Scalability Coefficients for Two-Level Polytomous Item Scores: An Introduction and an Application

Daniela R. Crisan, Janneke E. van de Pol, and L. Andries van der Ark

Abstract First, we made an overview of nonparametric item response models and the corresponding scalability coefficients in Mokken scale analysis for single-level item scores and two-level dichotomous item scores. Second, we generalized these models and coefficients to two-level polytomous item scores. Third, we applied the new scalability coefficients to a real-data example, and compared the outcomes with results obtained using single-level reliability analysis and single-level Mokken scale analysis. Results suggest that coefficients from single-level analyses do not provide accurate information about scalability of two-level item scores.

Keywords Mokken scale analysis • Multilevel analysis • Nonparametric item response theory • Scalability coefficients

1 Introduction

For most tests, a single rater provides the item scores that are used to estimate a particular subject' trait value. Typically, the rater and the subject are the same person but for several clinical or pedagogical tests the rater may be, for example, the parent or the supervisor of the subject. The item scores are not nested and called single-level item scores. For some tests, multiple raters provide the item scores that are used to estimate a particular subject's trait value. Examples include teachers whose

D.R. Crisan (✉)
Department of Psychometrics and Statistics, University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, The Netherlands
e-mail: d.r.crisan@rug.nl

J.E. van de Pol
Department of Education, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands
e-mail: j.e.vandepol@uu.nl

L.A. van der Ark
Research Institute of Child Development and Education, University of Amsterdam, P.O. Box 15776, 1001 NG, Amsterdam, The Netherlands
e-mail: l.a.vanderark@uva.nl

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_11

teaching skills are rated by all students in the classroom; hospitals for which the quality of health care is rated by multiple patients; or students whose essays are rated by multiple assessors. In these cases, the raters are nested within the subjects, and the resulting item scores are called two-level item scores.

Nonparametric item response theory (NIRT) models are flexible unidimensional item response theory (IRT) models that are characterized by item response functions that do not have a parametric form. For an introduction to NIRT models, we refer to Sijtsma and Molenaar (2002). NIRT models have been defined for dichotomous single-level item scores (Mokken 1971), polytomous single-level item scores (Molenaar 1997), and dichotomous two-level item scores (Snijders 2001), but not yet for polytomous two-level item scores.

NIRT models are attractive for two reasons. First, for single-level dichotomous item scores, NIRT models allow stochastic ordering of the latent trait by means of the unweighed sum score of the test (Grayson 1988; Hemker, Sijtsma, Molenaar & Junker 1997). This in an attractive property because for most tests the unweighed sum scores is used as a measurement value. For polytomous single-level item scores, NIRT models imply a weak form of stochastic ordering (Van der Ark & Bergsma 2010). It is unknown whether these properties carry over NIRT models for two-level item scores. Second, there are many methods available to investigate the fit of NIRT models (Mokken 1971; Sijtsma & Molenaar 2002; Van der Ark 2007). Because all well-known unidimensional item response models are a special case of the nonparametric graded response model (a NIRT model for single-level polytomous item scores) (Van der Ark 2001), investigating the fit of NIRT models is a logical first step in parametric IRT modelling: If the nonparametric graded response model does not fit, parametric IRT models will not fit either.

The set of methods to investigate the fit of NIRT models are called *Mokken scale analysis*. The most popular coefficients from Mokken scale analysis are the scalability coefficients (Mokken 1971). For a set of I items, there are $I(I-1)/2$ item-pair scalability coefficients H_{ij}, I item scalability coefficients H_i, and one total scalability coefficient H. Coefficient H reflects the accuracy of the ordering of persons using their sum scores (Mokken, Lewis & Sijtsma 1986); hence, the larger H, the more accurate is the ordering.

The remainder of this paper is organized as follows. First, we discuss NIRT models and scalability coefficients for dichotomous single-level, polytomous single-level, and dichotomous two-level item scores. Second, we generalize the NIRT model and scalability coefficients to polytomous two-level item scores, demonstrate how the scalability coefficients are estimated, and briefly discuss results from a simulation study investigating the scalability coefficients for both dichotomous and polytomous item scores (Crisan 2015). Third, we present a real-data example: We analyzed two-level polytomous item scores from the Appreciation of Support Questionnaire (Van de Pol, Volman, Oort & Beishuizen 2015), and compared the outcomes with results obtained using traditional reliability analysis. Finally, we elaborate on the implications of our findings and discuss future research directions.

2 NIRT Models and Scalability Coefficients

Let a test consists of I items, indexed by i or j. Let each item have $m + 1$ ordered response categories scored $0, \ldots, m$ indexed by x or y. If $m = 1$, the items scores are dichotomous, if $m > 1$ the item scores are polytomous. Suppose the test is used to measure the trait level of S subjects, indexed by s or t, and subject s has been rated by R_s raters, indexed by p or r. If $R_s = 1$ for all subjects, we have single-level item scores, and the index for the rater is typically omitted. Furthermore. Let X_{sri} denote the score of subject s by rater r on item i, and let X_{s++} denote the total score of subject s; that is, $X_{s++} = \sum_{i=1}^{I} \sum_{r=1}^{R_s} X_{sri}$. Finally, let θ denote a latent trait driving the item responses, and let θ_s denote the latent trait value of subject s.

2.1 NIRT Models and Scalability Coefficients for Single-Level Dichotomous Item Scores

The monotone homogeneity model (MHM) (Mokken 1971; Molenaar 1997; Sijtsma & Molenaar 2002) is a NIRT model for single-level dichotomous item scores. $P(X_{si} = x_{si}|\theta_s)$ denote the probability that subject s has score $x_{si} \in \{0, 1\}$ on item i. The MHM consists of three assumptions.

- Unidimensionality: θ is unidimensional;
- Local independence: item-scores are independent conditional on θ, that is,

$$P(X_{s1} = x_{s1}, X_{s2} = x_{s2}, \ldots, X_{sI} = x_{sI}|\theta_s) = \prod_{i=1}^{I} P(X_{si} = x_{si}|\theta_s); \qquad (1)$$

- Monotonicity: For each item i, there is a nondecreasing function $p_i(\cdot)$ such that the probability of obtaining item score 1 given latent trait value θ_s is $p_i(\theta_s) = P(X_{si} = 1|\theta_s)$.

Function $p_i(\theta)$ is known as the *item response function*. Under the MHM, item response function are allowed to intersect. If, additionally to the three assumptions, the restriction of non-intersecting of the IRFs is imposed, then the more restrictive double monotonicity model is defined (Mokken 1971).

The scalability coefficients are based on the Guttman model. Without loss of generality, let the I items be put in descending order of mean item score and be numbered accordingly, so that $P(X_i = 1) > P(X_j = 1)$ for $i < j$. The Guttman model does not allow that the easier (more popular) item has score 0 and the more difficult (less popular) item has score 1, and thus excludes item-score pattern $(X_i, X_j) = (0, 1)$, which is known as a *Guttman error*. For items i and j, let $F_{ij} = P(X_i = 0, X_j = 1)$ denote the probability of obtaining a Guttman error, and

let $E_{ij} = P(X_i = 0)P(X_j = 1)$ denote the expected probability of a Guttman error under marginal independence. Item-pair scalability coefficient H_{ij} is then defined as

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}. \tag{2}$$

If the MHM holds $0 \leq H_{ij} \leq 1$ for all $i \neq j$. H_{ij} equals the ratio of the covariance of X_i and X_j and the maximum covariance of X_i and X_j given the marginal item score distribution. Item scalability coefficient H_i is

$$H_i = 1 - \frac{\sum_{i \neq j} F_{ij}}{\sum_{i \neq j} E_{ij}}. \tag{3}$$

If the MHM holds $0 \leq H_i \leq 1$ for all i. H_i can be viewed as a nonparametric analogue of the discrimination parameter (Van Abswoude, Van der Ark & Sijtsma 2004). As a heuristic rule for inclusion in a scale, H_i is often required to exceed 0.3. Finally, total-scale scalability coefficient H is

$$H = 1 - \frac{\sum_i \sum_j F_{ij}}{\sum_i \sum_j E_{ij}}. \tag{4}$$

As a heuristic rule, $0.3 < H \leq 0.4$ is considered a weak scale, $0.4 < H \leq 0.5$ is considered a moderate scale, and $H > 0.4$ is considered a strong scale.

2.2 *NIRT Models and Scalability Coefficients for Single-Level Polytomous Item Scores*

The nonparametric graded response model (a.k.a. the MHM for polytomous items (Molenaar 1997) is the least restrictive NIRT model for polytomous items. As the MHM, it consists of the assumptions unidimensionality, local independence, and monotonicity but monotonicity is defined differently. For item score x ($x = 1, \ldots, m$) for each item i there is a nondecreasing function $p_{ix}(\cdot)$ such that the probability of obtaining at least item score x given latent trait value θ_s is $p_{ix}(\theta_s) = P(X_{si} \geq x|\theta_s)$. Function $p_{ix}(\theta)$ is known as the *item step response function*. Under the nonparametric graded response model, ISRFs from the same item cannot intersect by definition but ISRFs from different items are allowed to intersect. If, additionally to the three assumptions the restriction of non-intersecting of the ISRFs is imposed, then we have the more restrictive double monotonicity model for polytomous items (Molenaar 1997).

Scalability coefficients for polytomous item scores are more complicated than for dichotomous item scores, which are a special case. They are best explained using an

Table 1 Frequency table for two polytomous items with three response categories

		Item 2			
	Response	0	1	2	$P(X_1 \geq x)$
Item 1	0	**2** (0)	1 (2)	0 (4)	1
	1	**3** (0)	0 (1)	0 (2)	3/4
	2	**3** (0)	**2** (0)	**1** (0)	1/2
$P(X_2 \geq x)$		1	1/3	1/12	

Note: Frequencies not pertaining to Guttman errors are in boldface, frequencies pertaining to Guttman errors are in normal font, Guttman weights are between parentheses. The last row and column show the marginal cumulative probabilities

example. Table 1 contains the scores of 12 subjects on two items, each having three ordered answer categories.

First, Guttman errors are determined. *Item steps* (Molenaar 1983) $X_i \geq x (i = 1, \ldots, I; x = 1, \ldots, m)$ are boolean expressions indicating whether or not an item score is at least x. $P(X_i \geq x)$ defines the popularity of item step $X_i \geq x$. The item steps are placed in descending order of popularity. For the data in Table 1, the order of the item-steps is:

$$X_1 \geq 1, X_1 \geq 2, X_2 \geq 1, X_2 \geq 2. \tag{5}$$

Items steps $X_1 \geq 0$ and $X_2 \geq 0$ are omitted because, by definition, $P(X_1 \geq 0) = P(X_2 \geq 0) = 1$. Item-score pattern (x, y) is a Guttman error if an item step that has been passed is preceded by an item step that has not been passed. Let z_g^{xy} indicate whether (score 1) or not (score 0) the gth ordered item step has been passed for item-score pattern (x, y). The values of z_g^{xy} are collected in vector $\mathbf{z}^{xy} = (z_1^{xy}, \ldots, z_G^{xy})$. To obtain item-score pattern $(0, 2)$ in Table 1, a subject must have passed item steps $X_2 \geq 1$ and $X_2 \geq 2$ but not item steps $X_1 \geq 1$ and $X_1 \geq 2$. Hence, for item-score pattern $(0, 2)$, $\mathbf{z}^{02} = (0, 0, 1, 1)$. Because item steps that have been passed are preceded by items steps that have not been passed, $(0, 2)$ is identified as a Guttman error. Similarly, for item-score pattern $(2, 1)$, $\mathbf{z}^{21} = (1, 1, 1, 0)$ and item-score pattern $(2, 1)$ is not a Guttman error. In Table 1, the four item-score patterns for which the frequencies are printed in normal font are Guttman errors, whereas the frequencies printed in bold font are not.

Second, the frequencies of the item-score patterns are weighed (Molenaar 1991); the weight being equal to the number of times an item step that has not been passed preceded an item step that has been passed. Weight w_{ij}^{xy} equals

$$w_{ij}^{xy} = \sum_{h=2}^{G} \left\{ z_h^{xy} \times \left[\sum_{g=1}^{h-1} (1 - z_g^{xy}) \right] \right\} \tag{6}$$

(Kuijpers, Van der Ark & Croon 2013; Ligtvoet, Van der Ark, te Marvelde & Sijtsma 2010). For example, for item-score pattern (0, 2), $\mathbf{z}^{02} = (z_1^{02}, z_2^{02}, z_3^{02}, z_4^{02}) = (0, 0, 1, 1)$. Using Eq. (6), the weight equals $w_{ij}^{02} = 4$. Table 1 shows the weights between parentheses.

Item-pair scalability coefficient H_{ij} for polytomous items is

$$H_{ij} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} P(X_i = x, X_j = y)}{\sum_x \sum_y w_{ij}^{xy} P(X_i = x) P(X_j = y)} \tag{7}$$

(Molenaar 1991). Because item-score patterns that are not Guttman errors have weight 0, the probabilities pertaining to these patterns do not count, and the numerator of Eq. (7) is simply the sum of observed weighed Guttman errors, and the denominator the sum of expected weighed Guttman errors. Similarly, item scalability coefficient H_i for polytomous items is

$$H_j = 1 - \frac{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x, X_j = y)}{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x) P(X_j = y)}, \tag{8}$$

and the total scale scalability coefficient H is

$$H = 1 - \frac{\sum \sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x, X_j = y)}{\sum \sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x) P(X_j = y)}. \tag{9}$$

Note that for dichotomous items, the Guttman error receives a weight 1, and Eqs. (7)–(9) reduce to Eqs. (2)–(4), respectively. In Table 1, because there are only two items, $H_{12} = H_1 = H_2 = H = 0.50$.

2.3 NIRT Models and Scalability Coefficients for Two-Level Dichotomous Item Scores

Snijders (2001) generalized the MHM for dichotomous items to two-level data. As in the MHM, each subject has a latent trait value θ_s. In addition, rater r is assumed to have a deviation (δ_{sr}), so the latent trait value for subject s as rated by rater r is $\theta_s + \delta_{sr}$. Deviation δ_{sr} can be considered a random rater effect together with the subject by rater interaction. It is assumed that the raters are a random sample from the population of raters, so deviations δ_{sr} can be considered independent and randomly distributed variables. As the MHM, Snijders' model for two-level data assumes unidimensionality, local independence, and monotonicity for the item response functions $p_i(\theta_s + \delta_{sr}) = P(X_{sri} = 1|\theta_s; \delta_{sr})$. In addition, a second nondecreasing item response function is defined $\pi_i(\theta_s) = P(X_{si} = 1|\theta_s) = E_\delta[p_i(\theta_s + \delta_{sr})]$. If $p_i(\theta_s + \delta_{sr})$ is nondecreasing, then so is $\pi_i(\theta_s)$, yet $\pi_i(\theta_s)$ will be flatter.

Snijders generalized scalability coefficients for dichotomous items [Eqs. (2)–(4)] to two-level data, resulting in *within-rater* and *between-rater* scalability coefficients.[1] The within-rater scalability coefficients H_{ij}^W, H_i^W, and H^W are in fact equivalent to the scalability coefficients that were defined for the MHM [Eqs. (2)–(4), respectively], where every rater-subject combination is considered a separate case.

Snijders defined the between-rater item-pair scalability coefficients

$$H_{ij}^B = 1 - \frac{P(X_{sri} = 1, X_{spj} = 0)}{P(X_{sri} = 1)P(X_{srj} = 0)} (p \neq r). \tag{10}$$

The joint probability in the numerator is computed for pairs of different raters p and r ($p \neq r$) nested within the same subject s. More specifically, the numerator represents the joint probability that rater r assigns score 1 on item i to subject s and rater p assigns score 0 on item j to subject s. Because the denominator consists of a product of two probabilities that are independent of r, replacing r with p in the second term of the denominator would not make any difference: the expected proportion of Guttman errors under marginal independence remains the same. Using a similar line of reasoning, the item between-rater scalability coefficients are

$$H_i^B = 1 - \frac{\sum_{j \neq i} P(X_{sri} = 1, X_{spj} = 0)}{\sum_{j \neq i} P(X_{sri} = 1)P(X_{srj} = 0)} (p \neq r) \tag{11}$$

and

$$H^B = 1 - \frac{\sum \sum_{j \neq i} P(X_{sri} = 1, X_{spj} = 0)}{\sum \sum_{j \neq i} P(X_{sri} = 1)P(X_{srj} = 0)} (p \neq r). \tag{12}$$

Within-rater scalability coefficients are useful for investigating the quality of the test as a unidimensional cumulative scale for subject-rater combinations. The between-rater scalability coefficients and the ratio of the within- and between-rater scalability coefficients are useful for investigating the extent to which item responses are driven by the subjects trait value rather than by rater effects. If Snijders' model holds, $0 < H^B \leq H^W$ (Snijders 2001); and larger values indicate greater scalability. In the extreme case that there is no rater variation ($\delta_{rs} = 0$ for all r and all s), $H^B = H^W$. As a heuristic rule, Snijders suggested $H^B > 0.1$ and $H^W > 0.2$ to be reasonable. The ratio of the two scalability coefficients reflect the relative effect of the subjects and the raters. Low values indicate that the effect of raters is large and many raters per subject are required to scale the subjects. Snijders suggested $H^B/H^W \geq 0.3$ could be labelled reasonable and $H^B/H^W \geq 0.6$ excellent. The measurement for scaling subjects is the mean total score of a subjects across all raters: $\overline{X}_{s++}$.

[1] Terminology is ours; Snijders used within-subject and between-subject scalability.

3 A Generalization to Two-Level Polytomous Item Scores

Given the work on scalability coefficients for single-level polytomous item scores (Sect. 2.2) and two-level dichotomous item scores (Sect. 2.3), a generalization to two-level polytomous item scores is rather straightforward. The within-rater scalability coefficients for polytomous item scores are the same as the scalability coefficients for single-level polytomous item scores [Eqs. (7)–(9)] when considering all rater-subjects combinations as individual cases.

The between-rater scalability coefficients are defined as follows:

$$H_{ij}^{B} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{spj} = y)}{\sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x) P(X_{srj} = y)} (p \neq r), \tag{13}$$

$$H_{i}^{B} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{spj} = y)}{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x) P(X_{srj} = y)} (p \neq r), \tag{14}$$

and

$$H^{B} = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{spj} = y)}{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x) P(X_{srj} = y)} (p \neq r). \tag{15}$$

It may be verified that in case of dichotomous item scores Eqs. (13)–(15) reduce to Equations reduces to (10)–(12), respectively.

3.1 Estimation of the Scalability Coefficients

Snijders (2001) proposed estimators for the scalability coefficients for dichotomous item scores, by substituting the probabilities in their defining formulas by relative frequencies. If the number of raters per subject (R_s) is not the same for all subjects, then the probabilities required to compute the scalability coefficients can be estimated by averaging the relative frequencies across subjects. Snijders' estimators can be generalized to polytomous item scores. Let $\mathbf{1}(X_{sri} = x)$ denote the indicator function that $X_{sri} = x$, and let $\widehat{P}_i(x)$ be the estimator for $P(X_{sri} = x)$; then,

$$\widehat{P}_i(x) = \frac{1}{S} \sum_s \frac{1}{R_s} \sum_r \mathbf{1}(X_{sri} = x). \tag{16}$$

Equation (16) determines the proportions of raters per subject with a score x on item i and then averages these proportions across subjects, yielding the estimated probability of a score equal to x on item i.

The joint probabilities in the numerators of the scalability coefficients can be estimated as follows. Let $\widehat{P}_{ij}^{W}(x, y)$ denote the estimated within-rater joint probability

that $X_{sri} = x$ and $X_{srj} = y$, and let $\widehat{P}^B_{ij}(x, y)$ denote the estimated between-rater joint probability that $X_{sri} = x$ and $X_{spj} = y$. Then,

$$\widehat{P}^W_{ij}(x, y) = \frac{1}{S} \sum_s \frac{1}{R_s} \sum_r \mathbf{1}(X_{sri} = x, X_{srj} = y), \tag{17}$$

and

$$\widehat{P}^B_{ij}(x, y) = \frac{1}{S} \sum_s \frac{1}{R_s(R_s - 1)} \sum\sum_{p \neq r} \mathbf{1}(X_{sri} = x, X_{spj} = y). \tag{18}$$

Finally, substituting the probabilities in the defining formulas of the scalability coefficients with the estimators in Eqs. (16)–(18) leads to the following estimators of the within- and between-subject scalability coefficients:

$$\widehat{H}^W_{ij} = 1 - \frac{\sum_x \sum_y w^{xy}_{ij} \widehat{P}^W_{ij}(x, y)}{\sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{19}$$

$$\widehat{H}^B_{ij} = 1 - \frac{\sum_x \sum_y w^{xy}_{ij} \widehat{P}^B_{ij}(x, y)}{\sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{20}$$

$$\widehat{H}^W_i = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^W_{ij}(x, y)}{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{21}$$

$$\widehat{H}^B_i = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^B_{ij}(x, y)}{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{22}$$

$$\widehat{H}^W = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^W_{ij}(x, y)}{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{23}$$

and

$$\widehat{H}^B = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^B_{ij}(x, y)}{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}. \tag{24}$$

Example 1 illustrates the computation of the scalability coefficients.

Example 1. Table 2 (upper panel) shows the frequencies of the scores on 2 items, each having 3 ordered response categories, assigned by 12 raters to 3 subjects: Four raters rated subject 1 ($R_1 = 4$), two raters rated subject 2 ($R_2 = 3$), and five raters rated subject 3 ($R_3 = 5$). Frequencies equal to zero are omitted. These frequencies equal $\sum_r \mathbf{1}(X_{sri} = x, X_{srj} = y)$ and are required for computing $\widehat{P}^W_{ij}(x, y)$ (Eq. (17); values in last row of Table 2, upper panel). For example, $\hat{P}^W_{12}(0, 0) = \frac{1}{3}(\frac{1}{4} \times 2 + 0 + 0) \approx 0.17$.

Table 2 Frequencies of observed item-score patterns per subject (upper panel), frequencies of observed item-score patterns where each item-score in a pattern is assigned by a different rater for each subject (middle panel), and marginal frequencies of observed item-score patterns per subject (lower panel)

	Item-score pattern (x, y)									
s	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	R_s
1	2	1		1						4
2				1				2		3
3				1			3		1	5
$\hat{P}^W_{12}(x, y)$	0.17	0.08	0.00	0.26	0.00	0.00	0.20	0.22	0.07	

	Item-score pattern (x, y)									
s	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	$R_s(R_s - 1)$
1	7	2		2		2				12
2							2	2		6
3				3		1	13		3	20
$\hat{P}^B_{12}(x, y)$	0.19	0.06	0.00	0.11	0.00	0.07	0.33	0.11	0.05	

	Item 1			Item 2			
s	$x = 0$	$x = 1$	$x = 2$	$x = 0$	$x = 1$	$x = 2$	R_s
1	3	1		3	1		4
2		1	2	1	2		3
3		1	4	4		1	5
$\hat{P}_i(x)$	0.25	0.26	0.49	0.63	0.31	0.07	

Note: unobserved item-score patterns are left blank

Table 2 (middle panel) shows the frequencies of the item-score patterns assigned by different raters (e.g., $\sum\sum_{p\neq r} \mathbf{1}(X_{sri} = x, X_{spj} = y)$). For example, score 7 (first row, first column) is obtained as follows. Subject 1 received four item-score patterns: (0,0); (0,0); (0,1); and (1,0). Within these four patterns, it occurs 7 times that one rater has score 0 on item 1 and a different rater has score 0 on item 2. Then, $\hat{P}^B_{12}(0, 0) = \frac{1}{3}(\frac{1}{12} \times 7 + 0 + 0) \approx 0.19$.

Table 2 (lower panel) shows he marginal frequencies of the item scores for each subject (i.e., $\sum_r \mathbf{1}(X_{sri} = x)$), required for estimating $\widehat{P}_i(x)$ [Eq. (16)]. For example, $\hat{P}_1(0) = \frac{1}{3} \times (\frac{1}{4} \times 3 + 0 + 0) = 0.25$. Using the weights from Table 1 yields $\hat{H}^W_{12} = \widehat{H}^W_1 = \widehat{H}^W_2 = \widehat{H}^W = 0.50$, and $\hat{H}^B_{12} = \widehat{H}^B_1 = \widehat{H}^B_2 = \widehat{H}^B = 0.15$.

3.2 Results from a Simulation Study

Crisan (2015) performed a simulation study to the effect of item discrimination, number of ordered answer categories, the variance ratio of θ and δ, the number of subjects, and the number of raters per subject on the magnitude of $\hat{H}_W$, $\hat{H}_B$, and the ratio of $\hat{H}_B$ and $\hat{H}_W$. We briefly reiterate the main results here.

The variance ratio of θ and δ had an extremely large positive effect on the magnitude of $\widehat{H}_B$ ($\eta^2 = 0.985$) and $\hat{H}_B/\hat{H}_W$ ($\eta^2 = 0.558$), whereas item discrimination had an extremely large positive effects on the magnitude $\hat{H}_W$ ($\eta^2 = 0.766$) and $\hat{H}_B$ ($\eta^2 = 0.280$). Finally number of ordered answer categories had a very large positive effect of the magnitude of $\widehat{H}_W$. The variance ratio of θ and δ and number of subjects had the largest effects on the precision of the estimated values of $\hat{H}_W$, $\hat{H}_B$, and $\hat{H}_B/\hat{H}_W$.

4 Real-Data Example

We analyzed item scores of the Appreciation of Support Questionnaire (ASQ) (Van de Pol et al. 2015). The ASQ consists of 11 polytomously scored items (Translated items in Table 3). For each item, the scores ranged from 0 ("I don't agree at all") to 4 ("I totally agree"). The data came from an experimental study on the effects of scaffolding on prevocational students' achievement, task effort, and appreciation of support (Van de Pol et al. 2015). Six hundred fifty nine grade-8 students in The Netherlands, nested in 30 teachers, used the ASQ to express their appreciation of their own teacher's support. The number of students per teacher ranged from 12 to 46 ($M = 21.97$, $SD = 5.91$).

We conducted traditional reliability analysis, traditional Mokken scale analysis, and two-level Mokken scale analysis. Traditional reliability analysis and traditional Mokken scale analysis are inappropriate analyses for these data. However, they

Table 3 The items if the appreciation of support questionnaire

Item	Content	*M*	*SD*	IRC
1	The advice that this teacher gave me and my group was very helpful	2.53	1.00	0.70
2	Because of the way in which this teacher helped me and my group, I could focus on my work with ease	2.24	1.02	0.67
3	I felt the teacher took me seriously because of the way he/she helped me and my group	2.75	0.97	0.61
4	Because of the way this teacher helped me and my group, I could really learn new things	2.37	1.03	0.71
5	Because of the way this teacher helped me and my group, I made an effort	2.42	0.93	0.71
6	The way in which this teacher helped me and my group really worked for me	2.22	0.98	0.72
7	I could really use the help that this teacher offered	2.49	1.01	0.75
8	I worked hard with this teacher	2.37	0.98	0.67
9	The way in which this teacher helped me and my group was pleasant	2.46	1.03	0.77
10	The explanation and help of this teacher was really helpful	2.39	0.99	0.77
11	Because of the explanation and help of this teacher, I could proceed	2.48	1.03	0.71

Note: M = Mean, SD = standard deviation, *IRC* = item rest correlation

are used to demonstrate the different outcomes. All analyses were conducted in `R` (R Core Team 2015) using the packages `psych` (Revelle 2015) and `CTT` (Willse 2014) for traditional reliability analysis, `mokken` (Van der Ark 2007) for one-level Mokken scale analysis, and code available from the first author for two-level Mokken scale analysis.

4.1 Reliability Analysis

In traditional reliability analysis the nested structure is ignored. The descriptive statistics of the item scores were all similar: mean item scores ranged between 2.22 and 2.75, the item standard deviations ranged between 0.97 and 1.03, and the item rest correlations ranged between 0.61 and 0.75 (Table 3). Cronbach's alpha was 0.93. These results suggest a very reliable test score with no indication that items should be revised. The test score had mean $M = 26.72$, standard deviation $SD = 8.41$.

4.2 One-Level Mokken Scale Analysis

In one-level Mokken scale analysis, the nested structure is also ignored. Table 4 shows the item-pair and item scalability coefficients plus standard errors (Kuijpers et al. 2013). Because all item-pair scalability coefficients were greater than 0, and all item scalability coefficients are greater than default lower bound $c = 0.3$, the 11 items form a Mokken scale. The total scalability coefficient equalled $H = 0.58(0.02)$, which qualifies as a strong scale. In addition, we investigated monotonicity using the method *manifest monotonicity* (Junker & Sijtsma 2000), local independence using Ellis' theoretical upper and lower bounds (Ellis 2014), and non-intersection using the method *pmatrix* (Mokken 1971). We found no evidence of any substantial violation of the MHM and the double monotonicity model.

4.3 Two-Level Mokken Scale Analysis

From the single-level Mokken scale analysis we concluded that the assumptions of the double monotonicity model are reasonable. The within-rater scalability coefficients are the same as the scalability coefficients in single-level Mokken scale analysis (Table 4). The between-rater scalability coefficients (Table 5; upper diagonal and penultimate row) are greater than Snijder's heuristic lower bound 0.1 suggesting a satisfactory consistency between the raters. The total-scale between-rater scalability coefficient equalled $H^B = 0.14$. The ratio of the between and

Table 4 Scalability coefficients and standard errors for the appreciation of support questionnaire

	Item										
Item	1	2	3	4	5	6	7	8	9	10	11
1		0.60	0.55	0.60	0.50	0.58	0.64	0.47	0.58	0.60	0.57
2	0.04		0.49	0.53	0.62	0.52	0.55	0.60	0.58	0.57	0.50
3	0.04	0.04		0.53	0.51	0.54	0.56	0.53	0.58	0.52	0.52
4	0.04	0.04	0.04		0.57	0.60	0.57	0.52	0.60	0.60	0.54
5	0.04	0.03	0.04	0.03		0.64	0.60	0.67	0.62	0.59	0.53
6	0.04	0.04	0.04	0.03	0.03		0.61	0.58	0.70	0.68	0.57
7	0.03	0.04	0.04	0.04	0.03	0.03		0.54	0.67	0.63	0.67
8	0.04	0.03	0.04	0.04	0.03	0.03	0.04		0.57	0.56	0.50
9	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03		0.68	0.60
10	0.04	0.03	0.04	0.03	0.03	0.03	0.04	0.03	0.03		0.67
11	0.03	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.03	0.03	
H_i	0.57	0.56	0.53	0.57	0.58	0.60	0.60	0.55	0.62	0.61	0.57
SE	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03

Note: item-pair scalability coefficients H_{ij} are in the upper-triangular matrix, the standard errors in the lower-triangular matrix. Item scalability coefficients H_i and standard errors are in the last two rows

Table 5 Between-subject H coefficients for the appreciation of support questionnaire

	Item										
Item	1	2	3	4	5	6	7	8	9	10	11
1		0.16	0.13	0.17	0.15	0.17	0.15	0.18	0.16	0.16	0.13
2	0.27		0.11	0.14	0.15	0.13	0.13	0.15	0.15	0.14	0.12
3	0.23	0.23		0.13	0.12	0.11	0.10	0.12	0.11	0.11	0.09
4	0.27	0.25	0.24		0.15	0.14	0.14	0.16	0.15	0.15	0.13
5	0.30	0.25	0.24	0.25		0.14	0.12	0.16	0.15	0.12	0.11
6	0.29	0.24	0.20	0.23	0.22		0.14	0.16	0.14	0.16	0.12
7	0.23	0.24	0.18	0.24	0.21	0.23		0.15	0.14	0.14	0.12
8	0.39	0.25	0.22	0.31	0.24	0.28	0.28		0.17	0.15	0.13
9	0.27	0.26	0.19	0.25	0.24	0.20	0.21	0.30		0.15	0.13
10	0.27	0.24	0.21	0.25	0.21	0.23	0.22	0.27	0.22		0.13
11	0.23	0.24	0.18	0.24	0.21	0.21	0.18	0.26	0.21	0.19	
H_i^B	0.16	0.14	0.11	0.14	0.14	0.14	0.13	0.15	0.15	0.14	0.12
H_i^B/H_i^W	0.27	0.25	0.21	0.25	0.23	0.23	0.22	0.28	0.23	0.23	0.21

Note: item-pair scalability coefficients H_{ij}^B are in the upper-triangular matrix, the ratio of H_{ij}^B and H_{ij}^W in the lower-triangular matrix. Item scalability coefficients H_i^W and H_i^B/H_i^W are in the last two rows

within scalability coefficients (lower diagonal and last row) ranged from 0.18–0.27. All values are less than 0.3, (Snijder's heuristic value of a reasonable scale). This suggests that the rater deviation is relatively large and more students may be required for the scaling of these teachers. The results from the two-level scaling analysis shows a less bright picture than the results from the one-level analyses. Finally, the mean and standard deviation of the subject scores $\overline{X}_s$ were $M = 26.8$ and $SD = 4.35$, respectively.

5 Discussion

This chapter presented a first step in reviving Mokken scale analysis for two-level data, a method that has been largely ignored since its introduction 15 years ago. Our main contribution is the generalization of Snijder's (Snijders 2001) scalability coefficients to polytomous items. We have some reservations because the scalability coefficients for two-level polytomous data were derived by analogy, and without formal proof that the properties of the scalability coefficients for two-level polytomous item scores behave as one would expect under a two-level polytomous NIRT model.

Furthermore, using guidelines from Snijders (2001) and Crisan (2015) in the analysis of a real-data example, we showed that ignoring the two-level structure may result in at least two problems: First, single-level analyses provide information about the raters' scores rather than the subjects scores, whereas the interest is in scaling subjects, not raters. This problem has not always been acknowledged. Second, interpreting the quality of the scale using single-level statistics may give an that is too optimistic. Therefore, it is important that Mokken scale analysis for two-level data is developed further. A possible next step is the derivation of standard errors for the scalability coefficients proposed in this paper. If that has been accomplished the bias and variance of both the point estimates and standard errors can be investigated. Second, it would be interesting to investigate whether other methods in Mokken scale analysis can be generalized to multi-level data. As a start, Snijders proposed using the intra-subject correlation coefficient to assess reliability in two-level item scores, which has been generalized to polytomous items by Crisan (2015). Finally, the current methods should be further extended so that a rater is allowed to assess multiple subjects, and the methods should be implemented in software; both would increase the range of possible applications.

Acknowledgements We would to thank Letty Koopman for commenting on the first draft of the paper.

References

Crisan, D. R. (2015). *Scalability coefficients for two-level dichotomous and polytomous data: A simulation study and an application* Unpublished master's thesis. Tilburg University, Tilburg.

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika, 79*, 303–316.

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

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*, 331–217.

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

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43*, 42–69.

Ligtvoet, R., Van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578–595.

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

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

Molenaar, I. W. (1983). *Item steps*. (Heymans Bulletins HB-83-630-EX). Groningen: University of Groningen.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12*(37), 97–117.

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, NY: Springer.

R Core Team (2015). R: A language and environment for statistical computing [computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Revelle, W. (2015). Psych: Procedures for personality and psychological research [computer software]. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psychVersion=1.5.8.

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

Snijders, T. A. B. (2001). Two-level nonparametric scaling for dichotomous data. In A. Boomsma, M. A. J. van Duijn, & T. A B. Snijders (Eds.), *Essays on item response theory* (pp. 319–338). New York, NY: Springer.

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

Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2015). The effects of scaffolding in the classroom: support contingency and student independent working time in relation to student achievement, task effort and appreciation of support. *Instructional Science, 43*, 615–641.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*, 273–282.

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

Van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*, 272–279.

Willse, J. T. (2014). CTT: Classical test theory functions. R package version 2.1 [computer software]. Retrieved from http://CRAN.R-project.org/package=CTT.