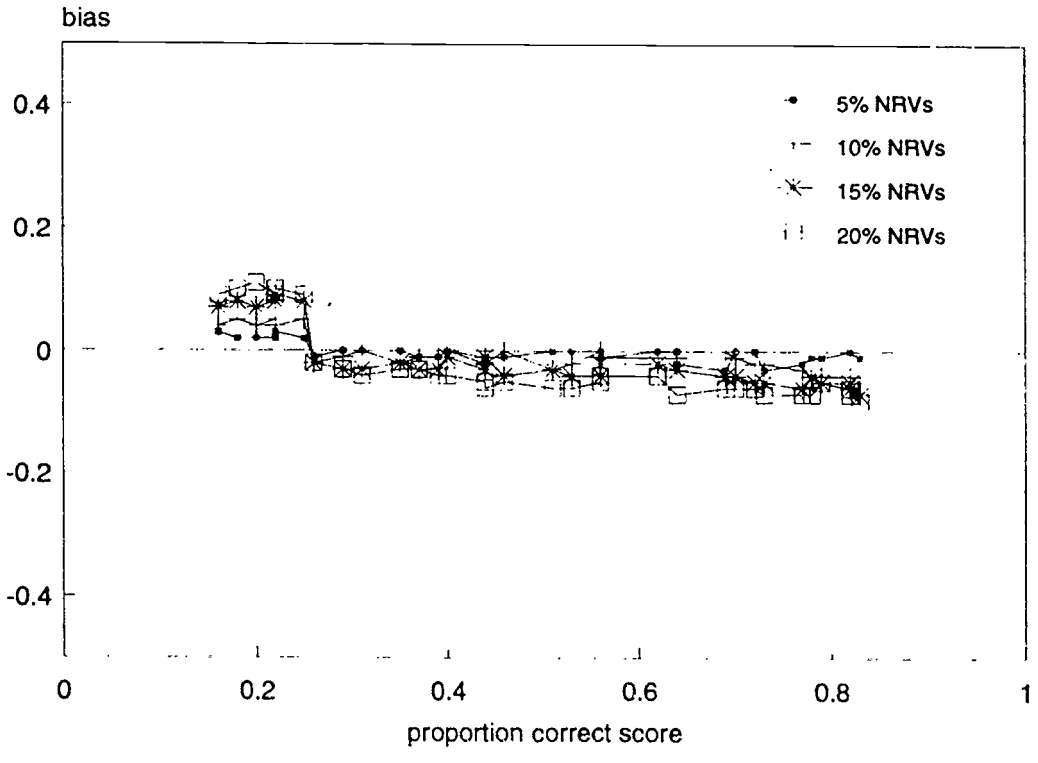

bias
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
proportion correct score
5% cheaters
10% cheaters
15% cheaters
20% cheaters


bias
0.4
0.2
0
-0.2
-0.4
5% guessers
10% guessers
15% guessers
20% guessers
0
0.2
0.4
0.6
0.8
1
proportion correct score


bias
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
proportion correct score
5% cheaters
10% cheaters
15% cheaters
20% cheaters


bias
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
proportion correct score
5% guessers
10% guessers
15% guessers
20% guessers


bias
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
proportion correct score
5% NRVs
10% NRVs
15% NRVs
20% NRVs


bias
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
proportion correct score
5% NRVs
10% NRVs
15% NRVs
20% NRVs

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

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

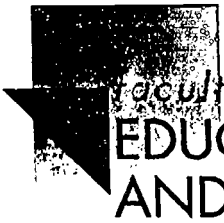
RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

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



faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands