PSYCHOMETRIKA—VOL. 73, NO. 2, 183–208
JUNE 2008
DOI: 10.1007/s11336-007-9034-z

MOKKEN SCALE ANALYSIS FOR DICHOTOMOUS ITEMS USING MARGINAL MODELS

L. ANDRIES VAN DER ARK, MARCEL A. CROON, AND KLAAS SIJTSMA

TILBURG UNIVERSITY

Scalability coefficients play an important role in Mokken scale analysis. For a set of items, scalability coefficients have been defined for each pair of items, for each individual item, and for the entire scale. Hypothesis testing with respect to these scalability coefficients has not been fully developed. This study introduces marginal modelling as a framework to derive the standard errors for the scaling coefficients and test hypotheses about these coefficients. Several examples demonstrate the possibilities of marginal modelling in Mokken scale analysis. These possibilities include testing whether Mokken's criteria for a scale are satisfied, testing whether scalability coefficients of different items are equal, and testing whether scalability coefficients are equal across different groups.

Key words: marginal models, Mokken scale analysis, scalability coefficients, test construction.

1. Introduction

Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002) is used for scaling items and measuring respondents on an ordinal scale. Mokken scale analysis consists of two parts. The first part is the evaluation of a set of items with ordered scores as a scale according to particular scaling criteria that are related to the monotone homogeneity model (Mokken, 1971; Sijtsma & Molenaar, 2002). This can be done in a confirmatory way for a set of items that are hypothesized to form a scale or in an exploratory way when an experimental set of items is analyzed to find out whether they constitute one or more scales. When none of the items satisfy the criteria of Mokken scale analysis, the result is that no scales can be constructed, but it happens more frequently that one or a few items in the set are unscalable whereas the majority of the items is scalable. The unscalable items are left out of the analysis. The scales that are produced by Mokken scale analysis are referred to as *Mokken scales*. The second part of Mokken scale analysis takes the scales found in the first part, and investigates several other interesting properties of the monotone homogeneity model that were not assessed explicitly in the first part of the analysis. This second part does not play a role in this study. Mokken scale analysis can be conducted using the stand-alone software package MSP5.0 for Windows (Molenaar & Sijtsma, 2000) and the R package mokken.0.2 (Van der Ark, 2007).

Mokken scales are defined by means of scalability coefficients (Mokken, 1971, pp. 148–153). The first part of Mokken scale analysis involves the testing of hypotheses about these scalability coefficients and the evaluation of their numerical values. The hypotheses involve testing whether scalability coefficients satisfy the criteria for a Mokken scale (Mokken, 1971, p. 184), and testing whether scalability coefficients are equal across items or across groups. We demonstrate that currently available methods do not allow us to test several interesting hypotheses about the scaling coefficients that are relevant in Mokken scale analysis, and we propose to use the *marginal modelling* framework for this purpose and also for testing hypotheses for which other solutions already exist (Mokken, 1971).

Requests for reprints should be sent to L. Andries van der Ark, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: a.vdark@uvt.nl

The paper is organized as follows. First, the principles of marginal modelling are explained. Second, Mokken scale analysis is discussed, including the monotone homogeneity model, the scalability coefficients, and the definition of a scale. Third, the scalability coefficients are discussed and it is shown how these coefficients can be reformulated so that they can be incorporated in marginal models. For the sake of readability, several important but rather cumbersome derivations have been diverted to appendices. Fourth, we give an overview of relevant hypotheses in Mokken scale analysis and we show how these hypotheses can be tested using marginal models. As an example, the marginal models were applied to data from a cognitive balance-task test (Van Maanen, Been, & Sijtsma, 1989). Fifth, the strengths and weaknesses of the marginal modelling approach are discussed, and recommendations are given for its practical use and for future improvements.

2. Marginal Models

Assume that a test consists of J dichotomously scored items, indexed by j and i. The random variable representing the scores on item j is denoted by X_j, and its realization by x_j ($x_j \in \{0, 1\}$). A vector containing the J item-score variables is denoted $(X_1, X_2, \ldots, X_J)$. The total score on the test is denoted by $X_+ = \sum_{j=1}^{J} X_j$. The popularity or the easiness of an item is defined as the probability that a randomly drawn respondent from the population of interest endorses a positively worded statement or answers an item correctly, respectively, and is denoted by π_j^1. The probability that a randomly drawn respondent does not endorse a positively worded statement or answers an item incorrectly, is denoted by π_j^0. The joint probability of scores on X_i and X_j is denoted by π_{ij}^{uv} [$u, v = 0, 1$; π_{ij}^{uv} can assume values for four different score pairs: $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$]. Without loss of generality, the items are ordered by decreasing popularity or easiness and numbered accordingly, such that

$$\pi_1^1 \geq \pi_2^1 \geq \cdots \geq \pi_J^1. \tag{1}$$

Equation (1) arbitrarily defines the most popular item to be item 1, the next popular item to be item 2, and so on. Equation (1) does not in any way restrict the data. Finally, the test data can be collected in a J-dimensional contingency table with $L = 2^J$ cells.

Consider the example in Table 1 (upper left-hand panel), which shows the cross classification of $J = 2$ items in a two-way contingency table. The observed frequencies in the contingency table are denoted by n_{ij}^{uv} ($u, v = 0, 1$) and the marginal frequencies are denoted by n_i^u, n_j^v, and n. Assuming a fixed sample size n, let m_{ij}^{uv} be the theoretically expected frequency satisfying $m_{ij}^{uv} = n \times \pi_{ij}^{uv}$ ($u, v = 0, 1$), with marginal frequencies m_i^u, m_j^v, and $m = n$. Sample estimates of m_{ij}^{uv} and π_{ij}^{uv} are denoted by $\widehat{m}_{ij}^{uv}$ and $\widehat{\pi}_{ij}^{uv}$, respectively. Without any constraints imposed upon the data, $\widehat{m}_{ij}^{uv} = n_{ij}^{uv}$ and $\widehat{\pi}_{ij}^{uv} = n_{ij}^{uv}/n$. In Table 1 (upper left-hand panel), $\widehat{\pi}_i^1 = 58/178 = 0.33$ and $\widehat{\pi}_j^1 = 44/178 = 0.25$. Because $\widehat{\pi}_i^1 > \widehat{\pi}_j^1$, item i is assumed to be more popular than item j in the population. The order of the indices i and j in the subscripts of, for example, n_{ij}^{uv}, in general indicates that in the sample item i is more popular than item j.

Marginal models for categorical data (Bartolucci & Forcina, 2002; Bartolucci, Forcina, & Dardanoni, 2001; Bergsma, 1997a; Bergsma & Rudas, 2002; Lang & Agresti, 1994; Rudas & Bergsma, 2004) constitute a family of models that impose restrictions on certain marginals (i.e., subsets) of contingency tables. These restrictions can have several forms. To illustrate this, we take the contingency table in the upper left-hand panel of Table 1 as a starting point.

The first example of a marginal model imposes equality constraints on two cell frequencies by hypothesizing that $\pi_{ij}^{00} = \pi_{ij}^{11}$. Estimation of this marginal model of equal diagonal probabil-

TABLE 1.

Example of a contingency table with observed frequencies for a dichotomous item pair (upper left-hand panel), the estimated expected frequencies under a marginal model of equal diagonal probabilities (upper right-hand panel), the estimated expected frequencies under a marginal model of homogeneous item popularity (lower left-hand panel), and the estimated expected frequencies under a marginal model with $\gamma = .8$ (lower right-hand panel).

		Item j		Total
		0	1	
Item i	0	102	18	120
	1	32	26	58
Total		134	44	178

		Item j		Total
		0	1	
Item i	0	64	18	82
	1	32	64	96
Total		96	82	178

		Item j		Total
		0	1	
Item i	0	102	25	127
	1	25	26	51
Total		127	51	178

		Item j		Total
		0	1	
Item i	0	106.684	13.316	120.000
	1	27.316	30.684	58.000
Total		134.000	44.000	178

ities yields estimated expected frequencies $\widehat{m}_{ij}^{uv}$ that are as close as possible to the observed frequencies n_{ij}^{uv} (e.g., using a maximum likelihood or least-squares criterion) but with $\widehat{m}_{ij}^{00} = \widehat{m}_{ij}^{11}$. Table 1 (upper right-hand panel) shows the maximum likelihood estimates of the expected frequencies.

Throughout the paper we assume a multinomial sampling distribution that has the effect of reproducing the sample size n (here $m = n = 178$) in the marginal model. The fit of the marginal model is evaluated by comparing the observed and expected frequencies using commonly known fit statistics for contingency tables such as the likelihood ratio statistic, G^2 (see Appendix A). Let C denote the number of nonredundant constraints on the frequencies in the contingency table. For large n, G^2 approaches a chi-square distribution with C degrees of freedom ($df = C$). In the first example, it may be verified that $G^2 = 64.352$; because there is one nonredundant constraint (i.e., $m_{ij}^{00} - m_{ij}^{11} = 0$), it follows that $df = 1$ and, as a result, $p < .0001$.

The second example of a marginal model imposes equality constraints on the marginal frequencies in Table 1 by hypothesizing that $\pi_i^1 = \pi_j^1$, which implies $\pi_i^0 = \pi_j^0$. Estimation of this marginal model of homogeneous item popularity yields estimated expected frequencies $\widehat{m}_{ij}^{uv}$ such that $\widehat{m}_i^0 = \widehat{m}_j^0$ and $\widehat{m}_i^1 = \widehat{m}_j^1$. Table 1 (lower left-hand panel) shows the maximum likelihood estimates of the expected frequencies. It may be verified that $G^2 = 3.973$; because there is one nonredundant constraint (i.e., $m_i^0 - m_j^0 = 0$), it follows that $df = 1$ and, as a result, $p = .0462$.

The third example of a marginal model imposes equality constraints on functions of the cell frequencies in Table 1 by restricting Goodman and Kruskal's (1954) γ coefficient to a value that is hypothesized between two variables in a particular study. This application is interesting because it allows us to illustrate marginal modelling in greater detail than the previous, simpler examples. Coefficient γ can be written as a function of the expected cell frequencies,

$$\gamma = \frac{m_{ij}^{00} m_{ij}^{11} - m_{ij}^{01} m_{ij}^{10}}{m_{ij}^{00} m_{ij}^{11} + m_{ij}^{01} m_{ij}^{10}}.$$

Bergsma and Croon (2005) described several interesting restrictions on γ that can be estimated using marginal models. A simple restriction is the arbitrary equality constraint $\gamma = .8$. For this marginal model the expected frequencies m_{ij}^{uv} $(u, v = 0, 1)$ are estimated under the constraint that $\gamma = .8$. Table 1 (lower right-hand panel) shows the maximum likelihood estimates of the expected frequencies. It may be verified that $G^2 = 3.207$; because there is one nonredundant constraint (i.e., $\gamma - 0.8 = 0$), it follows that $df = 1$ and, as a result, $p = .0733$.

In general, marginal models can be applied to multiway contingency tables with L cells. Let $\mathbf{n}$ be the $(L \times 1)$ vector of observed frequencies in the contingency table, and let $\mathbf{m}$ be the $(L \times 1)$ vector of expected frequencies given the marginal model. It is assumed that the order of the elements in both $\mathbf{n}$ and $\mathbf{m}$ corresponds to the following ordering of the item-score patterns collected in the $L \times J$ matrix $\mathbf{R}$, defined as

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 & 1 \\ 0 & 0 & 0 & \ldots & 1 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 1 & 0 & 1 \\ 0 & 0 & 0 & \ldots & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \ldots & 1 & 1 & 0 \\ 1 & 1 & 1 & \ldots & 1 & 1 & 1 \end{pmatrix}. \tag{2}$$

Given the ordering with respect to popularity or easiness in equation (1), the scores in the first column of $\mathbf{R}$ correspond to the most popular item, the scores in the second column to the next most popular item, and so on, and the scores in the last column correspond to the least popular item. Suppose that the marginal model consists of C nonredundant equality constraints, which are functions of $\mathbf{m}$. The first inequality constraint is denoted by $g_1(\mathbf{m})$, the second by $g_2(\mathbf{m})$, and the last by $g_C(\mathbf{m})$. Setting each function equal to zero yields $g_1(\mathbf{m}) = 0, g_2(\mathbf{m}) = 0, \ldots, g_C(\mathbf{m}) = 0$. In vector notation these equality constraints can be written as

$$\mathbf{g}(\mathbf{m}) = \begin{pmatrix} g_1(\mathbf{m}) \\ \vdots \\ g_C(\mathbf{m}) \end{pmatrix} = \mathbf{0}. \tag{3}$$

For the first example with respect to equal diagonal probabilities (Table 1, upper right-hand panel), equation (3) equals $\mathbf{g}(\mathbf{m}) = g_1(\mathbf{m}) = m_{ij}^{00} - m_{ij}^{11} = 0$; for the second example with respect to homogeneous item popularity (Table 1, lower left-hand panel), equation (3) equals $\mathbf{g}(\mathbf{m}) = g_1(\mathbf{m}) = m_i^1 - m_j^1 = (m_{ij}^{10} + m_{ij}^{11}) - (m_{ij}^{01} + m_{ij}^{11}) = m_{ij}^{10} - m_{ij}^{01} = 0$; and for the third example that imposes restriction $\gamma = .8$ (Table 1, lower right-hand panel), equation (3) equals $\mathbf{g}(\mathbf{m}) = g_1(\mathbf{m}) = (m_{ij}^{00} m_{ij}^{11} - m_{ij}^{01} m_{ij}^{10})/(m_{ij}^{00} m_{ij}^{11} + m_{ij}^{01} m_{ij}^{10}) - .8 = 0$.

Bergsma (1997b) developed syntax for Mathematica (Wolfram, 1999) that produces maximum likelihood estimates and asymptotic standard errors for $\mathbf{m}$. In the process of maximum likelihood estimation, the Jacobian of $\mathbf{g}(\mathbf{m})$ with respect to $\log(\mathbf{m})$ must be computed (see Appendix A). For different marginal models this Jacobian can have very different forms. Bergsma (1997a, p. 66) proposed to write the constraints in equation (3) in a single general matrix formula using a *recursive exp-log notation* (see also Kritzer, 1977). Once written in recursive exp-log notation, the derivation of the Jacobian is straightforward (Bergsma, 1997a, p. 68; see also Appendix A), and a simple recursive algorithm, which can be easily implemented in software, suffices to compute the Jacobian irrespective of the marginal model.

Given that $\mathbf{A}_1, \ldots, \mathbf{A}_q$ are q design matrices, the general form of the recursive exp-log notation of a marginal model is

$$\mathbf{g}(\mathbf{m}) = \mathbf{A}_q \exp\bigl(\mathbf{A}_{q-1} \log\bigl(\mathbf{A}_{q-2} \ldots \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})\bigr)\bigr)\bigr). \tag{4}$$

For a particular marginal model, the appropriate design matrices must be derived in order to write $\mathbf{g}(\mathbf{m})$ in a recursive exp-log notation. There are no explicit rules for deriving design matrices and the same marginal model can often be written in different recursive exp-log notations. Finding the most parsimonious recursive exp-log notation may require some effort.

For the three examples of marginal models in Table 1, the expected frequencies are collected in the vector $\mathbf{m} = (m_{ij}^{00}, m_{ij}^{01}, m_{ij}^{10}, m_{ij}^{11})^{\mathrm{T}}$ (the superscript T denotes the transpose). The first example concerning equal diagonal probabilities has one design matrix, which is $\mathbf{A}_1 = (1\ 0\ 0\ -1)$, and the recursive exp-log notation of the model constraints in equation (3) is equal to

$$\mathbf{g}(\mathbf{m}) = g_1(\mathbf{m}) = \mathbf{A}_1 \mathbf{m} = (1\ 0\ 0\ -1) \begin{pmatrix} m_{ij}^{00} \\ m_{ij}^{01} \\ m_{ij}^{10} \\ m_{ij}^{11} \end{pmatrix} = m_{ij}^{00} - m_{ij}^{11} = 0.$$

The second example with respect to homogeneous item popularity also has one design matrix, which is $\mathbf{A}_1 = (0\ 1\ -1\ 0)$. The recursive exp-log notation of equation (3) is $\mathbf{A}_1 \mathbf{m} = \mathbf{0}$, which results in $m_{ij}^{01} - m_{ij}^{10} = m_i^1 - m_j^1 = 0$.

For the third example that imposes $\gamma = .8$ upon the table, the design matrices were derived by Bergsma and Croon (2005), who showed that $\gamma = \mathbf{A}_5 . \exp(\mathbf{A}_4 . \log(\mathbf{A}_3 . \exp(\mathbf{A}_2 . \log(\mathbf{A}_1 . \mathbf{m}))))$, with

$$\mathbf{A}_1 = \mathbf{I}_{4\times 4}, \qquad \mathbf{A}_2 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \qquad \mathbf{A}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix},$$

$$\mathbf{A}_4 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{A}_5 = (1 \quad -1).$$

Hence the recursive exp-log notation of equation (3) is

$$\mathbf{g}(\mathbf{m}) = g_1(\mathbf{m}) = \mathbf{A}_5 . \exp\bigl(\mathbf{A}_4 . \log\bigl(\mathbf{A}_3 . \exp\bigl(\mathbf{A}_2 . \log(\mathbf{A}_1 . \mathbf{m})\bigr)\bigr)\bigr) - 0.8 = 0.$$

In Appendix A it is shown how maximum likelihood estimates of $\mathbf{m}$ are obtained subject to the constraints in equation (3), when these constraints are written in the recursive exp-log notation of equation (4).

3. Mokken Scale Analysis

The main purpose of this study is to use marginal models and the recursive exp-log notation to test hypotheses about scalability coefficients in the context of Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002). Before we explain this application, subsequently we introduce the monotone homogeneity model, the scalability coefficients, the relationships between the monotone homogeneity model and the scalability coefficients, the definition of a scale, two types of Mokken scale analysis, and some existing results for the distribution of the scalability coefficients.

3.1. The Monotone Homogeneity Model

The monotone homogeneity model (Mokken, 1971, Chap. 4; Sijtsma & Molenaar, 2002, pp. 22–23; Sijtsma & Meijer, 2007) is a nonparametric item response theory (IRT) model for ordinal person measurement (related theory was developed, e.g., by Molenaar, 1997; Ramsay, 1991; Scheiblechner, 2007; and Stout, 1990). Before we discuss the assumptions of this model, first we introduce some notation. Let θ denote the latent variable underlying performance on each of the items in the test. Let the probability of obtaining score x_j on item j be denoted by $P(X_j = x_j|\theta)$. This conditional response probability is known as the item response function (IRF). Further, let the joint probability of a particular score pattern on the J items in the test be denoted by $P(X_1 = x_1, \ldots, X_J = x_J|\theta)$. The monotone homogeneity model is based on the following three assumptions.

Unidimensionality. The responses to the items are driven by a unidimensional latent variable denoted θ.

Local Independence. The joint distribution of the item scores conditional on θ can be written as the product of the J conditional marginal distributions: $P(X_1 = x_1, \ldots, X_J = x_J|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta)$.

Monotonicity. As latent variable θ increases, the probability of a positive response to an item increases or stays the same across intervals of θ; that is, for two values of θ, say, θ_a and θ_b, and arbitrarily assuming that $\theta_a < \theta_b$, monotonicity means that $P(X_j = 1|\theta = \theta_a) \leq P(X_j = 1|\theta = \theta_b)$ for $j = 1, \ldots, J$.

For dichotomous items, the monotone homogeneity model implies the stochastic ordering of latent variable θ by total score X_+; that is, for an arbitrary value t of θ, the probability $P(\theta > t|X_+ = x_+)$ is nondecreasing in x_+ (Hemker, Sijtsma, Molenaar, & Junker, 1997; also, see Grayson, 1988). This property guarantees an ordinal person scale: Persons with higher X_+ scores on average have higher θ values.

Mokken (1971, pp. 119–120) showed that for a J-item test the monotone homogeneity model implies that all interitem covariances or, equivalently, all interitem product-moment correlations, are nonnegative. Let σ_{ij} denote the covariance between items i and j; then, the monotone homogeneity model implies

$$\sigma_{ij} \geq 0 \quad \text{for all } i < j. \tag{5}$$

Equation (5) is used throughout. Nonnegative interitem covariance is a special case of a more general interitem covariance result, known as conditional association, and proven to be true by Holland and Rosenbaum (1986) under more general conditions—multidimensional latent variables and continuous item scores, and local independence and monotonicity adapted to these conditions. In Holland and Rosenbaum's (1986) conditional association framework, nonnegative interitem covariance in equation (5) is referred to as pairwise nonnegative association (Ellis & Van den Wollenberg, 1993). Other observable consequences, such as manifest monotonicity (Junker & Sijtsma, 2000), can be used to test the monotonicity assumption, but like conditional association (except pairwise nonnegative association) they do not play a role in this study.

3.2. Scalability Coefficients

The Guttman (1950) model is the basis of the scalability coefficients H_{ij}, H_j, and H (Mokken, 1971; cf. Loevinger, 1948). Given an ordering of the J items according to decreasing popularity (equation (1)), the Guttman model assumes that a respondent who endorses the less popular item in a pair of items also endorses the more popular item. Thus, if $\pi_i^1 > \pi_j^1$, for any respondent the Guttman model excludes the item-score pattern $(X_i, X_j) = (0, 1)$. This item-score pattern is called a *Guttman error*, and the other three item-score patterns [(0, 0), (1, 0), and

(1, 1)] are called *conformal patterns*. Data that do not contain Guttman errors are in agreement with the Guttman model.

In a 2×2 contingency table for the scores on items i and j (with $\pi_i^1 > \pi_j^1$ and sample of size n), the expected number of Guttman errors, denoted F_{ij}, equals $F_{ij} = n \times \pi_{ij}^{01}$, and the expected number of Guttman errors under marginal independence, denoted by E_{ij}, equals $E_{ij} = n \times \pi_i^0 \times \pi_j^1$. The scalability coefficient for items i and j, denoted by H_{ij}, is computed from

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}} = 1 - \frac{\pi_{ij}^{01}}{\pi_i^0 \times \pi_j^1} = 1 - \frac{n \times m_{ij}^{01}}{m_i^0 \times m_j^1}. \tag{6}$$

For the example in Table 1 (upper left-hand panel), $\widehat{F}_{ij} = 18$ and $\widehat{E}_{ij} = 29.663$, yielding $\widehat{H}_{ij} = .3932$. To facilitate its interpretation, coefficient H_{ij} can be written as a normed covariance (e.g., Sijtsma & Molenaar, 2002, p. 55). Let $\sigma_{ij}^{\max}$ be the maximum covariance between items i and j, given the marginal distributions of X_i and X_j. Given that items i and j have positive variance, equation (6) is equal to

$$H_{ij} = \frac{\sigma_{ij}}{\sigma_{ij}^{\max}}. \tag{7}$$

The scalability coefficient for an individual item j, denoted H_j, $j = 1, \ldots, J$ (Mokken, 1971, p. 151), is defined as

$$H_j = 1 - \frac{\sum_{i \neq j} F_{ij}}{\sum_{i \neq j} E_{ij}} = 1 - \frac{n\left(\sum_{i=1}^{j-1} m_{ij}^{01} + \sum_{i=j+1}^{J} m_{ji}^{01}\right)}{\sum_{i=1}^{j-1} m_i^0 m_j^1 + \sum_{i=j+1}^{J} m_j^0 m_i^1}. \tag{8}$$

Coefficient H_j can also be written in terms of interitem covariances and corresponding maximum covariances, given the marginal distributions of the item scores, as

$$H_j = \frac{\sum_{i \neq j} \sigma_{ij}}{\sum_{i \neq j} \sigma_{ij}^{\max}}.$$

Let rest score $R_{(j)}$ be defined as the total score on $J - 1$ items excluding item j, then one can also write (Sijtsma & Molenaar, 2002, p. 57)

$$H_j = \frac{\sigma_{X_j R_{(j)}}}{\sigma_{X_j R_{(j)}}^{\max}}. \tag{9}$$

Equation (9) shows that coefficient H_j expresses the strength of the relationship between item j and the other items in the test, comparable with a regression coefficient in a regression model.

For a set of J items, Mokken (1971, p. 149) proposed the total-scale coefficient H, which is defined as

$$H = 1 - \frac{\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} F_{ij}}{\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} E_{ij}} = 1 - \frac{n\left(\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} m_{ij}^{01}\right)}{\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} m_i^0 m_j^1}. \tag{10}$$

Coefficient H can also be written in terms of interitem covariances and item rest-score covariances, which results in

$$H = \frac{\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \sigma_{ij}}{\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \sigma_{ij}^{\max}} = \frac{\sum_{j=1}^{J} \sigma_{X_j R_{(j)}}}{\sum_{j=1}^{J} \sigma_{X_j R_{(j)}}^{\max}}.$$

If the data obey a perfect Guttman scalogram, $H = 1$, but this value is never found in practice.

Sijtsma and Molenaar (2002, Theorem 4.2; see also Hemker, Sijtsma, & Molenaar, 1995) showed that H_{ij}, H_j, and H are related such that

$$\min_{i,j}(H_{ij}) \leq \min_{j}(H_j) \leq H \leq \max_{j}(H_j) \leq \max_{i,j}(H_{ij}). \tag{11}$$

3.3. Relationships Between the Monotone Homogeneity Model and the Scalability Coefficients

The monotone homogeneity model implies observable consequences with respect to the scalability coefficients H_{ij}, H_j, and H. These observable consequences are used in data analysis to investigate whether the data support the fit of the monotone homogeneity model (Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Meijer, 2007).

In particular, Mokken (1971, pp. 148–153; see also Sijtsma & Molenaar, 2002, Theorem 4.3) showed that the monotone homogeneity model implies that

$$\begin{aligned} &0 \leq H_{ij} \leq 1 \quad \text{for all } i < j, \\ &0 \leq H_j \leq 1 \quad \text{for all } j, \quad \text{and} \\ &0 \leq H \leq 1. \end{aligned} \tag{12}$$

Thus, negative scalability coefficients are in conflict with the monotone homogeneity model. These observable consequences are the basis of Mokken scale analysis.

3.4. Definition of a Scale and Two Types of Mokken Scale Analysis

3.4.1. Definition of a Scale. A set of items is a scale (Mokken, 1971, p. 184; Molenaar & Sijtsma, 2000; Sijtsma & Molenaar, 2002, p. 68), in this study called a Mokken scale if, for product-moment correlation ρ, and for any constant value $0 < c \leq 1$,

$$\rho_{ij} > 0 \text{ (or, equivalently, } H_{ij} > 0) \quad \text{for all } i < j, \quad \text{and} \tag{13}$$

$$H_j \geq c > 0 \quad \text{for all } j. \tag{14}$$

Equation (13) is the first criterion of a Mokken scale, and equation (14) is the second criterion of a Mokken scale. Compared to equations (5) and (12), strict inequality is not crucial here due to continuity of the scales of ρ and H_j. Except for the strict inequalities, the monotone homogeneity model implies both equation (13) and $H_j > 0$ (which is part of equation (14)).

However, the monotone homogeneity model does not imply a specific positive values of c. Thus, the inclusion of positive c in the definition of a Mokken scale can be a source of confusion and needs to be explained. To understand the role of positive c, one may note that the monotone homogeneity model, and special cases of this model such as the one-, two-, and three-parameter logistic models, allow items in a scale which have (nearly) flat IRFs. Such items contribute little, if anything, to a reliable person ordering and may even attenuate the reliability of this ordering; thus, these items are unwanted in a scale. The inclusion of a positive c in the definition of a Mokken scale prevents the selection of such items in a scale by rejecting items with H_js which are smaller than c. Thus, Mokken scale analysis aims to produce "high-quality" scales, the definition of which depends on the researcher's choice of lower bound c.

Mokken (1971, p. 184) proposed to always set c at least to .3. One may note that equation (11) implies that $H \geq \min_j(H_j)$; thus, for lower bound $c = .3$, the total-scale $H \geq .3$. The choice of c controls the quality of the individual items in the scale and of the total scale and, therefore, of the total-scale score X_+ for ordering persons on latent variable θ. Mokken (1971, p. 185) proposed the following rules of thumb for the interpretation of H. A set of items is unscalable for all practical purposes if $H < .3$; and a scale is considered weak if $.3 \leq H < .4$, moderate if $.4 \leq H < .5$, and strong if $H \geq .5$.

3.4.2. Two Types of Mokken Scale Analysis. Mokken scale analysis can have two forms (Mokken, 1971, pp. 187–199). The first possibility is that the researcher evaluates a given set of J items with respect to the definition of a scale for a chosen value of c. This is confirmatory Mokken scale analysis. The second possibility is to use an automated item selection algorithm (Mokken, 1971, pp. 190–199; Sijtsma & Molenaar, 2002, Chap. 5). This algorithm selects items one by one to obtain one or more scales (depending on the data structure) that agree with the definition of a Mokken scale. In each selection step, the item is chosen from the items not already selected, that not only agrees with equations (13) and (14) but also produces the greatest total-scale H coefficient with the items already selected in previous steps. This is exploratory Mokken scale analysis.

In the remainder of this paper we discuss the use of marginal modelling for testing hypotheses about the scalability coefficients. The term Mokken scale analysis refers to the use of scalability coefficients for scale construction both in a confirmatory and in an exploratory context.

3.5. Results for the Distribution of the Scalability Coefficients

Results for the distribution of the scalability coefficients are available for the null case (which refers to the null hypothesis that $H = 0$) and the nonnull case (which refers to the null hypothesis that $H = w$, w is some positive constant) (Mokken, 1971, pp. 160–169). Results for the null case Mokken (1971, pp. 160–164) are the following. Let S_{ij} be the sample covariance of items i and j, and let S_i and S_j be the sample standard deviations of items i and j, respectively; then for large n, in the null case, the statistics

$$Z_{ij} = \frac{S_{ij}}{S_i S_j}\sqrt{n-1},$$

$$Z_j = \frac{\sum_{i\neq j} S_{ij}}{S_j \sum_{i\neq j} S_i}\sqrt{n-1},$$

and

$$Z = \frac{\sum_{i=1}^{J-1}\sum_{j=i+1}^{J} S_{ij}}{\sum_{i=1}^{J-1}\sum_{j=i+1}^{J} S_i S_j}\sqrt{n-1},$$

converge to a standard normal distribution. In the available software for Mokken scale analysis, $H_{ij} = 0$ is tested against the alternative that $H_{ij} > 0$ to decide whether items satisfy the first criterion of a Mokken scale that $\rho_{ij} > 0$ (equation (13)). Results for the nonnull case yield asymptotic standard errors for $\widehat{H}$ (Mokken, 1971, pp. 164–169). These results are not available in current software for Mokken scale analysis.

4. A Marginal Modelling Approach to the Scalability Coefficients

Coefficient H_{ij} can be written in the recursive exp-log notation, which is useful for testing hypotheses involving H_{ij}. Let $\mathbf{m} = (m_{ij}^{00}, m_{ij}^{01}, m_{ij}^{10}, m_{ij}^{11})^{\mathrm{T}}$, and let $\mathbf{A}_1$ and $\mathbf{A}_2$ be the following design matrices:

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{A}_2 = (1 \quad -1 \quad -1 \quad 1).$$

TABLE 2.

Estimated expected frequencies for the data in Table 1 under the marginal model imposing $H_{ij} = .5$ on the table.

		Item j		Total
		0	1	
Item i	0	103.716	14.360	118.074
	1	30.990	28.935	59.924
Total		134.706	43.294	178

Then, H_{ij} in equation (6) equals

$$H_{ij} = 1 - \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})\bigr). \tag{15}$$

This can be verified by writing the term $\log(\mathbf{A}_1\mathbf{m})$ in equation (15) as

$$\log(\mathbf{A}_1\mathbf{m}) = \log\left[\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} m_{ij}^{00} \\ m_{ij}^{01} \\ m_{ij}^{10} \\ m_{ij}^{11} \end{pmatrix}\right] = \log\begin{pmatrix} n \\ m_i^0 \\ m_j^1 \\ m_{ij}^{01} \end{pmatrix},$$

and noting that

$$\exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m})\bigr) = \exp\left[\begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \log\begin{pmatrix} n \\ m_i^0 \\ m_j^1 \\ m_{ij}^{01} \end{pmatrix}\right] = \frac{n \times m_{ij}^{01}}{m_i^0 \times m_j^1}.$$

In the case of J items, there are $K = \frac{1}{2}J(J-1)$ item pairs; hence, there are K coefficients H_{ij}. The recursive exp-log notation for the vector $\mathbf{H}_{ij} = (H_{12}, H_{13}, \ldots, H_{J-1,J})^{\mathrm{T}}$ containing all K item-pair coefficients H_{ij} $(i < j)$ is derived in Appendix C.

Based on previous results, a researcher may have reason to believe that for two particular key items in a test $H_{ij} = w$, with $0 \leq w < 1$, and (s)he may wish to test this hypothesis on a sample from another population. Using the recursive exp-log notation for H_{ij} (equation (15)), it may be verified that the marginal model imposing $H_{ij} = w$ on the contingency table has one nonredundant constraint, which can be written in terms of equation (3) as

$$g_1(\mathbf{m}) = 1 - w - \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m})\bigr) = 0.$$

For the observed frequencies in Table 1 (upper left-hand panel), choosing $w = .5$ as an example, the marginal model with constraint $H_{ij} = .5$ yields the estimated expected frequencies shown in Table 2. This results in $G^2 = 1.2207$, $df = 1$, and $p = .2692$.

Item coefficient H_j can be written in a recursive exp-log notation, which is derived in Appendix D for the vector $\mathbf{H}_j = (H_1, H_2, \ldots, H_J)^{\mathrm{T}}$ containing all H_js. Total-scale coefficient H can be written in a recursive exp-log notation, which is derived in Appendix E.

5. Hypotheses in Mokken Scale Analysis

The use of marginal modelling for testing hypotheses in Mokken scale analysis is illustrated by means of the binary data from 484 children who were administered a 25-item balance-task

TABLE 3.
$\hat{\pi}_j^1$-Values for each of the five balance-task scales.

	Scales				
Item	Weight	Distance	Conflict Weight	Conflict Distance	Conflict Balance
1	.318	.118	.444	.684	.921
2	.326	.126	.496	.698	.950
3	.343	.153	.558	.698	.955
4	.382	.165	.791	.702	.963
5	.415	.428	.833	.727	.967

test (Van Maanen et al. 1989). It was hypothesized that the tasks could be divided into five dimensionally different subscales based on the type of task. The subscales are named Distance, Weight, Conflict Weight, Conflict Balance, and Conflict Distance. For a convenient presentation in the tables, in each of the five scales the items are numbered $1, \ldots, 5$. Table 3 shows the proportions-correct (i.e., the $\hat{\pi}_j^1$s) of the 25 items.

5.1. Testing the First Criterion of a Mokken Scale

The first criterion of a Mokken scale is $\rho_{ij} > 0$ for all $i < j$, which is identical to $H_{ij} > 0$ for all $i < j$ (equation (13)). In this section it is explained how marginal modelling can be used to test the global hypothesis that all K item-pair coefficients H_{ij}s are 0. This global test is a novel statistical tool in Mokken scale analysis. To appreciate its usefulness, first we discuss the exploratory analysis and then the confirmatory analysis. In doing this, we only discuss details of exploratory Mokken scale analysis that are relevant here, and skip many other details.

For exploratory Mokken scale analysis, assuming that already $r-1$ items have been selected into a scale (and without worrying how this has been accomplished; for the details, see Mokken, 1971, pp. 190–199; Sijtsma & Molenaar, 2002, Chap. 5), the rth candidate item for selection must have positive correlations (or, equivalently, positive pairwise scalability coefficients) with each of the $r-1$ items already selected (Mokken, 1971, p. 192, third step). This requirement assures us that the first criterion (equation (13)) of a scale is satisfied for the r items selected thus far. If, for the rth item, each of the $r-1$ item-pair coefficients is significantly greater than 0, the first criterion is satisfied, and if this result is also found for other candidate items, each of these items remains in competition to be included in the scale (which of these candidates eventually is the rth item to be selected depends on the second criterion (equation (14)) and other decision rules not discussed here).

The tests of $H_{ij} = 0$ against $H_{ij} > 0$ are conducted by testing the marginal independence of X_i and X_j. This is a simple procedure which can be done with little computational effort. The type I error rate is controlled by a Bonferroni correction, which is very conservative here because the test statistics are dependent, and because tests are accumulated across different steps in the automated item selection algorithm (Mokken, 1971, pp. 196–198).

In confirmatory Mokken scale analysis, the researcher has to test the first criterion for each item pair separately, but here we propose to use a marginal model to test for all K H_{ij} coefficients simultaneously whether they are equal to zero, thus circumventing the Bonferroni correction. Formally, $\mathbf{H}_{ij} = (H_{12}, H_{13}, \ldots, H_{J-1,J})^{\mathrm{T}}$ contains all K coefficients H_{ij} $(i < j)$. If the global null hypothesis that $\mathbf{H}_{ij} = \mathbf{0}$ is rejected, the researcher has to check next whether the sample values of the item-pair scalability coefficients are positive; that is, whether $\widehat{\mathbf{H}}_{ij} > \mathbf{0}$. Only the combination of a rejected global null hypothesis and positive sample H_{ij}s leads to the conclusion that the first criterion (equation (13)) of a Mokken scale is satisfied. If not all sample H_{ij}s are positive, the next step is to identify items that may be rejected from the scale. This is done in the

same way as when the global null hypothesis that $\mathbf{H}_{ij} = \mathbf{0}$ is not rejected. We suggest identifying candidate items for rejection by testing for separate item pairs $H_{ij} = 0$ against the alternative that $H_{ij} > 0$, just as with the exploratory procedure. Item pairs for which the null hypothesis is not rejected are identified, and for each item involved in such a pair it is counted how often it is involved in negative $\widehat{H}_{ij}$s with other items. Items that are frequently involved in negative sample item-pair scalability coefficients are candidates for removal from the test. We now concentrate on the new global test, based on marginal modelling, that $\mathbf{H}_{ij} = \mathbf{0}$.

Let $\mathbf{u}_K$ denote a vector of length K that contains 1s, and let $\mathbf{A}_1$ and $\mathbf{A}_2$ be design matrices (derived in Appendix C). In Appendix C it is shown that

$$\mathbf{H}_{ij} = \mathbf{u}_K - \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})\bigr). \tag{16}$$

Hence, the recursive exp-log notation of the K restrictions (see equation (3)) for marginal model $\mathbf{H}_{ij} = \mathbf{0}$ is

$$\mathbf{g}(\mathbf{m}) = \mathbf{u}_K - \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})\bigr) = \mathbf{0}_K. \tag{17}$$

If the marginal model in equation (17) is rejected and if in the sample $\widehat{H}_{ij} > 0$ for all $i < j$, then the first criterion (equation (13)) for a Mokken scale is met for all J items.

One advantage of this global test is that it does not require a Bonferroni correction. Another advantage is that it allows the first criterion for a Mokken scale to be strengthened, for example, by requiring that all H_{ij}s are greater than a positive value d so as to avoid values of H_{ij} close to 0. Values close to 0 may allow undesirable multidimensionality in a scale, and are not excluded by the second criterion for a Mokken scale, $H_j \geq c > 0$ all j (equation (14)). What is a reasonable choice for d? Because, by equation (11), we have that $\min_{i,j}(H_{ij}) \leq H_j \leq \max_{i,j}(H_{ij})$, it seems reasonable to choose an a priori lower bound d for H_{ij} smaller than c. In this example, we arbitrarily set $d = .1$.

Let $\mathbf{d}_K$ be a vector of length K with all elements equal to d. Then the marginal model equals $\mathbf{H}_{ij} = \mathbf{d}_K$. Using the recursive exp-log notation for $\mathbf{H}_{ij}$ in equation (16), it may be verified that the recursive exp-log notation of the K restrictions (see equation (3)) for this marginal model is

$$\mathbf{g}(\mathbf{m}) = \mathbf{u}_K - \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})\bigr) - \mathbf{d}_K = \mathbf{0}_K. \tag{18}$$

The marginal model with $d = 0$ (equation (17)) and the stronger marginal model with $d = .1$ (equation (18)) were tested on the balance-scale data. For each balance scale, the $\widehat{H}_{ij}$s and their standard errors, and the likelihood ratio statistic G^2 and corresponding p-value, are shown in Table 4. For $d = 0$, using $\alpha = .05$ the null model was rejected for all scales and, in addition, all sample $\widehat{H}_{ij}$s were found to be greater than zero. Thus, the first criterion of the Mokken scale ($\rho_{ij} > 0$; equation (13)) was assumed to be satisfied. For $d = .1$, implying the statistical test that simultaneously all $H_{ij} > .1$, four scales were found to satisfy this more demanding criterion but for the Conflict Balance scale the marginal model in equation (18) was not rejected.

5.2. Testing the Second Criterion of a Mokken Scale

The second criterion of a Mokken scale is that $H_j \geq c > 0$ for all $j = 1, \ldots, J$ (equation (14)). The current practice is that for each item the null hypothesis is tested that $H_j = 0$. When this null hypothesis is rejected, it is checked in the data whether $\widehat{H}_j$ exceeds lower bound c. If for each item the null hypothesis is rejected and $\widehat{H}_j > c$ for all j, the second criterion for a Mokken scale is assumed to be satisfied. Currently, there is no test available for the null hypothesis that $H_j = c$ against the alternative that $H_j > c$ and, sometimes, when the automated item selection procedure is used, an item scalability coefficient is greater than c when the item enters

TABLE 4.

Estimated scalability coefficients $\widehat{H}_{ij}$ with standard errors between parentheses for each of the five balance-task scales (upper panel); fit statistics (G^2, p-value) for the marginal model defining $H_{ij} = 0$ for $i = 1 \ldots, 4$; $j = i + 1, \ldots, 5$ (middle panel); and fit statistics for the marginal model defining $H_{ij} = .1$ for $i = 1 \ldots, 4$; $j = i + 1, \ldots, 5$ (lower panel).

	Scales									
Item pair	Weight		Distance		Conflict Weight		Conflict Distance		Conflict Balance	
1, 2	.658	(.045)	.454	(.085)	.438	(.064)	.730	(.046)	.221	(.110)
1, 3	.557	(.051)	.362	(.091)	.447	(.086)	.718	(.046)	.214	(.111)
1, 4	.632	(.051)	.341	(.102)	.427	(.095)	.696	(.047)	.211	(.111)
1, 5	.589	(.053)	.448	(.103)	.500	(.098)	.756	(.045)	.254	(.123)
2, 3	.561	(.049)	.482	(.064)	.272	(.079)	.672	(.046)	.185	(.100)
2, 4	.529	(.052)	.470	(.072)	.221	(.088)	.632	(.047)	.240	(.108)
2, 5	.580	(.051)	.411	(.075)	.510	(.086)	.787	(.041)	.397	(.124)
3, 4	.576	(.048)	.439	(.071)	.614	(.047)	.647	(.047)	.139	(.083)
3, 5	.575	(.049)	.399	(.073)	.653	(.050)	.700	(.045)	.161	(.094)
4, 5	.499	(.049)	.257	(.067)	.594	(.047)	.669	(.046)	.096	(.080)
Model: All $H_{ij} = .0$										
G^2	589.366		208.155		390.232		684.838		34.706	
p	.000		.000		.000		.000		.000	
Model: All $H_{ij} = .1$										
G^2	366.968		106.484		250.617		439.107		13.191	
p	.000		.000		.000		.000		.213	

the scale, but then drops below c as subsequent items enter the scale (e.g., Sijtsma & Molenaar, 2002, pp. 79–80).

The marginal modelling approach offers a solution. A marginal model may be tested in which, simultaneously, all $H_j = c$. $\mathbf{H}_j = (H_1, \ldots, H_J)^{\mathrm{T}}$ contains all H_js, and let $\mathbf{c}_J$ be a vector of length J with all elements equal to lower bound c. The marginal model is then $\mathbf{H}_j = \mathbf{c}$. If the marginal model is rejected and all sample $\widehat{H}_j$s exceed c, the second criterion is assumed to be satisfied.

Let $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, and $\mathbf{A}_4$ be design matrices (derived in Appendix D). Appendix D shows that

$$\mathbf{H}_j = \mathbf{u}_J - \exp\bigl(\mathbf{A}_4 \log\bigl(\mathbf{A}_3 \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m})\bigr)\bigr)\bigr). \tag{19}$$

Using the recursive exp-log notation for $\mathbf{H}_j$ in equation (19), it may be verified that the recursive exp-log notation of the J restrictions (see equation (3)) for the marginal model is

$$\mathbf{g}(\mathbf{m}) = \mathbf{u}_J - \exp\bigl(\mathbf{A}_4 \log\bigl(\mathbf{A}_3 \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m})\bigr)\bigr)\bigr) - \mathbf{c}_J = \mathbf{0}_J. \tag{20}$$

The marginal model in equation (20) with $c = .3$ (which is the default value in software for Mokken scale analysis) and the marginal model with the more demanding criterion $c = .4$ were tested on the balance-task data. For each balance-task scale, Table 5 shows the estimates of the H_js and their standard errors, and the likelihood ratio statistic G^2 and corresponding p-value. For lower bound $c = .3$, for four scales the marginal model was rejected. In addition, all the $\widehat{H}_j$s exceeded .3. Thus, the four scales meet the second criterion of a Mokken scale. The exception was the Conflict Balance scale, for which the marginal null model was not rejected. Thus, Conflict Balance does not meet the second criterion of a Mokken scale.

For $c = .4$, for the Distance scale the marginal model was not rejected, and for the Conflict Balance scale this marginal model was rejected but all $\widehat{H}_j$s were smaller than .4. Thus, for

TABLE 5.

Estimated scalability coefficients $\widehat{H}_j$ with standard errors between parentheses for each of the five balance-task scales (upper panel); fit statistics (G^2, p-value) for the marginal model defining $H_j = .3$ for $j = 1, \ldots, 5$ (middle panel); and fit statistics for the marginal model defining $H_j = .4$ for $j = 1, \ldots, 5$ (lower panel).

	Scales									
Item	Weight		Distance		Conflict Weight		Conflict Distance		Conflict Balance	
1	.610	(.037)	.403	(.066)	.450	(.052)	.725	(.037)	.225	(.092)
2	.584	(.033)	.456	(.045)	.359	(.050)	.704	(.032)	.261	(.077)
3	.567	(.031)	.427	(.044)	.538	(.035)	.683	(.030)	.172	(.072)
4	.557	(.033)	.381	(.047)	.518	(.038)	.660	(.032)	.164	(.073)
5	.559	(.035)	.372	(.054)	.588	(.042)	.727	(.030)	.213	(.075)
Model: All $H_j = .3$										
G^2	100.209		14.546		68.578		175.373		6.332	
p	.000		.013		.000		.000		.275	
Model: All $H_j = .4$										
G^2	40.982		5.022		36.590		102.063		11.284	
p	.000		.413		.000		.000		.046	

these two scales the more demanding second criterion of a Mokken scale was not satisfied. For the other three scales, the null model was rejected and all $\widehat{H}_j$s exceeded .4; hence, the more demanding second criterion of a Mokken scale was satisfied.

5.3. *Testing the Strength of the Scale*

Testing the strength of the scale can be considered equivalent with testing for the total-scale coefficient that $H \leq c$ against the alternative that $H > c$. If the null model is rejected for $c = .3$ and if in the sample $\widehat{H} > .3$, then the scale can be considered to be at least a weak scale; if the null model is rejected for $c = .4$ and if $\widehat{H} > .4$, then the scale can be considered to be at least a moderate scale; and if the null model is rejected for $c = .5$ and if $\widehat{H} > .5$, then the scale can be considered to be a strong scale. The statistical test can be performed using the asymptotic standard errors derived by Mokken (1971, pp. 164–169). From the asymptotic standard errors a $(1 - \alpha)\%$ confidence interval is constructed, and if c exceeds the upper bound of the confidence interval, the null hypothesis is rejected. This test is not available in the current software.

Alternatively, the test may be conducted using a marginal model. Let $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, and $\mathbf{A}_4$ be design matrices. These matrices are derived in Appendix E. Appendix E shows that H can be written as

$$H = 1 - \exp\bigl(\mathbf{A}_4 \log\bigl(\mathbf{A}_3 \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})\bigr)\bigr)\bigr). \tag{21}$$

Using equation (21) it can be verified that the recursive exp-log notation of the restriction (see equation (3)) in the null model is

$$g_1(\mathbf{m}) = 1 - \exp\bigl(\mathbf{A}_4 \log\bigl(\mathbf{A}_3 \exp\bigl(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})\bigr)\bigr)\bigr) - c = 0. \tag{22}$$

It may be noted that, in principle, in equation (22) lower bound c may be replaced by any constant $w > 0$.

The marginal models with $c = .3$, $c = .4$, and $c = .5$ (equation (22)) were tested on the balance-scale data. For each scale, Table 6 shows the estimate of coefficient H and its standard error, and the likelihood ratio statistic G^2 and corresponding p-value. Using the rules of thumb for the interpretation of values of H, Weight and Conflict Distance were strong scales,

TABLE 6.

For each of the five scales of the balance-task test: The estimated scalability coefficient $\widehat{H}$ with standard error between parentheses (first row); and the fit statistics (G^2, p-value) for the marginal models defining $H = .3$, $H = .4$, and $H = .5$.

	Scales				
	Weight	Distance	Conflict Weight	Conflict Distance	Conflict Balance
Coefficient H	.576 (.028)	.410 (.040)	.502 (.032)	.700 (.027)	.205 (.068)
Model: $H = .3$					
G^2	97.141	8.170	39.559	169.855	1.573
p	.000	.004	.000	.000	.210
Model: $H = .4$					
G^2	38.647	0.057	9.971	96.676	5.828
p	.000	.812	.002	.000	.016
Model: $H = .5$					
G^2	7.244	4.921	0.003	45.041	12.647
p	.007	.027	.956	.000	.000

Conflict Weight a moderate scale, Distance a weak scale, and Conflict Balance was found to be unscalable.

5.4. Testing Equality of Item Coefficients

Coefficient H_j expresses the contribution of item j to the ordering of respondents by means of total score X_+. Thus, it can be argued that coefficient H_j is a nonparametric IRT analogue to the discrimination power of an item (Van Abswoude, Van der Ark, & Sijtsma, 2004). The marginal modelling framework can be used to test whether the H_js of different items are equal. This may be interesting when one wants to know whether the items are different with respect to their contribution to the accuracy of the person ordering. Large differences may also provide the researcher with indications that different latent variables may drive the responses to different items (Sijtsma & Meijer, 2007). Currently, such a test is not available.

A statistical test for the null hypothesis "$H_1 = \cdots = H_J$" requires a slight modification of the marginal model in equation (20). Let $\mathbf{A}_5$ be a $(J-1) \times J$ matrix with element (j, j) equal to 1 for $j = 1, \ldots, J-1$; and element $(j, j+1)$ equal to -1 for $j = 1, \ldots, J-1$; the remaining elements are equal to 0. Using equation (19), it may be verified that

$$\mathbf{A}_5\mathbf{H}_j = \begin{pmatrix} H_1 - H_2 \\ H_2 - H_3 \\ \vdots \\ H_{J-1} - H_J \end{pmatrix},$$

which should be equal to $\mathbf{0}_{J-1}$ if all H_js are equal. Then using the design matrices $\mathbf{A}_1, \ldots, \mathbf{A}_4$ from equation (20) (see Appendix D), the marginal model for equal item coefficients is

$$\mathbf{g}(\mathbf{m}) = \mathbf{A}_5\big(\exp\big(\mathbf{A}_4 \log\big(\mathbf{A}_3 \exp\big(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m})\big)\big)\big)\big) = \mathbf{0}_{J-1}. \tag{23}$$

Using the marginal model in equation (23), the null hypothesis that $H_1 = H_2 = H_3 = H_4 = H_5$ was tested for each of the five balance-task scales. It may be noted that if all H_js are equal, then equation (11) implies that $H = H_j$. For each scale, Table 7 shows the estimated total-scale H and its standard error, under the marginal model of equal H_js, and the likelihood ratio statistic G^2 and corresponding p-value. For Conflict Weight the null model of equal H_js was rejected. For the other four scales the null model was not rejected, thus providing support for equal item contributions to the person ordering.

TABLE 7.

For the marginal model defining $H_1 = \cdots = H_5$: For each of five balance-task scales, estimated coefficient H with standard error between parentheses (upper panel); and fit statistic G^2, p-value (lower panel).

	Scales				
	Weight	Distance	Conflict Weight	Conflict Distance	Conflict Balance
H	.571 (.027)	.416 (.039)	.509 (.031)	.690 (.027)	.191 (.062)
G^2	3.015	4.850	25.039	9.114	4.119
p	.555	.303	.000	.058	.390

5.5. *Multiple-Group Hypotheses*

Mokken (1971, pp. 164–169) provided the asymptotic sampling theory for testing the null hypothesis that the H values for the same test in different groups are equal. Under this null hypothesis, the same test orders respondents from different groups with equal accuracy. For example, the balance-task test was administered to both boys and girls, and it may be interesting to test if the test orders boys and girls equally well. MSP (Molenaar & Sijtsma, 2000) allows the possibility to compare the H_j and H values of different groups, but not to test hypotheses about (in-)equality of H in different groups.

Assume that there are G groups, and let superscript g index these groups. Then the null hypothesis of interest is "$H^1 = \cdots = H^G$". The recursive exp-log notation requires the following definitions. Let $\mathbf{A}_1^1, \ldots, \mathbf{A}_1^G$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$, and $\mathbf{A}_5$ be design matrices (derived in Appendix F). Let $\mathbf{m}^*$ be a vector of length LG in which the vectors of expected frequencies from groups $1, \ldots, G$ are stacked, such that $\mathbf{m}^* = (\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^G)$. The symbol $\bigoplus$ indicates the direct product (see Appendix F). In Appendix F it is shown that

$$\begin{pmatrix} H^1 \\ H^2 \\ \vdots \\ H^G \end{pmatrix} = \mathbf{u}_G - \exp\left(\bigoplus_{g=1}^{G} \mathbf{A}_4 \log\left(\bigoplus_{g=1}^{G} \mathbf{A}_3 \exp\left(\bigoplus_{g=1}^{G} \mathbf{A}_2 \log\left(\bigoplus_{g=1}^{G} \mathbf{A}_1^g \mathbf{m}^*\right)\right)\right)\right). \tag{24}$$

In Appendix F it is also shown that the recursive exp-log notation for the marginal model with "$H^1 = H^2 = \cdots = H^G$" is

$$\begin{aligned} \mathbf{g}(\mathbf{m}) &= \begin{pmatrix} H^1 - H^2 \\ H^2 - H^3 \\ \vdots \\ H^{G-1} - H^G \end{pmatrix} \\ &= \mathbf{A}_5 \exp\left(\bigoplus_{g=1}^{G} \mathbf{A}_4 \log\left(\bigoplus_{g=1}^{G} \mathbf{A}_3 \exp\left(\bigoplus_{g=1}^{G} \mathbf{A}_2 \log\left(\bigoplus_{g=1}^{G} \mathbf{A}_1^g \mathbf{m}^*\right)\right)\right)\right) = \mathbf{0}_G. \end{aligned} \tag{25}$$

The marginal model in equation (25) was used to test equal H for boys (indexed $g = 1$) and girls ($g = 2$); that is, $H^1 = H^2$. For each balance-task scale, Table 8 shows coefficient H^g and its standard error, and the likelihood ratio statistic G^2 and corresponding p-value. For each scale, the sample $\widehat{H}$ value was higher for girls than for boys but only for Conflict Distance was the difference significant. Notice that for $G = 2$, if estimated standard errors are available, this result can be approximated using a t-test.

A generalization of the multigroup hypothesis to coefficients H_j and H_{ij} is straightforward if the item ordering is the same in all subgroups. If the item ordering is different for some subgroups, the design matrices must be adapted.

TABLE 8.

For the marginal model defining $H^1 = H^2$: For each of the five balance-task scales, scalability coefficients H for boys and girls with standard error between parentheses (upper panel); and fit statistics G^2, p-value (lower panel).

	Scales									
	Weight		Distance		Conflict Weight		Conflict Distance		Conflict Balance	
Boys	.544	(.037)	.348	(.053)	.439	(.044)	.626	(.044)	.114	(.045)
Girls	.591	(.042)	.467	(.059)	.560	(.046)	.749	(.034)	.356	(.134)
G^2	0.713		2.392		3.556		5.015		3.342	
p	.398		.122		.059		.025		.068	

6. Discussion

Marginal modelling offers a framework for testing many interesting hypotheses relevant to Mokken scale analysis that could not be tested before. In particular, new and exciting possibilities of the marginal modelling approach are:

(1) The availability of global tests that evaluate all interitem scalability coefficients H_{ij} simultaneously and all item-scalability coefficients H_j simultaneously. This offers new opportunities for assessing item and test quality.
(2) The possibility to test whether scalability coefficients are equal to a particular value. This is important for ascertaining item and test quality at a level deemed necessary by the researcher. This result also offers the possibility to test hypotheses about expected values of scalability coefficients (such as those derived from previous research).
(3) The comparison of scalability coefficients between different groups. This provides the opportunity to assess whether the measurement quality of a test is the same in different groups.

This paper has presented several useful examples but the array of possibilities has not yet been fully explored. Exploring these possibilities and implementing the most useful ones in user-friendly software is the first topic for future research.

One possible limitation of the marginal modelling approach is that for the global tests assessing all scalability coefficients simultaneously and to a lesser degree for tests of coefficient H alone, the size of the matrices can grow rapidly as the number of items increases. The experience accumulated thus far did not reveal computational problems for tests up to $J = 15$. Matrix $\mathbf{R}$ (equation (2)), which is required to solve the marginal modelling problem, has $L = 2^{15} = 32760$ rows. For larger J, the maximum likelihood estimation of the models becomes impractical. One solution may be to use an estimation procedure that only evaluates the observed item-score patterns so that the size of vector $\mathbf{m}$ does not exceed n. An example is the minimum information discrimination approach (e.g., Kullback, 1971; Read & Cressie, 1988, pp. 34–40). Applying alternative estimation procedures to marginal modelling of the scalability coefficients for Mokken scale analysis is the second topic for future research.

The methods presented here are only applicable to dichotomous items. Thus, a useful generalization is to Mokken scale analysis for polytomous items. Whereas, for dichotomous items, some of the interesting hypotheses tested in Mokken scale analysis could also be tested without the use of marginal models, this is often not possible for polytomous items. Examples are the computation of standard errors and testing the strength of the scale. The generalization of results for dichotomous items to polytomous items has proven to be problematic in many ways (e.g., Hemker et al., 1997; Sijtsma & Meijer, 2007), and this may also be true in the marginal modelling framework. The derivation of the design matrices for marginal

models is more complicated and the magnitude the computational problems is more troublesome. The generalization of the methods to polytomous items is the third topic for future research.

The syntax files for the marginal models used here are available upon request from the first author. Currently, researchers wishing to apply the marginal models presented in this paper need to have Mathematica installed on their computer.

Appendix A. Estimation of Marginal Models

Appendix A discusses the details of the optimization algorithm for estimating and testing the marginal models discussed in this paper (see also Bergsma & Croon, 2005). Suppose that a sample of n respondents provided responses to J items that are dichotomously scored. The number of different item-score patterns (see equation (2)) is $L = 2^J$. (In the more general case where item j has v_j ordered item scores, the number of different item-score patterns is given by $L = \prod_j v_j$.) Vectors $\mathbf{n}$ and $\mathbf{m}$ are both of length L, and contain the observed frequencies and expected frequencies of the item-score patterns, respectively. The marginal models discussed can be specified by a set of C equations that impose constraints on the theoretical expected frequencies in $\mathbf{m}$, which are collected in equation (3),

$$\mathbf{g}(\mathbf{m}) = \begin{pmatrix} g_1(\mathbf{m}) \\ \vdots \\ g_C(\mathbf{m}) \end{pmatrix} = \mathbf{0}.$$

Each constraint is defined recursively in terms of appropriate scalar functions and matrices as in equation (4).

Bergsma (1997a, pp. 89–95) developed a Fisher scoring algorithm to find the maximum likelihood (ML) estimates of the constrained theoretical expected frequencies in $\mathbf{m}$ (or, equivalently, the constrained cell probabilities). Assuming multinomial sampling and a vector $\boldsymbol{\mu}$ that contains C unknown Lagrangian multipliers, the augmented likelihood or Lagrangian is

$$L(\mathbf{m}, \boldsymbol{\mu}) = \mathbf{n}^{\mathrm{T}} \log(\mathbf{m}) - \boldsymbol{\mu}^{\mathrm{T}} \mathbf{g}(\mathbf{m}).$$

The ML estimates of the expected frequencies in vector $\mathbf{m}$ are obtained by means of an iterative procedure that determines a saddlepoint of this Lagrangian.

Let $\mathbf{G} = \mathbf{G}(\mathbf{m})$ be the Jacobian of $\mathbf{g}(\mathbf{m})$ with respect to $\log \mathbf{m}$. Hence, $\mathbf{G}$ is a $C \times C$ matrix with elements $g_{rs} = \partial g_r(\mathbf{m}) / \partial \log m_s$. Derivation of $\mathbf{G}$ can be done using the same recursive exp-log notation that was used to specify $\mathbf{g}(\mathbf{m})$ in equation (4). First, let $\phi(x)$ be a function that either indicates an exponential ($\phi(x) = \exp(x)$, $\phi'(x) = \exp(x)$), a logarithm ($\phi(x) = \log(x)$, $\phi'(x) = 1/x$), or a translation ($\phi(x) = x + c$, where c is some constant value, $\phi'(x) = 1$). Second, let $\mathbf{f}_0(\mathbf{m}), \mathbf{f}_1(\mathbf{m}), \mathbf{f}_2(\mathbf{m}), \ldots, \mathbf{f}_q(\mathbf{m})$ be a series of $q + 1$ functions, in which

$$\begin{aligned} &\mathbf{f}_0(\mathbf{m}) = \mathbf{m}, \\ &\mathbf{f}_i(\mathbf{m}) = \phi[\mathbf{A}_i \mathbf{f}_{i-1}(\mathbf{m})] \quad \text{for } i = 1, \ldots, q. \end{aligned} \tag{26}$$

The last function in equation (26) is

$$\mathbf{f}_q(\mathbf{m}) = \mathbf{g}(\mathbf{m})$$

as specified in equation (4). Third, the following recursive relationship can be derived for the partial derivatives of the functions $\mathbf{f}_i(\mathbf{m})$. Let $\mathbf{D}(\mathbf{v})$ be a diagonal matrix with vector $\mathbf{v}$ on its main

diagonal, then

$$\frac{\partial \mathbf{f}_0(\mathbf{m})}{\partial \log \mathbf{m}} = \mathbf{D}(\mathbf{m}),$$

and

$$\frac{\partial \mathbf{f}_i(\mathbf{m})}{\partial \log \mathbf{m}} = \mathbf{D}\left[\boldsymbol{\phi}'(\mathbf{A}_i \mathbf{f}_{i-1})\right]\mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{m})}{\partial \log \mathbf{m}} \quad \text{for } i = 1, \ldots, q. \tag{27}$$

Note that if ϕ indicates an exponential, then equation (27) equals

$$\frac{\partial \mathbf{f}_i(\mathbf{m})}{\partial \log \mathbf{m}} = \mathbf{D}[\exp(\mathbf{A}_i \mathbf{f}_{i-1})]\mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{m})}{\partial \log \mathbf{m}};$$

if ϕ indicates a logarithm, then equation (27) equals

$$\frac{\partial \mathbf{f}_i(\mathbf{m})}{\partial \log \mathbf{m}} = \mathbf{D}^{-1}(\mathbf{A}_i \mathbf{f}_{i-1})\mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{m})}{\partial \log \mathbf{m}};$$

and if ϕ indicates a translation, then equation (27) equals

$$\frac{\partial \mathbf{f}_i(\mathbf{m})}{\partial \log \mathbf{m}} = \mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{m})}{\partial \log \mathbf{m}}.$$

Fourth, the Jacobian can be obtained as

$$\mathbf{G} = \frac{\partial \mathbf{f}_q(\mathbf{m})}{\partial \log \mathbf{m}}.$$

Differentiating $L(\mathbf{m}, \boldsymbol{\mu})$ with respect to $\log \mathbf{m}$ yields

$$\mathbf{l}(\mathbf{m}, \boldsymbol{\mu}) = \mathbf{n} - \mathbf{m} - \mathbf{G}\boldsymbol{\mu}.$$

Under suitable regularity conditions, the ML estimator $\widehat{\mathbf{m}}$ is a vector $\mathbf{m}$ for which there is a Lagrange multiplier vector $\boldsymbol{\mu}$ such that the simultaneous equations

$$\mathbf{l}(\mathbf{m}, \boldsymbol{\mu}) = 0$$

and

$$\mathbf{g}(\mathbf{m}) = 0$$

are satisfied.

Then the expected value of the derivative matrix of the vector $(\mathbf{l}(\mathbf{m}, \boldsymbol{\mu}), \mathbf{g}(\mathbf{m}))$ with respect to $(\mathbf{m}, \boldsymbol{\mu})$ is

$$\mathbf{V}(\mathbf{m}) = \begin{pmatrix} -\mathbf{D}(\mathbf{m}) & \mathbf{G} \\ \mathbf{G}^{\mathrm{T}} & 0 \end{pmatrix}.$$

Let $\mathbf{n}^+$ be equal to the vector $\mathbf{n}$ with zeros replaced by a small positive constant (say, 10^{-10}), and define the Fisher scoring starting values

$$\begin{pmatrix} \log \mathbf{m}^{(0)} \\ \boldsymbol{\mu}^{(0)} \end{pmatrix} = \begin{pmatrix} \log \mathbf{n}^+ \\ 0 \end{pmatrix}$$

and, for $k = 0, 1, \ldots,$

$$\begin{pmatrix} \log \mathbf{m}^{(k+1)} \\ \boldsymbol{\mu}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \log \mathbf{m}^{(k)} \\ \boldsymbol{\mu}^{(k)} \end{pmatrix} - \mathbf{V}\big(\mathbf{m}^{(k)}\big)^{-1} \cdot \begin{pmatrix} \mathbf{l}(\mathbf{m}^{(k)}, \boldsymbol{\mu}^{(k)}) \\ \mathbf{g}(\mathbf{m}^{(k)}) \end{pmatrix}.$$

Then, as $k \to \infty$, $\mathbf{m}^{(k)}$ should go to $\widehat{\mathbf{m}}$. Straightforward matrix algebra yields the simplified form

$$\begin{aligned} \log \mathbf{m}^{(k+1)} &= \log \mathbf{m}^{(k)} + \mathbf{D}\big(\mathbf{m}^{(k)}\big)^{-1} \mathbf{l}\big(\mathbf{m}^{(k)}, \boldsymbol{\mu}^{(k+1)}\big), \\ \boldsymbol{\mu}^{(k+1)} &= -\big(\mathbf{G}^{\mathrm{T}} \mathbf{D}\big(\mathbf{m}^{(k)}\big)\mathbf{G}\big)^{-1} \big(\mathbf{G}^{\mathrm{T}} \mathbf{D}\big(\mathbf{m}^{(k)}\big)^{-1} \big(\mathbf{n} - \mathbf{m}^{(k)}\big) + \mathbf{g}\big(\mathbf{m}^{(k)}\big)\big). \end{aligned}$$

This algorithm does not always converge, and it can be helpful to introduce a step size $\text{step}^{(k)} \in \langle 0, 1]$ as follows:

$$\log \mathbf{m}^{(k+1)} = \log \mathbf{m}^{(k)} + \text{ step}^{(k)} \mathbf{D}\big(\mathbf{m}^{(k)}\big)^{-1} \mathbf{l}\big(\mathbf{m}^{(k)}, \boldsymbol{\mu}^{(k+1)}\big).$$

Note that the update of $\boldsymbol{\mu}$ is left unchanged.

The step size should be chosen so that the new estimate $\mathbf{m}^{(k+1)}$ is "better" than the old estimate $\mathbf{m}^{(k)}$. A criterion for deciding this is obtained by defining the following quadratic form measuring the "distance" from convergence:

$$\delta\big(\mathbf{m}^{(k)}\big) = \mathbf{l}\big(\mathbf{m}^{(k)}, \boldsymbol{\mu}^{(k+1)}\big) \mathbf{D}\big(\mathbf{m}^{(k)}\big)^{-1} \mathbf{l}\big(\mathbf{m}^{(k)}, \boldsymbol{\mu}^{(k+1)}\big).$$

Convergence is reached at $\mathbf{m}$ if and only if $\delta(\mathbf{m}) = 0$ and therefore, if possible, the step size should be chosen so that $\delta(\mathbf{m}^{(k+1)}) < \delta(\mathbf{m}^{(k)})$ for all k. This is possible if the tentative solution is sufficiently close to the ML estimate. Otherwise, a recommendation which seems to work very well in practice is to "jump" to another region by taking a step size equal to one.

After convergence of the estimation procedure, the null hypothesis that the model specified by the C constraints $\mathbf{g}(\mathbf{m}) = 0$ provides an acceptable fit to the data can be tested against the saturated model by means of a log likelihood ratio test. The test statistic is

$$G^2 = 2\mathbf{n}^{\mathrm{T}} \log(\mathbf{n}/\widehat{\mathbf{m}}),$$

which asymptotically follows a chi-square distribution with $df = C$.

Appendix B. Definition of Paired Row Product of Two Matrices

Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of order $n \times m$, let $\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}}, \ldots, \mathbf{a}_n^{\mathrm{T}}$ be the rows of $\mathbf{A}$, and let $\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2^{\mathrm{T}}, \ldots, \mathbf{b}_n^{\mathrm{T}}$ be the rows of $\mathbf{B}$. The paired row product of $\mathbf{A}$ and $\mathbf{B}$ is the elementwise or Hadamard product of row i of $\mathbf{A}$ and row j of $\mathbf{B}$ for $i = 1, \ldots, n-1,\ j = i+1, \ldots, n$, and is denoted $\mathbf{A} \circledast \mathbf{B}$. If $\mathbf{A} \circledast \mathbf{B} = \mathbf{C}$, then $\mathbf{C}$ is a $\frac{1}{2}n(n-1) \times m$ matrix with rows $\mathbf{c}_1^{\mathrm{T}}, \mathbf{c}_2^{\mathrm{T}}, \ldots, \mathbf{c}_{(1/2)n(n-1)}^{\mathrm{T}}$. Let $k = (i-1)n - \frac{1}{2}i(i-1) + (j-i)$, then

$$\mathbf{c}_k = \mathbf{a}_i \bullet \mathbf{b}_j \quad \text{for } i = 1, \ldots, n-1,\ j = i+1, \ldots, n,$$

where • denotes the Hadamard product. For example,

$$\begin{pmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \\ \mathbf{a}_3^{\mathrm{T}} \\ \mathbf{a}_4^{\mathrm{T}} \end{pmatrix} \circledast \begin{pmatrix} \mathbf{b}_1^{\mathrm{T}} \\ \mathbf{b}_2^{\mathrm{T}} \\ \mathbf{b}_3^{\mathrm{T}} \\ \mathbf{b}_4^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^{\mathrm{T}} \bullet \mathbf{b}_2^{\mathrm{T}} \\ \mathbf{a}_1^{\mathrm{T}} \bullet \mathbf{b}_3^{\mathrm{T}} \\ \mathbf{a}_1^{\mathrm{T}} \bullet \mathbf{b}_4^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \bullet \mathbf{b}_3^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \bullet \mathbf{b}_4^{\mathrm{T}} \\ \mathbf{a}_3^{\mathrm{T}} \bullet \mathbf{b}_4^{\mathrm{T}} \end{pmatrix}.$$

Appendix C. Recursive Exp-Log Notation for All Item-Pair Scalability Coefficients Simultaneously

In Appendix C, the recursive exp-log notation for equation (16) is derived. For the purpose of illustration, the design matrices are elaborated for a three-item test (hence, there are $K = 3$ item pairs and $L = 8$ item-score patterns). In these design matrices, dashed lines are displayed to facilitate readability.

Let $\mathbf{u}_L$ be a vector of length L that consists of ones; let $\mathbf{U} = \mathbf{u}_J \mathbf{u}_L^{\mathrm{T}}$, and let $\mathbf{R}$ be the $L \times J$ matrix that contains all possible item-score patterns defined in equation (2). The symbol $\circledast$ denotes the paired row product (Appendix B). Then, the $(1 + 2J + K) \times L$ design matrix $\mathbf{A}_1$ is a concatenation of four submatrices, that is,

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{u}_L^{\mathrm{T}} \\ \mathbf{U} - \mathbf{R}^{\mathrm{T}} \\ \mathbf{R}^{\mathrm{T}} \\ (\mathbf{U} - \mathbf{R}^{\mathrm{T}}) \circledast \mathbf{R}^{\mathrm{T}} \end{pmatrix}. \tag{28}$$

It may be verified that for three items (denoted by a, b, and c) of decreasing popularity, we have that

$$\log(\mathbf{A}_1 \mathbf{m}) = \log \left[\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hdashline 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hdashline 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} m_{abc}^{000} \\ m_{abc}^{001} \\ m_{abc}^{010} \\ m_{abc}^{011} \\ m_{abc}^{100} \\ m_{abc}^{101} \\ m_{abc}^{110} \\ m_{abc}^{111} \end{pmatrix} \right] = \log \begin{pmatrix} n \\ \hdashline m_a^0 \\ m_b^0 \\ m_c^0 \\ \hdashline m_a^1 \\ m_b^1 \\ m_c^1 \\ \hdashline m_{ab}^{01} \\ m_{ac}^{01} \\ m_{bc}^{01} \end{pmatrix}. \tag{29}$$

The $K \times (1 + 2J + K)$ design matrix $\mathbf{A}_2$ is a concatenation of three submatrices, that is, $\mathbf{A}_2 = (\mathbf{u}_K \ -\mathbf{Q}_1^T \ \mathbf{I}_K)$, in which $\mathbf{I}_K$ is the identity matrix of order K, and $\mathbf{Q}_1$ is a $K \times (2J)$ matrix containing zeros and ones. The rows of $\mathbf{Q}_1$ correspond to K item pairs, that is, item pair (i, j) (with $i = 1, \ldots, J-1, \ \ j = i+1, \ldots, J$) corresponds to the kth row of $\mathbf{Q}_1$ ($k = (i-1)J - \frac{1}{2}i(i-1) + (j-i)$, see also Appendix B). The columns of $\mathbf{Q}_1$ can be divided into two sets: the first J columns and the last J columns. Each row of $\mathbf{Q}_1$ has $2(J-1)$ zeros and two ones; the elements with value 1 are in the jth column of the first set of columns and in the ith

column of the last set of columns (i.e., column $J+i$). Using equation (29) it may be verified that for three items (a, b, and c) of decreasing popularity, we have that $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m}))$ equals

$$\exp\left[\left(\begin{array}{c:ccccc:ccc} 1 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{array}\right) \log\left[\begin{pmatrix} n \\ \hdashline m_a^0 \\ m_b^0 \\ m_c^0 \\ \hdashline m_a^1 \\ m_b^1 \\ m_c^1 \\ \hdashline m_{ab}^{01} \\ m_{ac}^{01} \\ m_{bc}^{01} \end{pmatrix}\right]\right]$$

$$= \begin{pmatrix} [n \times m_{ab}^{01}]/[m_a^0 \times m_b^1] \\ [n \times m_{ac}^{01}]/[m_a^0 \times m_c^1] \\ [n \times m_{bc}^{01}]/[m_b^0 \times m_c^1] \end{pmatrix}. \tag{30}$$

For three items, it may be verified that substituting the term $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m}))$ in equation (16) with the right-hand side of equation (30) produces coefficients H_{ij} as defined in equation (6).

Appendix D. Recursive Exp-Log Notation for All Item Scalability Coefficients Simultaneously

In Appendix D the recursive exp-log notation for equation (19) is derived. For the purpose of illustration, the design matrices are elaborated for a three-item test (hence, there are $K = 3$ item pairs and $L = 8$ item-score patterns). In the design matrices, dashed lines are displayed to facilitate readability.

Design matrix $\mathbf{A}_1$ was derived in Appendix C (equation (28)). The $(1+2K) \times (1+2J+K)$ design matrix $\mathbf{A}_2$ is the *direct sum* of the scalar 1, submatrix $\mathbf{Q}_1$ (Appendix C), and $\mathbf{I}_K$, that is,

$$\mathbf{A}_2 = 1 \oplus \mathbf{Q}_1 \oplus \mathbf{I}_K = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_K \end{pmatrix}. \tag{31}$$

Using equation (29) it may be verified that for three items (a, b, and c) in decreasing order of popularity, we have that $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m}))$ equals

$$\exp\left[\left(\begin{array}{c:ccccc:ccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \log\left[\begin{pmatrix} n \\ \hdashline m_a^0 \\ m_b^0 \\ m_c^0 \\ \hdashline m_a^1 \\ m_b^1 \\ m_c^1 \\ \hdashline m_{ab}^{01} \\ m_{ac}^{01} \\ m_{bc}^{01} \end{pmatrix}\right]\right] = \begin{pmatrix} n \\ \hdashline m_a^0 m_b^1 \\ m_a^0 m_c^1 \\ m_b^0 m_c^1 \\ \hdashline m_{ab}^{01} \\ m_{ac}^{01} \\ m_{bc}^{01} \end{pmatrix}. \tag{32}$$

The $(1+2J)\times(1+2K)$ design matrix $\mathbf{A}_3$ is the direct sum of the scalar 1, and twice the submatrix $\mathbf{Q}_2$, that is,

$$\mathbf{A}_3 = 1 \oplus \mathbf{Q}_2 \oplus \mathbf{Q}_2 = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}_2 \end{pmatrix},$$

where $\mathbf{Q}_2$ is a $J \times K$ matrix, where the rows correspond to the J items and the columns correspond to the K item pairs. Element i, j in $\mathbf{Q}_2$ equals 1 if the item corresponding to row i is in the item pair corresponding to column j and 0 otherwise. Using equation (32), it may be verified that for three items (a, b, and c) in decreasing order of popularity, we have that $\log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A_1 m})))$ equals

$$\log\left[\left(\begin{array}{c|ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right)\left(\begin{array}{c} n \\ \hline m_a^0 m_b^1 \\ m_a^0 m_c^1 \\ m_b^0 m_c^1 \\ \hline m_{ab}^{01} \\ m_{ac}^{01} \\ m_{bc}^{01} \end{array}\right)\right] = \log\left(\begin{array}{c} n \\ \hline m_a^0 m_b^1 + m_a^0 m_c^1 \\ m_a^0 m_b^1 + m_b^0 m_c^1 \\ m_a^0 m_c^1 + m_b^0 m_c^1 \\ \hline m_{ab}^{01} + m_{ac}^{01} \\ m_{ab}^{01} + m_{bc}^{01} \\ m_{ac}^{01} + m_{bc}^{01} \end{array}\right). \tag{33}$$

For the general case of J items, the middle part of the vector on the right-hand side of equation (33) is a subvector of length J with element j equal to $\sum_{i=1}^{j-1} m_i^0 m_j^1 + \sum_{i=j+1}^{J} m_j^0 m_i^1$. Similarly, the lower part of the vector on the right-hand side of equation (33) is a subvector of length J with element j equal to $\sum_{i=1}^{j-1} m_{ij}^{01} + \sum_{i=j+1}^{J} m_{ji}^{01}$.

The $J \times (1+2J)$ design matrix $\mathbf{A}_4$ is a concatenation of the unit vector, the negative of the identity matrix, and the identity matrix,

$$\mathbf{A}_4 = (\mathbf{1}_J \quad -\mathbf{I}_J \quad \mathbf{I}_J).$$

Using the right-hand side of equation (33), it may be verified that for three items (a, b, and c) ordered according to decreasing popularity, we have that $\exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A_1 m}))))$ equals

$$\begin{aligned} &\exp\left[\left(\begin{array}{c|ccc|ccc} 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 1 \end{array}\right)\log\left(\begin{array}{c} n \\ \hline m_a^0 m_b^1 + m_a^0 m_c^1 \\ m_a^0 m_b^1 + m_b^0 m_c^1 \\ m_a^0 m_c^1 + m_b^0 m_c^1 \\ \hline m_{ab}^{01} + m_{ac}^{01} \\ m_{ab}^{01} + m_{bc}^{01} \\ m_{ac}^{01} + m_{bc}^{01} \end{array}\right)\right] \\ &= \begin{pmatrix} [n(m_{ab}^{01} + m_{ac}^{01})]/[m_a^0 m_b^1 + m_a^0 m_c^1] \\ [n(m_{ab}^{01} + m_{bc}^{01})]/[m_a^0 m_b^1 + m_b^0 m_c^1] \\ [n(m_{ac}^{01} + m_{bc}^{01})]/[m_a^0 m_c^1 + m_b^0 m_c^1] \end{pmatrix}. \end{aligned} \tag{34}$$

For the general case of J items, the vector on the right-hand side of equation (34) is a vector of length J with element j equal to

$$\frac{n\left(\sum_{i=1}^{j-1} m_{ij}^{01} + \sum_{i=j+1}^{J} m_{ji}^{01}\right)}{\sum_{i=1}^{j-1} m_i^0 m_j^1 + \sum_{i=j+1}^{J} m_j^0 m_i^1}.$$

For three items, it may be verified that substituting the term $\exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A_1 m}))))$ in equation (19) with the right-hand side of equation (34) produces coefficients H_j as defined in equation (8).

Appendix E. Recursive Exp-Log Notation for the Scalability Coefficient for a Set of Items

In Appendix E the recursive exp-log notation for equation (21) is derived. For the purpose of illustration, the design matrices are elaborated for the case of three items (hence, there are $K = 3$ item pairs and $L = 8$ item-score patterns). In these design matrices, dashed lines are displayed to facilitate readability.

The recursive exp-log notation for scale coefficient H requires four design matrices, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, and $\mathbf{A}_4$, each consisting of submatrices. Design matrix $\mathbf{A}_1$ was derived in Appendix C (equation (28)), and design matrix $\mathbf{A}_2$ was derived in Appendix D (equation (31)). The $3 \times (1 + 2K)$ design matrix $\mathbf{A}_3$ is the direct sum of the scalar 1, and twice the row vector $\mathbf{u}_K^{\mathrm{T}}$, that is,

$$\mathbf{A}_3 = 1 \oplus \mathbf{u}_K^{\mathrm{T}} \oplus \mathbf{u}_K^{\mathrm{T}} = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_K^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{u}_K^{\mathrm{T}} \end{pmatrix}. \tag{35}$$

Using equation (32), it may be verified that for three items (a, b, and c) ordered according to decreasing popularity, we have that $\log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A_1 m})))$ equals

$$\log\left[\left(\begin{array}{c:ccc:ccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right)\begin{pmatrix} n \\ \hdashline m_a^0 m_b^1 \\ m_a^0 m_c^1 \\ m_b^0 m_c^1 \\ \hdashline m_{ab}^{01} \\ m_{ac}^{01} \\ m_{bc}^{01} \end{pmatrix}\right] = \log\begin{pmatrix} n \\ m_a^0 m_b^1 + m_a^0 m_c^1 + m_b^0 m_c^1 \\ m_{ab}^{01} + m_{ac}^{01} + m_{bc}^{01} \end{pmatrix}. \tag{36}$$

For the general case of J items, the middle element of the vector on the right-hand side of equation (36) (i.e., $m_a^0 m_b^1 + m_a^0 m_c^1 + m_b^0 m_c^1$) equals $\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} m_i^0 m_j^1$. Similarly, for the general case of J items the lower element of the vector on the right-hand side of equation (36) (i.e., $m_{ab}^{01} + m_{ac}^{01} + m_{bc}^{01}$) equals $\sum_{i=1}^{J-1} \sum_{j=i+1}^{J} m_{ij}^{01}$.

Design matrix $\mathbf{A}_4$ is a row vector with three elements, that is,

$$\mathbf{A}_4 = (1 \quad -1 \quad 1). \tag{37}$$

Using equation (36), it may be verified that for three items (a, b, and c) ordered according to decreasing popularity, we have that $\exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A_1 m}))))$ equals

$$\exp\left[(1 \quad -1 \quad 1)\log\begin{pmatrix} n \\ m_a^0 n_b^1 + m_a^0 m_c^1 + m_b^0 m_c^1 \\ m_{ab}^{01} + m_{ac}^{01} + m_{bc}^{01} \end{pmatrix}\right] = \frac{n(m_{ab}^{01} + m_{ac}^{01} + m_{bc}^{01})}{m_a^0 m_b^1 + m_a^0 m_c^1 + m_b^0 m_c^1}. \tag{38}$$

For the general case of J items, the ratio on right-hand side of equation (38) equals

$$\frac{n\left(\sum_{i=1}^{J-1}\sum_{j=i+1}^{J} m_{ij}^{01}\right)}{\sum_{i=1}^{J-1}\sum_{j=i+1}^{J} m_i^0 m_j^1}.$$

For three items, it may be verified that substituting the term $\exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A_1 m}))))$ in equation (21) with the right-hand side of equation (38) produces coefficient H as defined in equation (10).

Appendix F. Recursive Exp-Log Notation for the Scalability Coefficient for a Set of Items for Several Groups Simultaneously

In Appendix F the recursive exp-log notation for equations (24) and (25) are derived. The recursive exp-log notation for the vector containing scalability coefficients $H^1, \ldots, H^G$, requires four design matrices that are the same for each subgroup: $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$, and $\mathbf{A}_5$, and one design matrix that may be different for each subgroup: $\mathbf{A}_1^g$ $(g = 1, \ldots, G)$. Design matrix $\mathbf{A}_1^g$ (derived in Appendix C, equation (28)) identifies the frequencies pertaining to Guttman errors. If the subgroup g has a different item ordering than subgroup g' $(g \neq g')$, then the cells in the contingency table that pertain to Guttman errors are not the same for g and g' and $\mathbf{A}_1^g \neq \mathbf{A}_1^{g'}$.

Design matrix $\mathbf{A}_2$ was derived in Appendix D (equation (31)), and design matrices $\mathbf{A}_3$ and $\mathbf{A}_4$ were derived in Appendix E (equations (35) and (37), respectively). For a single group, H^1 is given by equation (21). Using equation (21), it may be verified that, for two groups,

$$\begin{pmatrix} H^1 \\ H^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \exp\left[\begin{pmatrix} \mathbf{A}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_4 \end{pmatrix} \times \log\left[\begin{pmatrix} \mathbf{A}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_3 \end{pmatrix} \exp\left[\begin{pmatrix} \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix} \log\left[\begin{pmatrix} \mathbf{A}_1^1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1^2 \end{pmatrix} \begin{pmatrix} \mathbf{m}^1 \\ \mathbf{m}^2 \end{pmatrix}\right]\right]\right]\right]. \quad (39)$$

A generalization of equation (39) to G subgroups yields equation (24).

Let $\mathbf{A}_5$ be a $(G-1) \times G$ matrix with element (g, g) equal to 1 for $g = 1, \ldots, G-1$, and element $(g, g+1)$ equal to -1 for $g = 1, \ldots, G-1$; then

$$\mathbf{A}_5 \begin{pmatrix} H^1 \\ H^2 \\ \vdots \\ H^G \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix} \begin{pmatrix} H^1 \\ H^2 \\ \vdots \\ H^G \end{pmatrix} = \begin{pmatrix} H^1 - H^2 \\ H^2 - H^3 \\ \vdots \\ H^{G-1} - H^G \end{pmatrix}. \quad (40)$$

The marginal model that implies that the vector on the right-hand side of equation (40) equals 0 can be found by substituting the vector $(H^1, \ldots, H^G)^{\mathrm{T}}$ in equation (40) with the right-hand side of equation (24), and setting it equal to $\mathbf{0}_G$:

$$\mathbf{g}(\mathbf{m}^*) = \mathbf{A}_5 \left\{ \mathbf{u} - \exp\left(\bigoplus_{g=1}^{G} \mathbf{A}_4^g \log\left(\bigoplus_{g=1}^{G} \mathbf{A}_3^g \exp\left(\bigoplus_{g=1}^{G} \mathbf{A}_2^g \log\left(\bigoplus_{g=1}^{G} \mathbf{A}_1^g \mathbf{m}^* \right) \right) \right) \right) \right\} = \mathbf{0}_G. \quad (41)$$

Because $\mathbf{A}_5 \mathbf{u} = \mathbf{0}$ equation (41) reduces to equation (25).

References

Bartolucci, F., & Forcina, A. (2002). Extended RC association models allowing for order restrictions and marginal modeling. *Journal of the American Statistical Association*, *97*, 1192–1199.

Bartolucci, F., Forcina, A., & Dardanoni, V. (2001). Positive quadrant dependence and marginal modeling in two-way dependence with ordered margins. *Journal of the American Statistical Association*, *96*, 1497–1505.

Bergsma, W.P. (1997a). *Marginal models for categorical data*. Tilburg: Tilburg University Press. http://stats.lse.ac.uk/bergsma/pdf/bergsma_phdthesis.pdf.

Bergsma, W.P. (1997b). marg_mod.nb [*Mathematica computer code*]. Retrieved from http://www.uvt.nl/mto/software2.html.

Bergsma, W.P., & Croon, M.A. (2005). Analyzing categorical data by marginal models. In L.A. van der Ark, M.A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 83–101). Mahwah, NJ: Erlbaum.

Bergsma, W.P., & Rudas, T. (2002). Marginal models for categorical data. *The Annals of Statistics*, *30*, 140–159.

Ellis, J.L., & Van den Wollenberg, A.L. (1993). Local homogeneity in latent trait models: A characterization of the homogeneous monotone latent trait model. *Psychometrika*, *58*, 417–429.

Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, *49*, 732–764.

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

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.

Hemker, B.T., Sijtsma, K., & Molenaar, I.W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*, 337–352.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331–347.

Holland, P.W., & Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*, 1523–1543.

Junker, B.W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65–81.

Kritzer, H.M. (1977). Analyzing measures of association derived from contingency tables. *Sociological Methods and Research*, *5*, 35–50.

Kullback, S. (1971). Marginal homogeneity of multidimensional contingency tables. *Annals of Mathematical Statistics*, *42*, 594–606.

Lang, J.B., & Agresti, A. (1994). Simultaneously modeling the joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, *89*, 625–632.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin*, *45*, 507–529.

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

Molenaar, I.W. (1997). Nonparametric models for polytomous responses. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer.

Molenaar, I.W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows* [software manual]. Groningen, The Netherlands: iec ProGAMMA.

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630.

Read, T.R.C., & Cressie, N.C. (1988). *Goodness of fit statistics for discrete multivariate analysis*. New York: Springer.

Rudas, T., & Bergsma, W.P. (2004). On applications of marginal models for categorical data. *Metron*, *62*, 1–23.

Scheiblechner, H. (2007). A unified nonparametric IRT model for d-dimensional psychological test data (d-ISOP). *Psychometrika*, *72*, 43–67.

Sijtsma, K., & Meijer, R.R. (2007). Nonparametric item response theory and related topics. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics*: Vol. 26. *Psychometrics* (pp. 719–746). Amsterdam: Elsevier.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Stout, W.F. (1990). A new item response modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.

Van Abswoude, A.A.H., Van der Ark, L.A., & Sijtsma, K. (2004). A comparative study of test dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*, 3–24.

Van der Ark, L.A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19.

Van Maanen, L., Been, P.H., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E.E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–288). Berlin: Springer.

Wolfram, S. (1999). *The Mathematica book* (4th ed.). Cambridge: Wolfram Media/Cambridge University Press.

Manuscript received 27 FEB 2007
Final version received 25 JUN 2007
Published Online Date: 8 NOV 2007