The Number of Guttman Errors as a Simple and Powerful Person-Fit Statistic

Rob R. Meijer

University of Twente

A number of studies have examined the power of several statistics that can be used to detect examinees with unexpected (nonfitting) item score patterns, or to determine person fit. This study compared the power of the U3 statistic with the power of one of the simplest person-fit statistics, the sum of the number of Guttman errors. In most cases studied, (a weighted version of) the latter statistic performed as well as the U3 statistic. Counting the number of Guttman errors seems to be a useful and simple alternative to more complex statistics for determining person fit. *Index terms: aberrance detection, appropriateness measurement, Guttman errors, nonparametric item response theory, person fit.*

A number of studies have investigated statistics that can be used to detect examinees with unexpected item score patterns on the basis of an item response theory (IRT) model (e.g., Drasgow, Levine, & McLaughlin, 1987; Levine & Rubin, 1979; Molenaar & Hoijtink, 1990; Tatsuoka, 1984) or in comparison to other persons (Harnisch & Linn, 1981; Miller, 1986; van der Flier, 1982). These person-fit statistics can be used to detect several kinds of nonfitting response behavior such as guessing, cheating, or extremely creative behavior. For a review and applications of person-fit statistics, see Hulin, Drasgow, & Parsons (1983, chap. 4) and Meijer & Sijtsma (1995). Most of the person-fit statistics have been developed for dichotomous item scores, in which a 1 indicates a correct or keyed response and a 0 indicates an incorrect or not keyed response.

Several studies have compared the power of person-fit statistics for detecting nonfitting item response patterns. For example, Drasgow et al. (1987) compared the power of several statistics that can be used if a parametric IRT model applies to the data; Harnisch & Linn (1981) compared several group-based statistics and concluded that their "modified caution index" was superior to the other statistics because it had the lowest correlation with the number-correct score.

This study compared the power of three statistics that can be used in a nonparametric IRT context. Two forms of one of the simplest person-fit statistics, the number of Guttman errors (i.e., the number of item pairs with a 0 on the easier item and a 1 on the more difficult item), were compared with the more complex U3 statistic (van der Flier, 1982). U3 was defined in a nonparametric IRT context and assumes invariant item ordering across the latent attribute scale. U3 was used because it has proven to be a useful statistic under varying test conditions in simulation and empirical research (Meijer, Molenaar, & Sijtsma, 1994; van der Flier, 1982).

Using simulated data, Meijer et al. (1994) investigated the power of U3 across varying test and person conditions. The percentage of simulees defined a priori as nonfitting responders that were detected increased with test length and the ratio of nonfitting to model-fitting simulees in the group. Furthermore, "cheating" simulees were easier to detect than "guessing" simulees. That study was extended here by comparing the power of U3 with the power of the (weighted) number of Guttman errors.

Method

Data matrices of 450 simulees were generated using the two-parameter logistic model [see Meijer et al. (1994) for details of the data simulation]. A

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 4, December 1994, pp. 311–314

0146-6216/94/040311-04$1.45

completely crossed $4 \times 2 \times 2 \times 2$ design was considered with: (1) four levels of uniform item discriminations ($a = .5$, 1.0, 2.0, and 5.0) for all k items; (2) two levels of test length ($k = 17$ and $k = 33$); (3) two levels of number of nonfitting response vectors (NNRV = 50 and NNRV = 25; 11% and 5.5% nonfitting simulees, respectively); and (4) two types of misfit ("cheating" and "guessing"). Cheating simulees answered most items according to the two-parameter logistic model, except for the three most difficult items in the 17-item test and the six most difficult items in the 33-item test. For these items, 0 scores were changed to 1 scores. Guessing simulees answered all items by blindly guessing the correct answer with a probability of .25.

The item difficulties from the two-parameter logistic model were equidistant with a distance of .25 for the 17-item test and .125 for the 33-item test. For each test, the median difficulty was 0. The latent trait was drawn from a standard normal distribution. For a more detailed description of the generating procedure and the rationale behind this design see Meijer et al. (1994).

The following person-fit statistics were used to detect nonfitting simulees: (1) the number of Guttman errors (G), (2) the number of Guttman errors divided by the maximum number of Guttman errors given the number-correct score (G^*), and (3) the U3 statistic. G^* was defined because the range of G depends on the number of items in the test and the number-correct score of a person; hence, G may be confounded with the number-correct score. To reduce this possible confounding, G was divided by the maximum number of Guttman errors given the number-correct score. For a person with a number-correct score $X = r$, this maximum equals $r(k - r)$.

Let π_g ($g = 1, ..., k$) denote the proportion of persons who respond correctly to item g. Furthermore, assume that k items in a test are numbered and ordered such that $\pi_g \geq \pi_h$ ($g = 1, ..., k-1$; $h = g + 1, ..., k$). Then G can be written as

$$G = \sum_{g=1}^{k-1} \sum_{h=g+1}^{k} f_{gh}, \tag{1}$$

where $f_{gh} = 1$ if a person has a Guttman error on items g and h and $f_{gh} = 0$ otherwise.

G^* can be written as

$$G^* = \frac{\sum_{g=1}^{k-1} \sum_{h=g+1}^{k} f_{gh}}{r(k-r)}. \tag{2}$$

Let $\mathbf{X}$ be a vector containing the observed binary item scores of a person, and let $\mathbf{X}^*$ be a vector of a person with a number-correct score $X = r$, with 1s in the first r positions and 0s in the last $k - r$ positions. This vector is called a Guttman vector because it fits the Guttman (1950) model. Let $\mathbf{X}'$ denote the item score vector with 0s in the first $k - r$ positions and 1s in the last r positions. Given $X = r$ it is the vector with the maximum number of Guttman errors, which is called a reversed Guttman vector. Finally, let P, in general, denote a probability, and let $P(\mathbf{X})$ denote the probability of an item score pattern $\mathbf{X}$ conditional on the number-correct score. U3 is defined as

$$\mathrm{U3} = \frac{\ln P(\mathbf{X}^*) - \ln P(\mathbf{X})}{\ln P(\mathbf{X}^*) - \ln P(\mathbf{X}')}. \tag{3}$$

U3 ranges from 0 to 1, where 0 indicates that the observed item score pattern is a Guttman vector, and 1 indicates that the observed pattern is a reversed Guttman vector. Increasing values indicate that patterns are further removed from perfect Guttman patterns.

With respect to the rate of detection within each cell, the percentage of a priori defined nonfitting simulees (valid nonfitting response vectors, VNRVs) successfully detected using G, G^*, and U3 was determined. This was done by first ordering all 450 simulees according to increasing G, G^*, and U3 values. Second, the percentages of a priori defined nonfitting simulees among the NNRV simulees with the highest G, G^*, and U3 values were determined.

Note that the percentage of normal persons that were incorrectly classified as nonfitting (false nonfitting response vectors, FNRV) and the percentage of nonfitting persons that were incorrectly classified as model-fitting (false model-fitting response vectors, FFRV) can be calculated easily if the number of VNRVs is known. If the percentage of VNRVs

increases, the percentages of FNRVs and FFRVs decrease. Therefore these percentages are not discussed further.

To investigate whether the number-correct score and the person-fit statistics were confounded, product-moment correlations were calculated between the number-correct score and each person-fit statistic.

Results

The percentages of VNRVs detected within each cell in the design are shown in Table 1. For G, G^*, and U3 an increase in a resulted in an increase in the rate of detection of both cheating and guessing simulees. Given a fixed a and test length, for NNRV = 50 the rate of detection for each statistic was at least as large in the cheating condition as in the guessing condition (e.g., for $a = 1$ and $k = 17$ using G, the percentage of VNRVs was 80% for cheating and 50% for guessing; using G^* these percentages were 70% and 62%; and using U3 these percentages were 68% and 62%, respectively). These trends held for NNRV = 25, with the exception of one cell ($a = .5$, $k = 17$, and G^* where it was 32% for cheating and 40% for guessing).

For a fixed a and cheating simulees, the rate of detection for G, G^*, and U3 was almost always higher for NNRV = 50 than for NNRV = 25 [similar results were obtained by Meijer et al. (1994) using U3]. For a fixed a and guessing simulees, this trend was weaker for all three statistics. Furthermore, Table 1 shows that an increase in the number of items in general resulted in an increase in the rate of detection for both cheating and guessing simulees.

Comparing the power (i.e., the percentage of VNRVs detected) of G, G^*, and U3 for all 32 conditions, in general the power of G, G^*, and U3 was similar in identical conditions. No statistic had the highest power across all cells. For $a = .5$, $a = 1$, and guessing simulees, G had less power than G^* and U3 (e.g., for NNRV = 50, $a = .5$, $k = 17$, and guessing, the percentage of VNRVs was 26% for G, 40% for G^*, and 42% for U3; for NNRV = 50, $a = 1$, $k = 17$, and guessing these percentages were 50%, 62%, and 62%, respectively). U3 was somewhat more effective than G and G^* for $a = .5$, $k = 17$, and cheating simulees (e.g., for NNRV = 50, $a = .5$, $k = 17$, and cheating, the percentage of VNRVs was 52% for U3, 48% for G, and 50% for G^*; for NNRV = 25 these percentages were 36% for U3 and 32% for both G and G^*); for $a = 2$, $k = 33$, NNRV = 50, and cheating simulees; and for $a = .5$, $k = 17$, NNRV = 50, and guessing simulees. However, the differences in power were small.

Table 2 shows the mean, SD, and the range of the product-moment correlations of G, G^*, and U3 with the number-correct score in both the guessing and cheating conditions. The mean correlation was approximately 0.0 in the cheating condition and approximately −.25 in the guessing condition.

Table 1
Percentage of Nonfitting Simulees Classified as Nonfitting (VNRVs) for G, G^*, and U3 Statistics, for Cheating and Guessing Conditions

	Cheating						Guessing					
NNRV	$k = 17$			$k = 33$			$k = 17$			$k = 33$		
and a	G	G^*	U3	G	G^*	U3	G	G^*	U3	G	G^*	U3
NNRV = 50												
.5	48	50	52	78	74	76	26	40	42	24	52	52
1.0	80	70	68	94	92	92	50	62	62	54	78	76
2.0	96	92	90	96	96	98	78	82	82	92	96	96
5.0	100	100	100	100	100	100	98	98	96	100	100	100
NNRV = 25												
.5	32	32	36	56	56	56	24	40	36	12	40	40
1.0	72	72	64	92	92	92	36	48	48	52	80	80
2.0	92	84	84	96	96	92	72	76	84	88	96	92
5.0	100	100	100	100	100	100	100	100	96	100	100	100

Table 2
Mean, SD, and Range of the Product-Moment Correlations of *G*, *G**, and U3 With the Number-Correct Score for Cheating and Guessing Conditions

	Cheating			Guessing		
Statistic	*G*	*G**	U3	*G*	*G**	U3
Mean	.02	.04	.04	−.26	−.27	−.25
SD	.03	.03	.04	.03	.03	.03
Range	−.07	−.02	−.05	−.33	−.34	−.32
	.08	.09	.12	−.18	−.18	−.19

Discussion

For most datasets simulated here, a simple count of the number of Guttman errors (*G*), or a simple count of the number of Guttman errors normed against the maximum number of Guttman errors given the number-correct score (*G**), were useful alternatives to the person-fit statistic, U3. Furthermore, the three statistics were not highly correlated with the number-correct score. These results supplement the results obtained by Meijer et al. (1994).

The present results are similar to those obtained by Harnisch & Linn (1981). On the basis of two empirical datasets, they preferred their modified caution index because it correlated −.02 and −.21 with the number-correct score in the two datasets. For the simulated datasets used here, the correlations with the number-correct score were of approximately the same magnitude. Thus, using this criterion to evaluate a person-fit statistic, counting the number of Guttman errors seems a useful and simple alternative to using U3 or the Harnisch and Linn caution index.

References

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59–79.

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 NJ: Princeton University Press.

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.

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

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Item, test, person, and group characteristics and their influence on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111–120.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8,* 261–272.

Miller, M. D. (1986). Time allocation and patterns of item response. *Journal of Educational Measurement, 23,* 147–156.

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

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13,* 267–298.

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.