KM 22(1986)
pag 131-145

A coefficient of deviance of response patterns

K. Sijtsma

Summary

In the present paper a coefficient is discussed which expresses the degree to which a person's response pattern on a set of test items deviates from other person's response patterns. Some results based on simulated data are presented, which show that the coefficient is rather successful with respect to detecting deviating response patterns.

Introduction

Besides summary statistics on a test, e.g., the raw score, item response patterns of individual subjects may be used in order to obtain information for diagnostic use of tests. Summary statistics give an impression on the overall level of ability or achievement at which a subject operates. Once information has been collected with respect to the degree to which an individual's response pattern deviates from some criterion, more detailed conclusions can be drawn with respect to that individual. Such conclusions entail, e.g., membership in a specific population (e.g., Van der Flier, 1982), deficiency of certain abilities needed to solve particular classes of problems that are represented in the test (e.g., Tatsuoka and Tatsuoka, 1983; Harnisch, 1983), and cheating or guessing on educational tests (e.g.

Vakgroep Arbeids- en Organisatiepsychologie, Vrije Universiteit, De Boelelaan 1081, Postbus 7161, 1081 HV Amsterdam. Tel. 020 - 5485517.

The author is grateful to Charles Lewis and Ivo W. Molenaar for their critical comments on an earlier draft of this paper.

Levine and Rubin, 1979). Consequently, subjects may be allocated to other groups, they may be given additional instruction, or their test performance may be reconsidered.

Early attempts at analysis of response patterns on items or tests have been made by, e.g., Du Mas (1946), Osgood and Suci (1952), Gaier and Lee (1953) and Cronbach and Gleser (1953) in the context of profile analysis. In the recent years two approaches to the analysis of item response patterns have developed.

One approach evaluates the likelihood of a response pattern given a probabilistic model (Wright and Stone, 1979; Levine and Rubin, 1979; Levine and Drasgow, 1982; Drasgow, 1982). In order to carry out this evaluation, estimates of the latent parameters of an item response model are needed. Tatsuoka and Linn (1983) and Tatsuoka (1984) propose coefficients which express the deviance of a binary-valued response vector from the corresponding vector containing an individual's success probabilities on the items in the test. In order to estimate these coefficients, one also needs the estimates of latent subject and item parameters, and consequently the test should conform to a parametric item response model.

The other approach evaluates an array of observed dichotomously scored item responses against statistics which are usually based on the group to which the individual belongs (Kane and Brennan, 1980; Van der Flier, 1980; 1982; Harnisch and Linn, 1981; Tatsuoka and Tatsuoka, 1982; Tatsuoka and Linn, 1983; Tatsuoka, 1984). In many cases, the latter approach concentrates on the fit of an individual subject's response pattern to the well-known deterministic Guttman model (1945; 1950).

In this paper we study deviating response patterns in the context of the second approach mentioned, i.e., deviation from ideal response patterns which conform to the Guttman model. The present study follows the tradition of the general nonparametric item response theory proposed by Mokken (1971; also, Mokken and Lewis, 1982). Mokken's approach concentrates on item analysis, and proposes several item coefficients and procedures of item selection. The role of both perfect and imperfect Guttman response patterns in nonparametric item response theory is discussed extensively (Mokken, 1971, p. 153-157).

No detailed analysis is made, however, of the problem of diagnosing individual response patterns. This study attempts to fill this gap in Mokken's development.

A nonparametric item response model

Mokken (1971, p. 115-169) has proposed a probabilistic approach for scaling persons and items on a single dimension. Actually, his approach to scaling entails two different nonparametric item response theories for dichotomously scored items. The models are called nonparametric because the item characteristic curves (ICC's) and the distribution of the latent attribute ξ are not specified in terms of a parametric family of functions. In the present paper, only the most general of his item response theories is presented, where the generality refers to the restrictions placed on the ICC's. We use this presentation to define some quantities to be used in this paper. The general model is based on three assumptions:

1. Measurement by means of the items in the test is unidimensional;
2. Responses to different items made by a single subject are stochastically independent given the position of the subject on the attribute ξ;
3. ICC's are monotonically nondecreasing in the latent attribute.

In order to explain the third assumption we need the success probability π of a subject, say v, on an item, say i, and denote it by π_{vi}. The third assumption means that for any two subjects, v and w, if there exists an item i for which $\pi_{vi} < \pi_{wi}$ then $\pi_{vj} \leq \pi_{wj}$ for all $j = 1, \ldots, k$, where k is the number of items in the test. Items conforming to this model are called monotonely homogeneous, and their ICC's may intersect. The third assumption implies that each item (partially) orders subjects in the same way.

It should be noted that π_{vi} is the expectation across independent replications of the binary item score X_{vi}. Let the observed raw scores be denoted by X_v and X_w respectively, then

$$\varepsilon(X_v/k|\xi_v) = 1/k \sum_{i=1}^{k} \pi_{vi} = \beta_v \leq$$

$$\varepsilon(X_w/k|\xi_w) = 1/k \sum_{i=1}^{k} \pi_{wi} = \beta_w , \quad (1)$$

using the third assumption. In random samples, β_v and β_w are estimated by their unbiased estimators (Mokken, 1971, p. 129)

$$\hat{\beta}_v = X_v/k \text{ and } \hat{\beta}_w = X_w/k . \quad (2)$$

A coefficient of subject scalability

Mokken (1971, p. 148) uses Loevinger's (1948) scalability coefficient H to assess the degree to which a data set conforms to the Guttman model. Since this coefficient concentrates on the scalability of items, no conclusions can be drawn with respect to the scalability of individual persons. In the present paper, we define H for two persons v and w, and arbitrarily assume that v is the less able person, and that his/her ability level is denoted by β_v so that $\beta_v < \beta_w$ under the monotonely homogeneous model. Furthermore, we let $\beta_{vw}(1, 0)$ denote the expected proportion of items answered positively by the less able person v and negatively by the more able person w, and $\beta^{(0)}_{vw}(1, 0)$ denote the error proportion to be expected if the response patterns of v and w were independent across items: $\beta^{(0)}_{vw}(1, 0) = \beta_v(1 - \beta_w)$. This would mean that the response patterns of two fixed persons v and w correlate zero, given their proportions of positive responses β_v and β_w. The error proportion $\beta_{vw}(1, 0)$ is estimated by the unbiased and consistent estimator $\hat{\beta}_{vw}(1, 0) = X_{vw}(1, 0)/k$, where $X_{vw}(1, 0)$ denotes the observed error frequency. The scalability coefficient for two persons is defined as

$$H_{vw} = 1 - \beta_{vw}(1, 0)/\beta^{(0)}_{vw}(1, 0) . \quad (3)$$

Of more interest is the scalability coefficient of one person v, H_v, with respect to the population to which he/she belongs. It should be realized that when $\beta_v > \beta_w$, the expected error proportion equals $\beta_{vw}(0, 1)$. When $\beta_v = \beta_w$, it is readily seen that $\beta_{vw}(1, 0) = \beta_{vw}(0,1)$, and there is no problem choosing an error proportion. For v, w = 1, ..., n, the person coefficient is defined as

$$H_v = 1 - \left[\sum_{w<v} \beta_{vw}(0, 1) + \sum_{v<w} \beta_{vw}(1, 0) \right] / \left[\sum_{w<v} \beta_{vw}^{(0)}(0, 1) + \sum_{v<w} \beta_{vw}^{(0)}(1, 0) \right] , \quad (4)$$

where by definition w < v implies $\beta_w < \beta_v$ and v < w implies $\beta_v \leq \beta_w$.

Van der Flier (1980, p. 43) presents H_v in an equivalent form, which we have adapted to the present notation:

$$H_v = \sum_{w \neq v} \left[\beta_{vw}(1, 1) - \beta_v \beta_w \right] / \left[\sum_{v<w} \beta_v (1 - \beta_w) + \sum_{w<v} \beta_w (1 - \beta_v) \right] , \quad (5)$$

where the numerator consists of the sum of the covariances between the response patterns of the subjects, and the denominator consists of the sum of the corresponding maximum covariances for fixed β_v (v = 1, ..., n). Tatsuoka and Linn (1983) discuss Sato's Caution Index, which is rather similar to H_v.

It can be seen in formula (4) that H_v equals one when the error proportions of person v all equal zero. It equals zero when the average covariance of the response pattern of v with the other response patterns equals zero (formula (5)). Negative values indicate negative correlations with one or more other response patterns.

Finally, it may be noted that the coefficients H_{vw} and H_v can be used without the assumption of monotone homogeneity since they can be interpreted as normed average covariances.

Recently, Jansen (1982; also, see Roskam, Van den Wollenberg & Jansen, 1986) has criticized the use of the item coefficient H in the context of the Mokken model. Others (e.g., Molenaar, 1982; Mokken, Lewis & Sijtsma, 1986) have critically responded to these criticisms. In this paper, we do not repeat the positions taken in this discussion. We only recall that Jansen (1982) has convincingly shown that H is not suited for the selection of items which conform to one of Mokken's item response models, from a larger set also containing items which violate these models. Sijtsma and Prins (1986) have recently investigated how H can be used properly to select items within the context of Mokken's approach.

Since the person coefficient H_v is used to detect deviating response patterns, but not to select sets of persons complying with some probabilistic test model, we think that the criticisms of H do not apply to H_v.

Properties of the person coefficient

It is well-known that the Guttman model implies perfect reproducibility of response patterns conditional upon the raw scores of persons. This means that given X_v, the X_v easiest items have been answered positively and the remaining items negatively. When $H_v = 1$ for some person v, it can be shown that this person's response pattern is perfectly reproducible. For this purpose, we consider a datamatrix (Figure 1(a)) containing the score pattern of person v. At the left of person v we order the persons for whom $\beta_w < \beta_v$, and at the right the persons characterized by $\beta_v \leq \beta_w$. The ordering of the items according to their difficulty is assumed to be unknown. We know, however, that $H_v = 1$, meaning that person v does not share error patterns with the other persons. This knowledge determines the configuration of the data matrix as indicated in Figure 1(a). The other elements can be chosen given the constraints imposed by the β_w ($w = 1, \ldots, n; w \neq v$). It is clear, however, that no matter which choice we make, items 2 and 4 will always be the easiest, implying that perfect reproducibility holds for person v.

Although $H_v = 1$ constitutes a sufficient condition for perfect reproducibility, it can be shown that it is not necessary. In Figure 1(b) a complete datamatrix is given. It is readily verified that $H_4 < 1$, although perfect reproducibility holds for the score pattern of person 4.

persons

items		w < v				v	v < w	
	1	0	0	0	0	0		
	2					1	1	1
items	3	0	0	0	0	0		
	4					1	1	1
		$\beta_w < \beta_v$				β_v	$\beta_v \leq \beta_w$	

Figure 1(a): Data matrix illustrating sufficiency of $H_v = 1$ for perfect reproducibility.

persons

		1	2	3	4	5	6	7	
	1	1	0	1	1	1	1	1	6
	2	0	1	1	1	1	1	0	5
items	3	0	0	0	0	1	1	1	3
	4	0	0	0	0	0	0	1	1
		1	1	2	2	3	3	3	

Figure 1(b): Data matrix illustrating that $H_v = 1$ is not necessary for perfect reproducibility.

Besides H_{vw} and H_v, an overall coefficient can be defined which expresses the global scalability of the group of persons with respect to each other:

$$H^{(P)} = \sum_{v<w}\sum \left[\beta_{vw}(1, 1) - \beta_v\beta_w\right] / \sum_{v<w}\sum \beta_v(1 - \beta_w) \ . \qquad (6)$$

This coefficient is the person counterpart of the overall scalability coefficient H of Loevinger (1948; see also Mokken, 1971, p. 148-153), which is defined as:

$$H = \sum_{i<j}\sum \left[\pi_{ij}(1, 1) - \pi_i\pi_j\right] / \sum_{i<j}\sum \pi_i(1 - \pi_j) \ , \qquad (7)$$

where $\pi_{ij}(1, 1)$ denotes the expected proportion of persons responding positively to both items i and j, and $\pi_i \leq \pi_j$ (i, j = 1, ..., k).

For given β_v and β_w (v, w = 1, ..., n), we define the constant

$$C_{vw} = \begin{cases} \beta_v(1 - \beta_w) \text{ if } \beta_v \leq \beta_w ; \\ (1 - \beta_v)\beta_w \text{ if } \beta_v > \beta_w . \end{cases}$$

Using this definition, we rewrite (6) as

$$H^{(P)} = \frac{\sum_v \left[\sum_{w\neq v} C_{vw} * \sum_{w\neq v} \{\beta_{vw}(1, 1) - \beta_v\beta_w\} / \sum_{w\neq v} C_{vw} \right]}{\sum_v \sum_{w\neq v} C_{vw}} =$$

$$= \frac{\sum_v H_v \sum_{w\neq v} C_{vw}}{\sum_v \sum_{w\neq v} C_{vw}} . \tag{8}$$

Several properties of $H^{(P)}$ can be deduced:

Property 1:

$$H^{(P)} \geq \min(H_v) \text{ , } v = 1, \ldots, n . \tag{9}$$

This property says that $H^{(P)}$ is always at least as large as the smallest of the individual person coefficients, which is readily seen in (8).

Property 2:

$$H^{(P)} = 1 \text{ iff } H_v = 1 \text{ , for all } v = 1, \ldots, n . \tag{10}$$

Necessity follows from the fact that $H^{(P)} = 1$ implies that there are no error patterns for any pair of persons. Sufficiency immediately follows from (8).

Property 3:

$$H^{(P)} = 1 \text{ iff Guttman-homogeneity} . \tag{11}$$

When $H^{(P)} = 1$, the persons can be (partially) ordered without any errors. It is readily seen that at the same time the items are also (partially) ordered without any errors.
Sufficiency follows by definition.

Corollary: From properties 2 and 3 it follows that

$$H_v = 1, \text{ for all } v = 1, \ldots, n,$$

$$\text{iff } \left\{ \begin{array}{l} X_{vi} = 1,\ i = 1, \ldots, X_v, \text{ and} \\ \\ X_{vi} = 0,\ i = X_v + 1, \ldots, k\ . \end{array} \right\} \text{ for all } v = 1, \ldots, n\ . \quad (12)$$

Thus, when all person coefficients are perfect the property of reproducibility holds. In this case also $H^{(P)} = H = 1$.

Finally, Mokken (1971, p. 157-169; see also Goodman, 1959) has derived the asymptotic sampling theory for the item coefficients H_{ij}, H_i and H. Since the person coefficients H_{vw}, H_v and $H^{(P)}$ are the analogues of the item coefficients when the data matrix is transposed, this sampling theory seems to be easily applied to the person coefficients as well. The approximation of the empirical to the theoretical sampling distribution will probably be problematic for at least two reasons. First, the number of items k is always small, and second, a test usually is deliberately constructed, and the items are (almost) never sampled randomly.

<u>Detectability of abberant patterns by means of H_v</u>

The effectiveness of H_v as a means for detecting abberant response patterns was studied for two different cases.

In the first case some persons in a larger sample are assumed to have guessed blindly for the correct answer on each item, while most persons have not. Several artificial data matrices were constructed, each containing response patterns which were either generated according to an item response model when persons had not guessed, or according to a constant success probability for each item when persons were assumed to have guessed. The simulated situation can be interpreted as one in which a group of students has studied a subject matter, and then takes an examination on the basis of the ability level acquired. Another group has not even looked at the books and consequently guesses blindly for each item with a constant success probability. The latter group may, e.g., be interested in collecting items for training for future examinations, and has no serious intention of passing the exam at the present administration.

In the second case, we consider a subgroup of respondents with ability parameters below the population mean: $\xi < \varepsilon(\xi) = 0$. During the exam, these respondents are assumed to have copied the correct answers to a few difficult items from their more able neighbours, or to have cheated in some other way. Thus, for a few difficult items $\pi_{vi} = 1$, and the success probabilities on the other items are determined by the item response model.

These cases are only two out of a large number of situations one might be interested in. Nevertheless, we hope that our cases shed some light on the usefulness of coefficient H_v.

Method

Subject parameters were randomly sampled from a standard normal distribution. In each cell of the design of this study, there are always 100 persons behaving according to the item response model, and the test always consists of nine items. The number of deviating persons is varied across three levels: 10, 20 and 40 persons, respectively. When no guessing or cheating is involved, item responses are simulated by means of the three-parameter logistic model:

$$\pi_{vi} = \gamma_i + (1 - \gamma_i) \exp[\alpha_i(\xi_v - \delta_i)] \;/\; \{1 + \exp[\alpha_i(\xi_v - \delta_i)]\} ,$$

where δ_i, α_i and γ_i denote the difficulty, discrimination and guessing parameters of item i, respectively.

In the case of guessing behaviour, tests consisting of two- and four-choice items are studied.

In the case of cheating on some items, the number of such items is varied: one and two items, respectively. These are always the most difficult items in the population where no cheating has taken place.

In the sample, the item scores on one or two items are all changed into ones for a subgroup of persons. As a consequence, when looking at sample statistics, these items often are no longer the most difficult items.

Within one data matrix guessing and discrimination ($\alpha_i = 2.0$ for each item in each cell) are always kept constant across items, whilst the difficulties are equidistant between -2.0 and 2.0. In the case of cheating $\gamma_i = .25$ for all items, except the items on which the cheating takes place. The data generation procedure has, e.g., been described in detail by Van den Wollenberg (1982).

Results

Persons are always a priori classified into n_D guessing and $n - n_D = 100$ non-guessing persons. Accordingly, the person coefficients are also classified into the n_D lowest and the $n - n_D$ highest coefficients. It may be noted that n and n_D change when $\beta_v = 0$ or 1 and consequently H_v is undefined. The resulting two-by-two cross-classification contains the false and valid positives and the false and valid negatives, respectively. The relations among these classifications are expressed by the phi-coefficient, which is always significantly positive (Table 1). This means that, generally speaking, the guessing persons are characterized by the lowest person coefficients, and the opposite is true for the other groups. It can also be seen in Table 1 that the mean person coefficient is always markedly lower among the guessing persons than among the non-guessing persons. Thus, the positive relations among the classifications are not caused by outliers.

Finally, although it is often significantly positive, the correlation between the raw score and the person coefficient is rather small, which means that the person coefficient adds information to the knowledge one has about persons based on their raw scores.

Table 1: Results with respect to the detectability of persons guessing blindly on each item.

		$n_D = 10$		$n_D = 20$		$n_D = 40$	
$\pi_{vi} = .25$	$\bar{H}_v$	.02	.45	.08	.40	.00	.37
	$\phi(D, H_v)$		.51*		.38*		.60*
	$r(X_v, H_v)$		.21*		.41*		.39*
$\pi_{vi} = .50$	$\bar{H}_v$	.02	.26	−.04	.19	.08	.20
	$\phi(D, H_v)$		.44*		.57*		.32*
	$r(X_v, H_v)$		.08		.22*		.33*

Note: n_D : number of deviating persons.

$\bar{H}_v$: mean of H_v in the deviating group (first entry) and the non-deviating group (second entry), respectively.

$\phi(D, H_v)$: phi-coefficient based on the cross-classification containing the frequencies of the false and valid positives and the false and valid negatives, respectively. An asterisk means significance at 5% level (one-sided).

$r(X_v, H_v)$: pm-correlation between the raw score and H_v. An asterisk means significance at 5% level (two sided).

Table 2: Results with respect to the detectability of persons cheating on a few difficult items.

		$n_D = 10$		$n_D = 20$		$n_D = 40$	
$k_c = 1$	$\bar{H}_v$	.09	.45	.20	.33	.17	.37
	$\phi(D, H_v)$		.56*		.27*		.31*
	$r(X_v, H_v)$		−.09		.27*		.21*
$k_c = 2$	$\bar{H}_v$	−.04	.39	−.01	.31	.23	.31
	$\phi(D, H_v)$		.45*		.69*		.11
	$r(X_v, H_v)$		.05		−.07		−.21*

Note: k_c : number of items on which cheating takes place.

Other notation is explained in Table 1.

Generally, the results for the cheating condition (Table 2) are closely comparable to the results for the guessing condition. In some cases the phi-coefficients are rather small. This can be attributed partly to the fact that the item(-s) on which cheating has taken place are not the most difficult ones in the sample. Another reason for small phi-coefficients is that behaviour is abberant on only one or two items.

A replication study shows that the results in the guessing condition may be considered valid, while the phi-coefficients in the cheating condition seem less stable. The general conclusions of this study are maintained, however.

Discussion

The results from our simulation study lead to the tentative conclusion that the coefficient H_v is rather successful in detecting abberant response patterns. An attractive feature of the coefficient is that it can be applied to data irrespective of the (non-) fit of an item response model.

A problem which was not solved in this paper, is where to put a borderline between deviant and non-deviant patterns. Van der Flier (1980) has presented a coefficient with a known sampling theory, making possible significance tests of the null hypothesis that a response pattern is not deviant. Such a significance test is certainly a useful tool, but it does not supply a proof that a person has guessed or cheated, or that something else has occurred which has caused the response pattern to be deviant. Therefore, additional evidence is needed in order to explain why a person should be considered aberrant. This remark applies to all formal methods to detect aberrant response patterns.

Compared to methods based on item response theory, the present method is attractive for its simplicity and independence of any underlying formal model. From the literature one can hardly draw any conclusions with respect to the relative effectiveness of different approaches. A comparative study on the effectiveness in detecting abberant patterns of different methods from different approaches would be highly interesting.

References

Cronbach, L.J., & Gleser, G.C. (1953). Assessing similarity between profiles. Psychological Bulletin, 50, 456-473.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.

Du Mass, F.M. (1946). A quick method of analyzing the similarity of profiles. Journal of Clinical Psychology, 2, 80-83.

Flier, H. van der (1980). Vergelijkbaarheid van individuele testprestaties. Lisse: Swets & Zeitlinger.

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

Gaier, E.L., & Lee, M.C. (1953). Pattern analysis: the configural approach to predictive measurement. Psychological Bulletin, 50, 140-148.

Goodman, L.A. (1959). Simple statistical methods for scalogram analysis. Psychometrika, 24, 29-43.

Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9, 139-150.

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. Princeton: Princeton University Press.

Harnisch, D.L. (1983). Item response patterns: applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Harnisch, D.L., & Linn, R.L. (1981). Analysis of item response patterns: Questionnable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

Jansen, P.G.W. (1982). De onbruikbaarheid van Mokkenschaalanalyse. Tijdschrift voor Onderwijsresearch, 7, 11-24.

Kane, M.T., & Brennan, R.L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, 4, 105-126.

Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-56.

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

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin*, *45*, 507-530.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Mokken, R.J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417-430.

Mokken, R.J., Lewis, C., & Sijtsma, K. (1986). The Mokken scale: A critical discussion: A rejoinder. *Applied Psychological Measurement*, *10*, (forthcoming).

Molenaar, I.W. (1982). De beperkte bruikbaarheid van Jansen's kritiek. *Tijdschrift voor Onderwijsresearch*, *7*, 25-30.

Osgood, C.E., & Suci, G.J. (1952). A measure of relation determined by both mean difference and profile information. *Psychological Bulletin*, *49*, 251-262.

Roskam, E.E., Wollenberg, A.L. van den, & Jansen, P.G.W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement*, *10*, (forthcoming).

Sijtsma, K., & Prins, P.M. (1986). Itemselectie in het Mokken model. *Tijdschrift voor Onderwijsresearch*, *11*, (ter perse).

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

Tatsuoka, K.K., & Linn, R.L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, *7*, 81-96.

Tatsuoka, K.K., & Tatsuoka, M.M. (1982). Detection of abberant response patterns. *Journal of Educational Statistics*, *7*, 215-231.

Tatsuoka, K.K., & Tatsuoka, M.M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 221-230.

Wollenberg, A.L. Van den (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123-140.

Wright, B.D., & Stone, M.H. (1979). *Best test design. Rasch measurement*. Chicago: MESA PRESS.