# Theoretical and Empirical Comparison of the Mokken and the Rasch Approach to IRT

**Rob R. Meijer and Klaas Sijtsma, Free University of Amsterdam**

**Nico G. Smid, Philips, Eindhoven, The Netherlands**

The Mokken model of monotone homogeneity, the Mokken model of double monotonicity, and the Rasch model are theoretically and empirically compared. These models are compared with respect to restrictiveness to empirical test data, properties of the scale, and accuracy of measurement. Application of goodness-of-fit procedures to empirical data largely confirmed the expected order of the models according to restrictiveness: Almost all items were in concordance with the model of monotone homogeneity, and fewer items complied with the model of double monotonicity and the Rasch model. The model of monotone homogeneity was found to be a suitable alternative to more restrictive models for basic testing applications; more sophisticated applications, such as equating and adaptive testing, appear to require the use of parametric models. *Index terms: goodness-of-fit, item response theory, measurement properties, Mokken model, Rasch model.*

In item response theory (IRT), test behavior is explained by taking into account attributes of both persons and items. An important notion in IRT is the item characteristic curve, or item response function (IRF). For dichotomously-scored items, it provides the probability of persons answering an item positively as a function of the latent trait. IRT test models are formulated so as to permit the derivation of consequences that can be checked empirically. Therefore, fit of the model to the data plays an important role in IRT.

In recent years much empirical research has used three IRT models: the Mokken models of monotone homogeneity and double monotonicity, and the Rasch model. See Mokken (1971), Stokman (1977), and Kingma and ten Vergert (1985) for empirical applications of the Mokken models. Empirical applications of the Rasch model can be found in van den Wollenberg (1979), Schmitt (1981), and de Jong-Gierveld and Kamphuis (1985).

Data are analyzed most often with a single model. As a result, a systematic comparison of models is often difficult to achieve. In this study, the three models were compared both theoretically and empirically. First, the models are briefly introduced and compared on three important characteristics: restrictiveness with respect to empirical data, measurement properties of the scale, and methods to determine reliability of measurement. Second, the fit of the three models to the same set of empirical data is investigated. Differences and resemblances between the models are illustrated on the basis of the results from the empirical analyses.

## Theoretical Considerations

The IRT models proposed by Mokken (1971) are called nonparametric because the IRFs are not parametrically defined, and because no assumptions are made concerning the distribution of the latent trait. The model of monotone homogeneity is based on the assumptions of unidimensionality, local stochastic independence, and monotonicity in the latent attribute. Unidimensional measurement

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 14, No. 3, September 1990, pp. 283-298*

*0146-6216/90/030283-16$2.05*

states that within a specified population, response behavior on the test can be explained by a single underlying attribute. This attribute is measured on a scale denoted by θ. Local stochastic independence implies that the response behavior of a person on an arbitrarily selected item $g$ is not influenced by his or her responses on previous items, nor will it affect response behavior on subsequent items. This assumption is a logical consequence of unidimensionality, but the reverse is not true.

The assumption of monotonicity in the latent attribute specifies that a higher attribute value implies an increasing probability of responding positively to an item. Because present psychological theories seem to offer no viable alternatives, it may indeed be plausible that someone with a higher ability, aptitude, attitude, or trait does not have a smaller probability of answering items positively than someone with a lower level of that variable.

Let $P$ denote a probability, in general, and let $P_g(\theta_i)$ denote the probability of person $i$ obtaining a positive response on item $g$. Formally, the model of monotone homogeneity implies for two persons $i$ and $j$ and an item $g$, that if $\theta_i < \theta_j$, then

$$P_g(\theta_i) \leq P_g(\theta_j) \quad , \tag{1}$$
for $i, j = 1, \ldots, n$, $i \neq j$, and $g = 1, \ldots, k$.

When items conform to the model, persons can be ordered according to the latent attribute (Mokken, 1971). Because θ is not estimated numerically in the model of monotone homogeneity, the ordering according to θ is replaced by the ordering according to the true score $T$ from classical test theory (CTT; Lord & Novick, 1968). Mokken (1971) showed that $T$ and θ have the same order. The number-correct score $X$ is used as an estimator for $T$; an ordering according to $X$, therefore, gives an estimate of the ordering according to θ. Grayson (1988) also showed that the number-correct score $X$ has a monotone likelihood ratio in θ for all models complying with monotone homogeneity.

The model of double monotonicity is based on the same set of assumptions as the model of monotone homogeneity, plus the additional assumption of monotonicity in the item difficulties. Together these assumptions imply that the IRFs do not intersect, but may touch or coincide. Formally, the property of nonintersecting IRFs can be stated as follows: Let δ denote the common item difficulty from IRT; then for two items $g$ and $h$ with $\delta_g < \delta_h$,

$$P_g(\theta) \geq P_h(\theta) \quad , \text{ for all } \theta. \tag{2}$$

In addition to ordering persons according to θ, the model allows an ordering of items according to their difficulties. Apart from ties, this order is the same in each subpopulation of the population where the model holds. In the model of double monotonicity, the true score $T$ is used for ordering persons, and $\pi_g$, the proportion of persons giving a positive response to item $g$, is used for ordering items. Mokken (1971) showed that the order according to $\pi_g$ is the reverse of the order according to the difficulty parameter $\delta_g$ from the parametric IRT models. Mokken and Lewis (1982), Niemöller and van Schuur (1983), and Sijtsma (1988) provide an introduction to the Mokken model.

In the one-parameter logistic (Rasch) model, test behavior is explained by the difficulty of an item and the attribute value of a person (Fischer, 1974; Rasch, 1960). This model is based on the same set of assumptions as the Mokken model of monotone homogeneity, plus the assumption of minimal sufficiency of the unweighted person and item sum scores for the estimation of the θ and δ parameters, respectively. Given these four assumptions, the probability of positively answering an item is given by:

$$P_g(\theta) = \frac{\exp(\theta - \delta_g)}{1 + \exp(\theta - \delta_g)} \quad . \tag{3}$$

Sometimes it is convenient to use the transformation $\theta = \ln \xi$ and $\delta_g = \ln \epsilon_g$, and Equation 3 then becomes

$$P_g(\xi) = \frac{\xi}{\xi + \epsilon_g} \quad . \tag{4}$$

This equation is used in the next section.

An important characteristic of the Rasch model is that it enables so-called specifically objective measurement (Fischer, 1974, 1987; Wright, 1977). Furthermore, it can be shown (e.g., Fischer, 1974) that, for both person θs and item δs, the Rasch model allows measurement on a difference or a ratio scale.

## Comparison of the Models

In ordering the models on the basis of their restrictiveness for empirical data, the model of monotone homogeneity is least restrictive and the Rasch model is most restrictive. This is clear from a consideration of the assumptions on which the models are based.

Table 1 shows that the Rasch model rests on the same set of assumptions as the model of monotone homogeneity, plus the additional assumption of sufficiency. Taken together, these four assumptions also imply nonintersecting IRFs. The Rasch model is thus a special case of the model of double monotonicity, and is therefore a more restrictive model. This ordering does not imply that nonparametric models are always clearly less restrictive than parametric models. For example, it is not obvious which of the two models is most restrictive—the model of double monotonicity that does not allow intersection of IRFs but does not restrict IRFs to the logistic function, or the three-parameter logistic model that does allow intersection but is limited to logistic IRFs. This ordering of models reflects the expectation that, empirically, fewer items comply with the Rasch model than with the Mokken models, and that fewer items will comply with the model of double monotonicity than with the model of monotone homogeneity.

**Table 1**
Assumptions of the Mokken Model of Monotone Homogeneity (MH), Mokken Model of Double Monotonicity (DM), and the Rasch Model (RM) [* Indicates That the Assumption Pertains to the Model]

| Assumption | MH | DM | RM |
|---|---|---|---|
| Unidimensionality | * | * | * |
| Local stochastic independence | * | * | * |
| Monotonicity in θ | * | * | * |
| Monotonicity in δ | | * | *[1] |
| Minimal Sufficiency | | | * |

[1]Monotonicity in δ holds in the Rasch model, but it is made redundant by the other four assumptions.

The model of monotone homogeneity allows ordering of persons with respect to θ. Because IRFs may intersect, the ordering of items according to their difficulties varies with θ. The overall ordering of $k$ items, according to $\pi_g$ ($g = 1, \ldots, k$), does not represent the correct ordering within subgroups. Ordering of both persons and items is possible when the model of double monotonicity applies;

however, attributes and difficulties are measured on separate scales. Attributes are measured on the true score scale, and difficulties are measured on the scale of proportions. For the Rasch model, measurement of items and persons takes place on a common difference or a common ratio scale. The measurement properties of the three models are summarized in Table 2.

**Table 2**
Measurement Properties of the Three Models

| Model | Persons | Items |
|---|---|---|
| MH | Ordinal | |
| DM | Ordinal | Ordinal |
| RM | Difference/Ratio | Difference/Ratio |

Measurement on a difference or a ratio scale is restricted to psychometric interpretations of the scale. From a psychological point of view, however, it is impossible to compare persons on a ratio or a difference scale in terms of psychological attributes, because psychological theories are too ambiguous and vague to allow comparison of persons on metric scales. For example, given a person $i$ with $\xi_i = 1$, and a person $j$ with $\xi_j = 2$, it does not make sense to conclude that person $j$ is twice as intelligent as person $i$.

The odds for success are needed in order to understand the meaning of comparing persons on the ratio scale from the Rasch model. The odds ($O_{ig}$) for success of a person $i$ on item $g$ is the ratio of the success probability on item $g$ and the failure probability on the same item (denoted by $P_g$ and $Q_g$, respectively):

$$O_{ig} = P_g(\xi_i)/Q_g(\xi_i) \quad . \tag{5}$$

For the Rasch model, Equation 5 becomes $O_{ig} = \xi_i/\epsilon_g$. The ratio of the odds of persons $i$ and $j$ is given by:

$$O_{ig}/O_{jg} = \xi_i/\xi_j \quad . \tag{6}$$

For example, on an intelligence test for two persons with $\xi_i = 1$ and $\xi_j = 2$, $O_{ig}/O_{jg} = .5$. The odds for success for person $j$ are twice the odds for success for person $i$. Obviously, it cannot be concluded that person $j$ is twice as intelligent as person $i$. The interpretation in terms of odds follows from the Rasch model, whereas the interpretation in psychological terms does not.

In IRT, reliability of measurement is determined by the accuracy with which the latent attribute is estimated (Lord, 1980). In parametric IRT, the information function has a role analogous to the reliability coefficient in CTT. Both provide a tool for determining the precision of measurement by means of $\hat{\theta}$ and $X$, respectively. Classical reliability only allows an overall evaluation of the precision of measurement, whereas the information function allows an evaluation of measurement precision as a function of the latent attribute.

The nonparametric Mokken models do not allow numerical estimates of person and item parameters that are needed to estimate the information function. Therefore, accuracy of measurement is determined with the reliability coefficient from CTT. For the model of monotone homogeneity, no specific methods to determine the reliability of measurement have been developed. Methods from CTT, such as coefficient $\alpha$ or $\lambda$-2 (Guttman, 1945), must be used to estimate accuracy of measurement.

Based on the model of double monotonicity, Mokken (1971), and Sijtsma and Molenaar (1987) developed methods allowing for an overall estimate of the precision of measurement using the number-correct score $X$. These methods make explicit use of the assumption that IRFs do not intersect. Establishment of a fitting model of double monotonicity should thus precede application of these

methods.

In the Rasch model, the information function is given by

$$I = \frac{1}{\sigma^2(\hat{\theta}|\theta)} \quad . \tag{7}$$

This function gives information about the precision of measurement using the reciprocal of the conditional variance of the maximum likelihood estimate of $\theta$. It has been shown (Lord, 1980) that the contribution of an item to the precision of measurement is independent of the other items in the test. The test information function allows an evaluation of measurement precision on each point of the scale.

Gustafsson (1977) discussed the index of examinee separation in the context of the Rasch model. In contrast to the information function, this index expresses the overall reliability of the maximum likelihood estimate $\hat{\theta}$, and can thus be seen as an IRT counterpart of the classical reliability coefficient for the number-correct score $X$.

## Method

### Data

The goodness-of-fit procedures of the three models are illustrated with an analysis of data from a widely-used Dutch verbal intelligence test—the Verbal Analogies Test (Drenth & van Wieringen, 1969). This test contains 40 items that measure "verbal intelligence." Each item consists of a sentence in which the first- and the last-item word have been removed. Two rows of five words each are listed below the sentence. The examinee has to identify one word from the first row (numbered 1 through 5) as the first word of the sentence, and one word from the second row (lettered A through E) as the last word of the sentence. An example of an item is:

. . . is to love as hostility is to . . .

| | | | | |
|---|---|---|---|---|
| 1. kiss | 2. enemy | 3. wedding | 4. hate | 5. lover |
| A. ally | B. law | C. quarrel | D. passion | E. friendship . |

The test was developed for examinees at or beyond the college/university level. The administration time of the test is 40 minutes. For the present analysis, a sample of 990 examinees was used; most of the examinees graduated from high school.

The dataset used archival data that were collected in the context of selection of personnel for computer occupations, such as systems analyst and programmer. Because the test has a time limit, and a persons-by-items matrix containing correct-incorrect scores was available, it was unknown whether each examinee had attempted each item. Therefore, a 0 score in the data matrix may have meant "omitted" (due to time limit), as well as "incorrectly answered."

To minimize the probability that part of the data analyzed was not attempted by all examinees, the analysis was restricted to the first 32 items. This decision was based on the observation that after Item 32, there was a sudden decrease of the items' $\pi_g$ values, the proportion of correct responses. Because the last eight items did not differ in task from the first 32 items, this sudden decrease was assumed to be caused by the time limit and not by real differences in difficulty.

### Goodness-of-Fit Methods

*Monotone homogeneity.* The data were first analyzed by methods for evaluating goodness-of-fit of the model of monotone homogeneity. Given the assumptions of this model, Mokken (1971) showed that inter-item covariances are nonnegative. Furthermore, the overall scalability coefficient $H$ (Mokken, 1971; Mokken & Lewis, 1982; see also Cliff, 1977, who denotes this coefficient by $c_{t3}$) and the two related coefficients for individual items (denoted by $H_g$) and pairs of items (denoted by $H_{gh}$) also

have nonnegative values. A positive value, however, does not guarantee monotone homogeneity (e.g., Mokken & Lewis, 1982). Positive values thus constitute necessary conditions for monotone homogeneity. For practical test construction purposes, Mokken (1971) recommended the value $H = .3$ as a lower bound for a set of items comprising a test; Mokken, Lewis, and Sijtsma (1986) discussed the rationale underlying this choice.

Because $H \geq 0$ is a necessary condition for monotone homogeneity, the empirical data were first investigated with respect to this restriction. Many researchers use $H = .3$ as a practical lower bound (Mokken, 1971), and therefore the data were also investigated using this lower bound.

A second procedure used to evaluate the assumption of monotone homogeneity in these data is that proposed by Molenaar (1982; 1983a). In this method, given an item $g$ ($g = 1, \ldots, k$), for each person the "rest" score $S = X - X_g$ is determined (see Rosenbaum, 1984, for the use of rest score rather than raw score). Persons having the same rest score together constitute a rest score group. The rest score groups are ordered according to increasing rest score $S$ ($S = 0, \ldots, k - 1$). The proportion of persons answering positively to item $g$ is estimated within each rest score group, and the proportion is denoted by $\pi_{sg}$. $\pi_{sg}$ is estimated by $\hat{\pi}_{sg} = n_{sg}/n_s$, where $n_{sg}$ denotes the number of persons in the sample having a rest score equal to $S$ who give a positive answer to item $g$, and $n_s$ is the number of persons having a rest score equal to $S$.

Provided the model of monotone homogeneity holds (given increasing rest score) for a fixed item $g$, the proportions $\pi_{sg}$ are nondecreasing. When a sample proportion $\hat{\pi}_{s-1,g}$ is followed by a smaller proportion $\hat{\pi}_{sg}$, the null hypothesis of equal proportions is tested against the alternative of $\hat{\pi}_{s-1,g} > \hat{\pi}_{sg}$. For small samples this is done with Fisher's exact probability test, and for large samples with the $\chi^2$ test for independent samples (Siegel, 1956, p. 96–111) in the two-by-two table of rest score groups $(S - 1, S)$ by item score $X_g$ (0,1).

Data analysis was done using the program MOKKEN SCALE (Niemöller & van Schuur, 1980; see Debets, Sijtsma, Brouwer, & Molenaar, 1989, for a program that handles polychotomous scores, as well). This program contains a bottom-up item selection procedure that starts by selecting the pair of items for which (1) $H_{gh}$ is significantly larger than 0, and (2) $H_{gh}$ is the largest among the coefficients for all possible item pairs. Then a third item $f$ is selected that (3) correlates positively with the items already selected, (4) has an $H_f$ coefficient with respect to the items selected that is significantly larger than 0, and (5) has an $H_f$ coefficient that is larger than a user specified value $c$; in this study, two analyses were performed using $c = 0$ and $c = .3$, respectively.

From the pool of items, an item is selected that maximizes the overall $H$ of $g$, $h$, and $f$. A fourth item is selected that satisfies conditions 3, 4, and 5, above, and that maximizes the overall coefficient of the four items selected. The program continues to select items as long as items are available that satisfy conditions 3, 4, and 5. During the selection process a large number of significance tests are computed, so there is a great danger of capitalization on chance. In order to reduce this danger, the significance level is automatically adapted throughout the procedure to the number of tests in each selection step. The data were analyzed once with, and once without, this selection procedure.

*Double monotonicity.* To determine whether the model of double monotonicity adequately explained response behavior, Mokken (1971) proposed a visual inspection method for the evaluation of the complete dataset with the order properties within two matrices. The **P** matrix has order $k \times k$ and contains the proportions $\pi_{gh}$ ($g,h = 1, \ldots, k, g \neq h$) of persons giving correct answers to a pair of items. The items across rows and columns are ordered according to increasing order of overall $\pi_g$ values. Given the assumption of double monotonicity, the rows and columns are nondecreasing (Mokken, 1971). The $\mathbf{P}^{(0)}$ matrix also has order $k \times k$, and contains the proportions $\pi_{\bar{g}\bar{h}}$ ($g,h = 1, \ldots, k, g \neq h$) of persons giving negative responses to a pair of items. Given double monotonicity,

the rows and columns in this matrix are nonincreasing. These expected ordering properties in the **P** and $\mathbf{P}^{(0)}$ matrices in the sample are used to evaluate double monotonicity.

Another possibility for investigating the model of double monotonicity is based on the item-by-item cross table (Molenaar, 1982; 1986) for test score groups. Let $\pi_{g\bar{h}}$ denote the proportion of persons having a positive response on item $g$ and a negative response on item $h$, and let $\pi_{\bar{g}h}$ denote the proportion of persons having a negative response on item $g$ and a positive response on item $h$. Given the model of double monotonicity, with $\pi_g \leq \pi_h$ in the population, it can be shown (Molenaar, 1982) that in each subgroup (e.g., test score group) $\pi_{g\bar{h}} \leq \pi_{\bar{g}h}$. Sample results in the opposite direction (i.e., $\hat{\pi}_{g\bar{h}} > \hat{\pi}_{\bar{g}h}$) may be indicative of violations of the overall item order. The null hypothesis of equal proportions, $\pi_{g\bar{h}}/(\pi_{g\bar{h}} + \pi_{\bar{g}h}) = .5$, against the alternative that $\pi_{g\bar{h}}/(\pi_{g\bar{h}} + \pi_{\bar{g}h}) > .5$, can be tested for test score groups with the McNemar test (Siegel, 1956, p. 63–67).

*Rasch homogeneity.* In order to select Rasch homogeneous sets of items from a larger set, Andersen's (1973) conditional likelihood ratio test was used for globally testing the assumptions of monotonicity in $\theta$ and sufficiency (van den Wollenberg, 1979). The sample of persons was divided into two disjoint subsamples of about equal size, one containing examinees having the higher raw scores, and the other containing examinees having lower raw scores. Item parameters were estimated in these subsamples. If the model held, the same parameters were estimated in the subsamples, and the $\chi^2$ distributed test statistic had a value reflecting only random fluctuations. If the model was globally rejected, in the next step of the analysis the test proposed by Molenaar (1983b; see also Glas, 1989, for new proposals) was used for detecting deviant items.

With this test, for an item $g$ within each test score group, the proportion of positive answers is compared with the expected proportion, given that the model holds. For item $g$, the standardized differences are combined across the lower and higher test score groups into the test statistic $U_g$. This test statistic is approximately standard normally distributed. As is shown below, practical use of $U_g$ is more complicated than simply removing items having significant $U_g$ values (see also Molenaar, 1983b).

After removal of items based on the $U_g$ analyses, response behavior on the remaining items was tested with respect to the assumptions of unidimensionality and local stochastic independence using Andersen's conditional likelihood ratio test. For this purpose, subgroups were constructed with splitter items (van den Wollenberg, 1982).

A splitter item is an item with which the sample is partitioned into two subsamples, one consisting of examinees responding positively to the item, and the other containing examinees responding negatively. Under the alternative hypothesis of multidimensionality, some of the items in the test measure the same attribute as the splitter item, and other items do not. Van den Wollenberg (1982) argued that in this situation, item parameter estimates belonging to items related to the splitter item differ systematically across the subsamples, resulting in a significant value of Andersen's test statistic. A comparison of item parameter estimates in both groups may reveal deviant items, and may suggest which items should be removed or which unidimensional subscales may be appropriate.

The splitter item technique resembles the use of the **P** and $\mathbf{P}^{(0)}$ matrices for evaluating the model of double monotonicity. For example, the **P** matrix contains all bivariate proportions $\pi_{fg} (g = 1, \ldots, k)$ in row $f$, and if $\pi_g \leq \pi_h$, then $\pi_{fg} \leq \pi_{fh}$. Division of $\pi_{fg}$ and $\pi_{fh}$ by $\pi_f$ shows that the ordering of item difficulties holds in the group responding correctly to item $f$. This ordering also holds in the group giving a negative response to item $f$ ($\pi_{\bar{f}g} \leq \pi_{\bar{f}h}$), and corresponds in the $\mathbf{P}^{(0)}$ matrix with $\pi_{\bar{f}\bar{g}} \geq \pi_{\bar{f}\bar{h}}$. Thus the technique of inspection of the ordering in rows $f$ in the **P** and $\mathbf{P}^{(0)}$ matrices is likely an adaptation of the splitter item technique for the model of double monotonicity (Molenaar, 1982). In practice, however, inspection of these matrices is never used as a check on unidimensionali-

ty, but rather as a check on the property of nonintersecting IRFs.

Another test for unidimensionality was provided by Martin-Löf (1973), and Wainer, Morgan, and Gustafsson (1980). To use the Martin-Löf test, the item set is divided into two disjoint subsets of items that hypothetically measure different attributes. Because the Verbal Analogies Test was intended to be unidimensional, it was difficult to group the items on an a priori basis. Therefore, the item set was divided into two subsets of items on the basis of their $\pi_g$ values. The idea behind this division was that difficult items may measure another dimension than easy items. An explicit procedure for evaluating unidimensionality is notably absent in both Mokken models.

### Precision of Measurement

In addition to an investigation of model fit, the precision of measurement was also examined for each model. Coefficient $\alpha$ was estimated for the model of monotone homogeneity. Mokken's Method 1 (Mokken, 1971), and a method proposed by Sijtsma and Molenaar (1987) were used to estimate precision of measurement for the model of double monotonicity. These methods are less biased than Mokken's Method 2 (Sijtsma, 1988; Sijtsma & Molenaar, 1987). For the Rasch model, precision of measurement was determined with the test information function and the index of examinee separation.

## Results

### Analysis of Monotone Homogeneity

The 32 items were analyzed as a test. Then the item selection algorithm was used to explicitly select items conforming to the model of monotone homogeneity.

Without selection, all items had positive $H_g$ values ranging from .09 to .38. For the complete set of 32 items, $H = .25$. The $H$ and the $H_g$ values resulting from the stepwise item selection using $H = 0$ as a lower bound are shown in Table 3. Only Item 22 violated the model of monotone homogeneity. This item correlated negatively with Items 3, 6, and 12.

The trend of the overall $H$ coefficient during the stepwise selection process revealed that $H$ decreased relatively quickly during the selection of the first few items, but then tended to stabilize. This is in agreement with results from a simulation study by Sijtsma and Prins (1986) in which items were selected from a set of equidistant Rasch items.

The $\pi_g$ values in Table 3 are spaced rather evenly. Although these $\pi_g$ values are not linearly related to the latent difficulty ($\delta$) parameters from the Rasch model, this spacing suggests approximately equidistant $\delta$ parameters, as in the simulation study. Furthermore, because $H_g$ is related to the discriminating power of an item, with the variance of the person distribution and the distance of the item difficulties held constant (Mokken et al., 1986), the small dispersion of most $H_g$ coefficients (in Table 3, 23 out of 31 coefficients have values between .2 and .3) may point in the direction of only modest differences in discriminating power across the items. Therefore, similar results concerning the trend of $H$ during the selection of items found by Sijtsma and Prins (1986) and found in this study may be explained by comparable latent item characteristics. The estimation of reliability of the unweighted total score with $\alpha$ for 31 items resulted in $\alpha = .86$.

Table 4 shows results for the item selection using the lower bound $H = .3$. This additional restriction resulted in the selection of two subscales. Table 4 shows for Item 31 that it is possible that some $H_g$ values could be smaller than .3 after item selection is completed. One explanation for such small values is that when an item is selected, this decision is based on its $H_g$ value with regard to the items already selected. When selected, the $H_g$ value is not smaller than the user-specified value, but the final $H_g$ value is estimated with respect to all items selected in the scale when the selection is com-

**Table 3**
$H$ and $H_g$ Scalability Coefficients for Stepwise Selection From the First 32 Items of the Verbal Analogies Test ($H \geq 0.0$) [$H_g$ Values Were Determined After the Selection Was Completed]

| Item | $H$ | $H_g$ | $\pi_g$ |
|---|---|---|---|
| 28 | | .39 | .24 |
| 21 | .65 | .26 | .62 |
| 4 | .51 | .31 | .81 |
| 27 | .47 | .35 | .35 |
| 5 | .45 | .29 | .55 |
| 16 | .44 | .32 | .66 |
| 3 | .43 | .32 | .93 |
| 2 | .41 | .28 | .87 |
| 9 | .40 | .29 | .50 |
| 7 | .39 | .28 | .73 |
| 31 | .38 | .26 | .14 |
| 6 | .37 | .30 | .74 |
| 8 | .36 | .26 | .56 |
| 26 | .36 | .28 | .37 |
| 15 | .35 | .28 | .66 |
| 30 | .35 | .28 | .34 |
| 19 | .34 | .26 | .52 |
| 23 | .33 | .27 | .47 |
| 18 | .33 | .26 | .41 |
| 11 | .32 | .25 | .55 |
| 14 | .32 | .22 | .67 |
| 29 | .31 | .25 | .29 |
| 17 | .31 | .23 | .64 |
| 10 | .30 | .22 | .59 |
| 20 | .29 | .22 | .55 |
| 13 | .29 | .21 | .68 |
| 1 | .28 | .20 | .69 |
| 25 | .28 | .22 | .58 |
| 32 | .27 | .19 | .34 |
| 12 | .26 | .15 | .60 |
| 24 | .25 | .12 | .62 |
| 22[a] | | .08 | .16 |

[a]Item 22 was rejected.

pleted. Due to items selected subsequently, it is then possible that the $H_g$ value of an item is smaller than the value at initial selection, and even smaller than the user-specified value $c$.

After the first scale is completed, the program continues selecting items from the set remaining—this selection round resulted in Scale 2 in Table 4. There is an option in the program allowing the items in Scale 2 to function as a start set for a new selection round from all items. With this option, a new scale was found containing the three items of Scale 2, plus 11 items from the original first scale. It is generally recommended to take item and subscale content into account when deciding what scale to use in practice. In this case, however, it was not possible to decide which end result from item selection should be preferred, because items could not be grouped on the basis of their content. Therefore, the longer Scale 1 was used as the final scale. $\alpha$ for Scale 1 was .81, and $\alpha$ for Scale 2 was .42.

**Table 4**
$H$ and $H_s$ Scalability Coefficients for Stepwise Selection of Scales 1 and 2 From the First 32 Items of the Verbal Analogies Test ($H \geq .30$)

| Item | $H$ | $H_s$ |
|---|---|---|
| Scale 1 | | |
| 28 | | .34 |
| 21 | .65 | .33 |
| 4 | .51 | .36 |
| 27 | .47 | .39 |
| 5 | .45 | .32 |
| 16 | .44 | .36 |
| 3 | .43 | .33 |
| 2 | .41 | .33 |
| 9 | .40 | .35 |
| 7 | .39 | .34 |
| 31 | .38 | .28 |
| 6 | .37 | .34 |
| 8 | .36 | .33 |
| 26 | .36 | .32 |
| 15 | .35 | .33 |
| 30 | .35 | .30 |
| 19 | .34 | .31 |
| Scale 2 | | |
| 29 | | .32 |
| 13 | .38 | .35 |
| 11 | .32 | .31 |

Over 200 sample violations of decreasing proportions were found in the columns of the rest score-by-item matrix for all 32 items. Applying the method proposed by Molenaar (1982), it was found that only two of these violations were significant at a .05 level (Item 5 in rest score groups 25 and 27, $\chi^2 = 4.5$, $df = 1$, $p = .03$; and Item 22 in rest score groups 6 and 11, $\chi^2 = 4.2$, $df = 1$, $p = .04$). Because these were only local violations of the assumption of monotone IRFs, all IRFs were considered to be in concordance with the model. Based on negative correlations, however, only Item 22 (not selected with stepwise selection) did not comply with the model of monotone homogeneity.

### Analysis of Double Monotonicity

A visual inspection of the **P** and $\mathbf{P}^{(0)}$ matrices suggested that several violations of monotone orderings were present in their rows and columns. Violations greater than .03 (Mokken, 1971) were considered too large to maintain the model of double monotonicity for the complete set of items.

Because of the large size of the entire **P** and $\mathbf{P}^{(0)}$ matrices, only parts of these matrices are shown in Table 5. For the **P** matrix it can be verified that violations are present for Items 5, 8, and 25. For the entire **P** matrix, Items 5, 8, 11, 16, 19, 22, 25, 27, and 32 violated the assumptions of double monotonicity. For the $\mathbf{P}^{(0)}$ matrix, no violations occurred for the items in Table 5. For the full matrix, Items 11, 16, and 26 violated double monotonicity.

Testing violations of the item ordering in the test score groups with the McNemar test revealed significant violations (in at least two score groups) involving Items 5, 8, 11, 12, 16, 19, 22, 24, 25,

**Table 5**
**P and $P^{(0)}$ Matrices for Nine Items**

| Item | 5 | 20 | 11 | 8 | 25 | 10 | 12 | 21 | 24 |
|---|---|---|---|---|---|---|---|---|---|
| **P Matrix** | | | | | | | | | |
| 5 | – | .37 | .36 | .36 | .37 | .37 | .37 | .40 | .34 |
| 20 | .37 | – | .35 | .37 | .35 | .36 | .36 | .39 | .36 |
| 11 | .36 | .35 | – | .35 | .34 | .38 | .36 | .38 | .37 |
| 8 | .36 | .37 | .35 | – | .36 | .39 | .35 | .40 | .37 |
| 25 | .37 | .35 | .34 | .36 | – | .38 | .38 | .42 | .38 |
| 10 | .37 | .36 | .38 | .39 | .38 | – | .37 | .40 | .39 |
| 12 | .37 | .36 | .36 | .35 | .38 | .37 | – | .41 | .40 |
| 21 | .40 | .39 | .38 | .40 | .42 | .40 | .41 | – | .41 |
| 24 | .34 | .36 | .37 | .37 | .38 | .39 | .40 | .41 | – |
| **$P^{(0)}$ Matrix** | | | | | | | | | |
| 5 | – | .27 | .27 | .26 | .25 | .24 | .22 | .24 | .19 |
| 20 | .27 | – | .25 | .26 | .23 | .23 | .21 | .22 | .19 |
| 11 | .27 | .25 | – | .25 | .22 | .25 | .21 | .21 | .20 |
| 8 | .26 | .26 | .25 | – | .23 | .24 | .20 | .23 | .19 |
| 25 | .25 | .23 | .22 | .23 | – | .22 | .20 | .22 | .18 |
| 10 | .24 | .23 | .25 | .24 | .22 | – | .18 | .20 | .18 |
| 12 | .22 | .21 | .21 | .20 | .20 | .18 | – | .20 | .18 |
| 21 | .24 | .22 | .21 | .23 | .22 | .20 | .20 | – | .17 |
| 24 | .19 | .19 | .20 | .19 | .18 | .18 | .18 | .17 | – |

26, 27, and 32. Except for Items 12 and 24, the other 10 items were also found to be deviant with the **P** and $P^{(0)}$ matrices. Based on these analyses, these 12 items were removed from the test. For the remaining 20 items, the reliability estimated with Mokken's Method 1 was .8. The method proposed by Sijtsma and Molenaar (1987) yielded an estimate of .81.

## Rasch Analysis

The assumptions of monotonicity and sufficiency were investigated with Andersen's (1973) conditional likelihood ratio test. The group of persons was divided into two disjoint subsamples on the basis of their raw scores. One group contained persons with scores ranging from 1 to 19, and the other group contained persons with scores ranging from 20 to 31. On the basis of this division, the null hypothesis of equal item parameters in the subgroups was rejected in favor of the alternative hypothesis of unequal parameters ($\chi^2 \approx 209$, $df = 31$; $p < .001$). Based on these results, it may be concluded that the assumptions of monotonicity and sufficiency were violated.

To find deviant items, individual items were investigated with the $U_g$ test at a .05 significance level. Items with $|U_g| > 1.96$ violate the Rasch model. It could be argued that a smaller significance level should be adopted in accordance with the number of significance tests carried out. Because one or only a few small score groups may be responsible for large positive or negative $U_g$ values, the observed and predicted frequencies in separate score groups were also studied for each item. When such local violations occurred due to small score groups, significant negative or positive values of $U_g$ were not considered to be relevant indicators of misfit. Therefore, the combination of a .05 significance level and the inspection of frequencies in score groups likely constituted a satisfactory protection against chance capitalization.

Items 5, 9, 15, 16, 27, and 28 measured the latent trait with a steeper IRF ($U_g < -1.96$) than the other items, as shown in Table 6. Items 1, 12, 14, 22, 24, and 32 had IRFs that were relatively flat

**Table 6**
$U_g$ Values for All Items From the Verbal Analogies Test (Analysis 1), After Removal of Items With Relatively High $U_g$ Values (Analysis 2), and After Removal of Items With Relatively High and Relatively Low $U_g$ Values (Analysis 3)

| | Analysis | | |
|---|---|---|---|
| Item | 1 | 2 | 3 |
| 1 | 2.6** | – | – |
| 2 | .7 | 1.2 | .3 |
| 3 | .4 | 1.3 | .2 |
| 4 | -1.6 | -.8 | -1.3 |
| 5 | -2.9** | -2.4** | – |
| 6 | -1.3 | -.3 | -.5 |
| 7 | -1.1 | -.7 | -.5 |
| 8 | -1.9 | -.3 | -1.1 |
| 9 | -3.5** | -2.5** | – |
| 10 | .4 | 1.0 | .5 |
| 11 | -.7 | 1.3 | .5 |
| 12 | 5.1** | – | – |
| 13 | 1.7 | 2.8** | 2.8** |
| 14 | 2.1** | – | – |
| 15 | -3.2** | -1.3 | – |
| 16 | -3.4** | -2.8** | – |
| 17 | .6 | 1.2 | 1.0 |
| 18 | .5 | 1.5 | 1.1 |
| 19 | -.8 | .1 | .7 |
| 20 | .4 | 1.8 | 1.1 |
| 21 | -1.4 | -1.2 | -.6 |
| 22 | 10.2** | – | – |
| 23 | -.9 | -.2 | -.9 |
| 24 | 6.8** | – | – |
| 25 | 1.3 | 2.5* | 1.4 |
| 26 | -1.6 | -2.9** | -2.2** |
| 27 | -4.6** | -4.6** | – |
| 28 | -4.2** | -3.8** | – |
| 29 | 1.4 | 2.6* | 1.5 |
| 30 | .2 | .6 | -1.2 |
| 31 | 1.8 | 3.1* | 2.5** |
| 32 | 4.0** | – | – |

*Local Violation.
**Global Violation.

($U_g > 1.96$). Items 22 and 24 had particularly large positive $U_g$ values.

The tasks that have to be fulfilled to answer an item correctly do not vary systematically across items. This makes removal of items on the basis of their content somewhat arbitrary. Based on psychometric considerations only, items with relatively flat IRFs were removed. Although items with relatively steep IRFs also violate the Rasch model and thus are candidates for removal from the test, they sharply discriminate between examinees and are useful in an intelligence test. Although these items might be considered for removal, as well, they were kept in the test in this first selection round

for practical reasons. Further, experience indicated that the $U_g$ value of items with relatively steep curves may improve when items with positive $U_g$ values were removed.

After removal of Items 1, 12, 14, 22, 24, and 32, the remaining items still violated the assumptions of monotonicity and/or sufficiency. The results for the Andersen test statistic were $\chi^2 \approx 75$, $df = 25$, $p < .001$. Table 6 shows the $U_g$ values after removal of the six weakly-discriminating items (Analysis 2). Removal of these items, as well as removal of items with relatively steep IRFs, gave a better fit of the model. The Andersen test for score groups (high-low partitioning) was $\chi^2 \approx 33$, $df = 19$, $p = .02$.

Although Items 13 and 31 still had relatively flat IRFs, and the IRF of Item 26 was too steep (Analysis 3 in Table 6), the selection of items was terminated to avoid chance capitalization. The remaining set of 20 items was analyzed with respect to unidimensionality, using the Martin-Löf test and the Andersen test using splitter items. For the Martin-Löf test, the set of 20 remaining items was divided into two disjoint subsets—one containing the relatively easy items (Items 2, 3, 4, 6, 7, 8, 11, 13, 17, and 21), and the other containing the relatively difficult items (Items 10, 18, 19, 20, 23, 25, 26, 29, 30, and 31). The results for the Martin-Löf statistic were $\chi^2 \approx 106$, $df = 109$, $p = .54$. It can thus be assumed that the remaining items allowed for unidimensional measurement.

To investigate unidimensionality with the Andersen test using splitter items, substantive clues for a sensible choice of such items were absent. Therefore, items with a difficulty $\pi_g$ of about .5 were selected so that the accuracy of the item parameter estimation was about the same in both groups. The results of the Andersen test are shown in Table 7 for Splitter Items 8, 11, 20, 25, and 30. Although the hypothesis of unidimensional measurement should formally be rejected at the .05 level on the basis of splitter items 8, 11, and 30, the $\chi^2$ values were not high enough to be of concern. When combined with the other results obtained with the Martin-Löf test and the Andersen test, these results appeared to suggest unidimensionality.

**Table 7**
Results of the Andersen Test Based on Splitter Items

| | Splitter Item | | | | |
|---|---|---|---|---|---|
| Statistic | 8 | 11 | 20 | 25 | 30 |
| $\pi_g$ | 0.56 | 0.55 | 0.55 | 0.58 | 0.34 |
| $\chi^2$ | 48.70 | 53.30 | 27.80 | 22.20 | 32.50 |
| $df$ | 18 | 18 | 18 | 18 | 18 |
| $p$ | <0.001 | <0.001 | 0.06 | 0.22 | 0.02 |

If the statistical tests used for the Rasch analysis would have been too powerful with a sample size of 990, a 20-item Rasch scale would not have been obtained as the final result; rather, the Verbal Analogies Test would have been broken down step-by-step into many fragments. Because this did not happen, the use of the statistical tests was deemed appropriate.

The information function for $\hat{\theta}$, based on the 20 items that were finally retained in the test, had the well-known bell-shaped appearance. The most precise measurement took place around $\hat{\theta} = .2$; at $\hat{\theta} = -3$ and $+3$, information was about one-third of the maximum information. The index of examinee separation was .78, which gives an overall impression of measurement accuracy with $\hat{\theta}$.

## Discussion

For empirical applications, the ordering of models with respect to restrictiveness is an important notion. The Mokken model of monotone homogeneity is least restrictive and the Rasch model is

most restrictive. This order usually will appear in analyses of empirical data: The Rasch model explains behavior on fewer items than the model of double monotonicity, and fewer items should be in concordance with this model than the model of monotone homogeneity.

The results of the empirical analyses were largely in agreement with this expectation (see Table 8). The model of monotone homogeneity ($H \geq 0$) did not apply for only 1 item (Item 22), and 12 items did not fulfill the requirements of the model of double monotonicity (Items 5, 8, 11, 12, 16, 19, 22, 24, 25, 26, 27, and 32). Furthermore, 12 items did not fit the Rasch model (Items 1, 5, 9, 12, 14, 15, 16, 22, 24, 27, 28, and 32).

**Table 8**
Items Selected by the Three Models
(* and + Indicate Different Scales)

| | MH | | | |
|---|---|---|---|---|
| Item | $H \geq 0$ | $H \geq .3$ | DM | RM |
| 1 | * | | * | |
| 2 | * | * | * | * |
| 3 | * | * | * | * |
| 4 | * | * | * | * |
| 5 | * | * | | |
| 6 | * | * | * | * |
| 7 | * | * | * | * |
| 8 | * | * | | * |
| 9 | * | * | * | |
| 10 | * | | * | * |
| 11 | * | + | | * |
| 12 | * | | | |
| 13 | * | + | * | * |
| 14 | * | | * | |
| 15 | * | * | * | |
| 16 | * | * | | |
| 17 | * | | * | * |
| 18 | * | | * | * |
| 19 | * | * | | * |
| 20 | * | | * | * |
| 21 | * | * | * | * |
| 22 | | | | |
| 23 | * | | * | * |
| 24 | * | | | |
| 25 | * | | | * |
| 26 | * | * | | * |
| 27 | * | * | | |
| 28 | * | * | * | |
| 29 | * | + | * | * |
| 30 | * | * | * | * |
| 31 | * | * | * | * |
| 32 | * | | | |

From Table 8 it is clear that all items for which the model of double monotonicity held also conformed to the model of monotone homogeneity ($H \geq 0$). In the set of Rasch items, however, the behavior on five items (Items 8, 11, 19, 25, and 26) was not explained by the model of double monotonicity. It is theoretically impossible that Rasch items do not conform to the model of double

monotonicity. The misfit of these five items can thus be explained by properties of the methods for checking the goodness-of-fit of the three models, such as the power of tests, and the arbitrariness of some decisions based on statistics.

With respect to the methods for evaluating the fit of the three models, a problem was caused by the way the items of the Verbal Analogies Test are constructed. The tasks that have to be performed to solve an item correctly do not vary systematically across items. This means that it was not possible to create subsets of items that might measure hypothetically different attributes. Consequently, statistical indicators had to be relied on for the analyses; item content was disregarded.

The results of the stepwise selection of items for the model of monotone homogeneity ($H \geq .3$) indicate that this additional restriction yielded sets of items that discriminate better among persons, in comparison with sets of items selected by using $H = 0$ as a lower bound. As Table 8 shows, if $H \geq .3$, fewer items form one scale; this table also shows that $H \geq .3$ may result in the rejection of some items that are included in the final Rasch scale.

The Rasch model has interesting measurement properties, such as specifically-objective measurement and measurement of items and persons on a metric scale. However, much of the research has shown that items do not easily fit this model. The present Rasch analysis also showed that many items did not fit the model, whereas the behavior on almost all items was explained by the model of monotone homogeneity.

In many testing applications, it often suffices to know the order of persons on an attribute (e.g., in selection problems). Therefore, the Mokken model of monotone homogeneity seems to be an attractive model for two reasons. First, ordinal measurement of persons is guaranteed when the model applies to the data. Second, the model is not as restrictive with respect to empirical data as are the Mokken model of double monotonicity and the Rasch model. If, in addition, an invariant ordering of items is required for all examinees (e.g., in intelligence testing), the model of double monotonicity may be appropriate. More sophisticated applications, however, such as equating, item banking, and adaptive testing, preferably require measurement on metric scales. If such applications are envisaged, response behavior must comply with the demands of parametric models such as the Rasch model.

## References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika, 42,* 375–401.

Debets, P., Sijtsma, K., Brouwer, E., & Molenaar, I. W. (1989). MSP: A computer program for item analysis according to a nonparametric IRT approach. *Psychometrika, 54,* 534–536.

Drenth, P. J. D., & van Wieringen, P. (1969). *Verbale Aanleg Test* [*Verbal Ability Test*]. Amsterdam: Swets & Zeitlinger.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [*Introduction to psychological test theory*]. Bern: Huber.

Fischer, G. H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika, 52,* 565–587.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* Unpublished doctoral dissertation, Universiteit Twente, Enschede.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Gustafsson, J. E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program.* Institute of Education, University of Göteborg, Sweden.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

de Jong-Gierveld, J., & Kamphuis, F. (1985). The development of a Rasch-type loneliness scale. *Applied Psychological Measurement, 9,* 289–299.

Kingma, J., & Ten Vergert, E. M. (1985). A nonparametric scale analysis of the development of conservation. *Applied Psychological Measurement, 9,* 375–387.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Martin-Löf, P. (1973). *Statistika Modeller. Anteckningar från seminarier Läsaret 1969-70 utarbetade av Rolf Sundberg, 2: a uppl.* [*Statistical models. Notes from seminars 1969-70 by Rolf Sundberg, 2nd ed.*]. Stockholm: Institute för Försäkrungsmatematik och Matematisk Statistik vid Stockholms Universitet.

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

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken Scale: A critical dicussion." *Applied Psychological Measurement, 10,* 279–285.

Molenaar, I. W. (1982). Een tweede weging van de Mokken schaal [A second weighing of the Mokken scaling procedure]. *Tijdschrift voor Onderwijsresearch, 7,* 172–181.

Molenaar, I. W. (1983a). Rasch, Mokken en schoolbeleving [Rasch, Mokken and school experience]. In S. Lindenberg & F. N. Stokman (Eds.), *Modellen in de sociologie.* Deventer: Van Loghum Slaterus.

Molenaar, I. W. (1983b). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48,* 49–72.

Molenaar, I. W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën [An exercise in item response theory for three ordered response categories]. In G. F. Pikkemaat & J. J. A. Moors (Eds.), *Liber Amicorum Jaap Muilwijk.* Groningen: Econometrisch Instituut.

Niemöller, B., & van Schuur, W. H. (1980). *Mokken test. STAP user's manual (Volume 4).* Amsterdam: University of Amsterdam Press.

Niemöller, B., & van Schuur, W.H. (1983). Stochastic models for unidimensional scaling: Mokken and Rasch. In D. McKay, N. Schofield, & P. Whiteley (Eds.), *Data analysis and the social sciences.* London: Frances Pinter Publications.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.

Schmitt, N. (1981). Rasch analysis of the Central Life Interest Measure. *Applied Psychological Measurement, 5,* 3–10.

Siegel, S. (1956). *Nonparametric statistics.* New York: McGraw Hill.

Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory.* Amsterdam: Free University Press.

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

Sijtsma, K., & Prins, P. M. (1986). Itemselectie in het Mokken model [Item selection in the Mokken model]. *Tijdschrift voor Onderwijsresearch, 11,* 121–129.

Stokman, F. N. (1977). *Roll calls and sponsorship: A methodological analysis of third world group formation in the United Nations.* Leyden: Sijthoff.

Wainer, H., Morgan, A., & Gustafsson, J. E. (1980). A review of estimation procedures for the Rasch model with an eye toward longish tests. *Journal of Educational Statistics, 5,* 35–64.

van den Wollenberg, A. L. (1979). *The Rasch model and time-limit tests.* Nijmegen: Stichting Studentenpers Nijmegen.

van den Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6,* 83–91.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

## Author's Address

Send requests for reprints or further information to Rob R. Meijer, Vrije Universiteit, Vakgroep Arbeids en Organisatiepsychologie, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands.