Nonparametric item response theory for multilevel test data

Koopman, L.

Publication date
2022

Appendices

Appendix A Small Data Example

This document contains a small constructed dataset (Table A1) to demonstrate the computations of the bivariate and univariate frequencies, the weighted Guttman errors, and the scalability coefficients, after which the application of the recursive exp-log notation is illustrated.

Table A1:
Transposed Data Matrix for $S = 3$ Subjects, with $R_s = 4, 5,$ and 6 Raters, Respectively, Who Respond to $I = 2$ Items with $m + 1 = 3$ Response Categories

	Subject	1				2					3					
	Rater	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Item	X_a	0	0	1	0	1	1	1	2	1	0	1	2	0	0	0
	X_b	0	0	1	0	2	2	2	1	2	1	2	2	1	1	0

A.1 Computing Proportions

Table A2 provides the bivariate within-rater proportions (Equation 3.2) for items X_a and X_b (Table A1). Value 3 in the cell belonging to subject 1 and item-score pattern $(0, 0)$ is obtained by counting the number of raters who scored value 0 on both X_a and X_b (rater 1, 2, and 4). The within-rater bivariate proportion is then computed as $p_{ab}^{00(W)} = [3/4 + 0/5 + 1/6]/3 = .306$ (Table A2, fourth row).

Table A3 shows the bivariate between-rater proportions (Equations 3.3). The number 6 in the cell belonging to subject 1 and item-score pattern $(0, 0)$ is obtained by counting the pairs of raters for which one rater scores 0 on X_a and the other rater scores 0 on X_b. These rater-pairs are $(1, 2), (1, 4), (2, 1), (2, 4), (4, 1)$, and $(4, 2)$. The between-rater bivariate proportion is then computed by $p_{ab}^{00(B)} = [6/12 + 0/20 + 3/30]/3 = .200$ (Table A3, fourth row).

Table A4 shows the univariate proportions (Equations 3.4). Value 3 in the cell belonging to subject 1 and item-score 0 on X_a is obtained by counting the number of raters scoring 0 on X_a (rater 1, 2, and 4). The univariate proportions are computed by

Table A2:
Bivariate Within-Rater Proportions, Expected Proportions, and Weights of Guttman Errors of Items X_a and X_b in Table A1

s	Item-score pattern $(X_a = x, X_b = y)$									R_s
	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	
1	3				1					4
2						4		1		5
3	1	3				1			1	6
$p_{ab}^{xy}(W)$	.306	.167	.000	.000	.083	.322	.000	.067	.056	
$p_{ab}^{xy}(E)$	.144	.150	.178	.124	.128	.153	.037	.039	.046	
$\widehat{w}_{ab}^{xy}$	0	0	1	1	0	0	3	1	0	

Note. Frequencies of unobserved item-score patterns are left blank.

Table A3:
Between-Rater Bivariate Proportions of Items X_a and X_b in Table A1

s	Item-score pattern $(X_a = x, X_b = y)$									$R_s(R_s - 1)$
	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	
1	6	3		3						12
2					4	12			4	20
3	3	9	8	1	3	1	1	3	1	30
$p_{ab}^{xy}(B)$	.200	.183	.089	.094	.100	.211	.011	.033	.078	

Note. Frequencies of unobserved item-score patterns are left blank.

summing the relative frequencies across subjects, $p_a^0 = [3/4 + 0/5 + 4/6]\,/3 = .472$ (Table A4, row 4). The expected bivariate proportions under marginal independence of the items are computed by multiplying the univariate proportions (Equation 3.5). For item-score pattern $(0, 0)$ this results in $p_{ab}^{00(E)} = .472 \times .306 = .144$ (Table A2, fifth row).

A.2 Estimating weighted Guttman errors

The item-step ordering (Section 3.4.2) is determined by the estimated item-step popularities (final row in Table A4), resulting in the following estimated item-step ordering:

$$Z_{b1}, Z_{a1}, Z_{b2}, Z_{a2} = Z_{(1)}, Z_{(2)}, Z_{(3)}Z_{(4)}. \tag{A1}$$

For each bivariate item-score pattern the weights are estimated by applying Equation 3.7. For item-score pattern (2, 0) the second and fourth item-step was passed, whereas the

Table A4:
Univariate Proportions and Estimated Popularities of Items X_a and X_b in Table A1

s	X_a			X_b			R_s
	$x=0$	$x=1$	$x=2$	$y=0$	$y=1$	$y=2$	
1	3	1		3	1		4
2		4	1		1	4	5
3	4	1	1	1	3	2	6
p_i^x	.472	.406	.122	.306	.317	.378	
$\widehat{P}(X_{sri} \geq x)$	1.000	.528	.122	1.000	.694	.378	

Note. Frequencies of unobserved item-score patterns are left blank.

first and third item-step was failed, thus vector $\mathbf{z}_{ab}^{20} = [0\ 1\ 0\ 1]$. The estimated weight is

$$\widehat{w}_{ab}^{20} = 1 \times [(1-0)] + 0 \times [(1-0)+(1-1)] + 1 \times [(1-0)+(1-1)+(1-0)] = 3. \quad \text{(A2)}$$

Table A2 (last row) shows the estimated weights for all item-score patterns. Because the weights are based on the marginal frequencies, they are identical for within- and between-rater coefficients.

A.3 Estimating two-level Scalability Coefficients

Scalability coefficients H_{ab}^W and H_{ab}^B are estimated by taking the ratio of the weighted sum of observed Guttman errors (F_{ij}^W and F_{ij}^B, respectively) and the weighted sum of expected Guttman errors under marginal independence of the items (F_{ij}^E) (Equation 3.8 and 3.9, respectively). Values F_{ij} are computed using the bivariate proportions and weights in Table A2 and A3: $\widehat{F}_{ab}^W = \sum\sum \widehat{w}_{ab}^{xy} p_{ab}^{xy(W)} = 0 \times .306 + 0 \times .167 + 1 \times .000 + 1 \times .000 + 0 \times .083 + 0 \times .322 + 3 \times .000 + 1 \times .067 + 0 \times .056 = 0.067$, and similarly $\widehat{F}_{ab}^B = \sum\sum \widehat{w}_{ab}^{xy} p_{ab}^{xy(B)} = 0.250$ and $\widehat{F}_{ab}^E = \sum\sum \widehat{w}_{ab}^{xy} p_{ab}^{xy(E)} = 0.453$. Inserting these values in Equation 3.8 and 3.9 results in the estimated coefficients

$$\widehat{H}_{ab}^W = 1 - 0.067/0.453 = 0.853, \quad \text{(A3)}$$

and

$$\widehat{H}_{ab}^B = 1 - 0.250/0.453 = 0.448. \quad \text{(A4)}$$

The ratio coefficient yields $\widehat{H}_{ab}^{BW} = 0.448/0.853 = 0.526$. Because we only have two items, $H_{ij} = H_i = H$ for all coefficients.

A.4 Vector n with frequencies of item-score patterns

Vector **n** contains the frequencies of all item-score patterns in lexicographical order (Equation A5). For Table A1, this results in

$$\mathbf{n} = \begin{pmatrix} n_{ab}^{00} \\ n_{ab}^{01} \\ n_{ab}^{02} \\ n_{ab}^{10} \\ n_{ab}^{11} \\ n_{ab}^{12} \\ n_{ab}^{20} \\ n_{ab}^{21} \\ n_{ab}^{22} \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 0 \\ 0 \\ 1 \\ 5 \\ 0 \\ 1 \\ 1 \end{pmatrix}. \tag{A5}$$

Vector **n** can be decomposed in three vectors, one for each subject,

$$\mathbf{n}_1 = \begin{pmatrix} n_{sab}^{100} \\ n_{sab}^{101} \\ n_{sab}^{102} \\ n_{sab}^{110} \\ n_{sab}^{111} \\ n_{sab}^{112} \\ n_{sab}^{120} \\ n_{sab}^{121} \\ n_{sab}^{122} \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{n}_2 = \begin{pmatrix} n_{sab}^{200} \\ n_{sab}^{201} \\ n_{sab}^{202} \\ n_{sab}^{210} \\ n_{sab}^{211} \\ n_{sab}^{212} \\ n_{sab}^{220} \\ n_{sab}^{221} \\ n_{sab}^{222} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \text{ and } \mathbf{n}_3 = \begin{pmatrix} n_{sab}^{300} \\ n_{sab}^{301} \\ n_{sab}^{302} \\ n_{sab}^{310} \\ n_{sab}^{311} \\ n_{sab}^{312} \\ n_{sab}^{320} \\ n_{sab}^{321} \\ n_{sab}^{322} \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{A6}$$

Because we only have two items, the values in $\mathbf{n}_1, \mathbf{n}_2$, and $\mathbf{n}_3$ are equal to the bivariate frequencies per subject in Table A2.

A.5 Design matrix $\mathbf{A}_1$

Design matrix $\mathbf{A}_1$ (Equation 3.27) is constructed according to the L item-score patterns of vector **n**, and is presented in Table A5. An example of the other design matrices are available upon request from the corresponding author.

Rows 1 to 9 in matrix $\mathbf{A}_1$ (Table A5) reflect submatrix $\mathbf{B}^B$, and pertain to the between-rater bivariate proportions $P(X_{sri} = x, X_{spj} = y)$. To compute the cell values in $\mathbf{B}^B$ we need, for each subject, a vector containing the between-rater proportions of the item-score patterns $\mathbf{p}_s^B = \mathbf{n}_s/(SR_s(R_s - 1))$. Note that Equation A6 provides vectors $\mathbf{n}_s$. Hence,

$$\mathbf{p}_1^B = \begin{pmatrix} .083 \\ .000 \\ .000 \\ .000 \\ .028 \\ .000 \\ .000 \\ .000 \\ .000 \end{pmatrix}, \mathbf{p}_2^B = \begin{pmatrix} .000 \\ .000 \\ .000 \\ .000 \\ .000 \\ .067 \\ .000 \\ .017 \\ .000 \end{pmatrix}, \text{ and } \mathbf{p}_3^B = \begin{pmatrix} .011 \\ .033 \\ .000 \\ .000 \\ .000 \\ .011 \\ .000 \\ .000 \\ .011 \end{pmatrix}. \tag{A7}$$

Table A5:

Values of Matrix $\boldsymbol{A}_1$ *for Items* X_a *and* X_b *in Table A1.*

$\mathbf{A}_1$		Item-score pattern l of vector $\mathbf{n}$								
		1	2	3	4	5	6	7	8	9
	(x, y)									
	(0,0)	.042	.011							
	(0,1)	.029	.022							
	(0,2)	.006	.022							
	(1,0)					.083	.002			
$\mathbf{B}^B$	(1,1)						.020			
	(1,2)						.042			
	(2,0)									.011
	(2,1)									.033
	(2,2)								.067	.011
	(x, y)									
	(0,0)	.076								
	(0,1)		.056							
	(0,2)									
	(1,0)									
	(1,1)					.083				
$\mathbf{B}^W$	(1,2)						.064			
	(2,0)									
	(2,1)								.067	
	(2,2)									.056
	(x)									
	(0)	.076	.056							
	(1)					.083	.064			
$\mathbf{U}$	(2)								.067	.056
	(0)	.076								
	(1)		.056			.083			.067	
	(2)						.064			.056

Note. Zero values are left blank.

The first cell of $\mathbf{B}^B$ links the first item-score pattern (n_{ab}^{00}) to the first between-rater bivariate proportion $P(X_{sra} = 0, X_{spb} = 0)$. The value of this cell is computed by $\mathbf{1}(X_{a(1)} = 0)[\sum_s (n_{sb}^0 - \mathbf{1}(X_{b(1)} = 0))p_{s(1)}^B]/n_{(1)} = 1[(3-1)0.083 + (0-1)0 + (1-1)0.028]$ / 4 = .042. The second cell of the first row links the second item-score pattern (n_{ab}^{01}) to the first between-rater bivariate proportion, and is computed as $\mathbf{1}(X_{a(2)} = 0)[\sum_s (n_{sb}^0 - \mathbf{1}(X_{b(2)} = 0))p_{s(2)}^B]/n_{(2)} = 1[(0-0)0 + (0-0)0 + (1-0)0.033]$ / 3 = .011.

Rows 10 to 18 in matrix $\mathbf{A}_1$ (Table A5) belong to submatrix $\mathbf{B}^W$, and pertain to the within-rater bivariate proportions $P(X_{sri} = x, X_{srj} = y)$. For submatrix $\mathbf{B}^W$ we need the

vector of proportions of the item-score patterns,

$$\mathbf{p} = \sum_{s=1}^{S} \mathbf{n}_s/(SR_s) = \begin{pmatrix} .306 \\ .167 \\ .000 \\ .000 \\ .083 \\ .322 \\ .000 \\ .067 \\ .056 \end{pmatrix}. \tag{A8}$$

The first cell of $\mathbf{B}^W$ links the first item-score pattern to the within-rater bivariate proportion $P(X_{sra} = 0, X_{srb} = 0)$. The value of this cell is computed by $\mathbf{1}(X_{a(1)} = 0, X_{b(1)} = 0)p_{(1)}/n_{(1)} = 1 \times 0.306 \ / \ 4 = .076$.

Rows 19 to 24 in matrix $\mathbf{A}_1$ (Table A5) represent submatrix $\mathbf{U}$ and pertain to the univariate proportions $P(X_{sri} = x)$. The first cell of $\mathbf{U}$ links the first item-score pattern to the univariate proportion $P(X_{sra} = 0)$ and is computed by $\mathbf{1}(X_{a(1)} = 0)p_{(1)}/n_{(1)} = 1 \times .306 \ / \ 4 = .076$.

Multiplying $\mathbf{A}_1$ with vector $\mathbf{n}$ (Equation 3.28) results in a vector with all bivariate and univariate proportions, which are shown in Tabel A2 to A4. For the first row of $\mathbf{A}_1$ this results in $0.042 \times 4 + .011 \times 3 + 0 \times 0 + 0 \times 0 + 0 \times 1 + 0 \times 5 + 0 \times 0 + 0 \times 1 + 0 \times 1 = .200$, which is the value of $p_{ab}^{00(B)}$, presented in the first cell of the last row in Table A3.

A.6 Standard errors

For completeness sake we provide the computed standard errors (Section 3.5.2) of the estimated scalability coefficient. For $\widehat{H}_{ij}^{W} = 0.853$, $\widehat{SE} = 0.175$, for $\widehat{H}_{ij}^{B} = 0.448$, $\widehat{SE} = 0.314$, and for $\widehat{H}_{ij}^{BW} = 0.526$, $\widehat{SE} = 0.421$.

Appendix B Data simulation method

This document contains a description of the method to generate data that are used for simulation study in Section 3.5.4. Data were simulated for S subjects, each rated by $R = 18$ raters, on $I = 7$ items with $m + 1 = 5$ answer categories.

B.1 Hierarchical rater model

We used a parametric hierarchical rater model (see Patz et al., 2002, for details) to generate data, combining a graded response model for the unobserved subject scores on level 2 (Samejima, 1969) with a signal detection model for the observed item scores on level 1. Model parameters were selected to result in similar values of the scalability coefficients as in the real-data example in Section 3.2. The model is parameterized as follows:

$$\begin{array}{lrcl} \text{Level 2} & \theta_s & \sim & i.i.d.\ N(0, \sigma_\theta^2), s = 1, .., S \\ & \xi_{si} & \sim & \text{Graded response model}, i = 1, ..., 7, \text{ for each } s \\ \text{Level 1} & \delta_{sr} & \sim & i.i.d.\ N(0, \sigma_\delta^2), r = 1, ..., 18, \text{ for each } s \\ & X_{sri} & \sim & \text{Signal detection model, for each } s, r, i \end{array} \tag{B9}$$

Level 2: Graded response model. Attribute value θ_s for subject s is sampled from a normal distribution with mean 0 and standard deviation 1.25. Ideal rating ξ_{si} for subject s and item i was obtained using a graded response model (Samejima, 1969), which is the parametric version of Mokken's model for monotone homogeneity (Hemker et al., 1996). We fixed item discrimination α to 0.75 and subsequent item steps to have an equidistant item difficulty β_{ix} between -3 and 3. According to the graded response model, the probability that $\xi_{si} \geq x$ $(x = 1, \ldots, 4)$ given θ_s is

$$P(\xi_{si} \geq x|\theta_s) = \frac{\exp\left[\alpha(\theta_s - \beta_{ix})\right]}{1 + \exp\left[\alpha(\theta_s - \beta_{ix})\right]}. \tag{B10}$$

Note that $P(\xi_{si} \geq 0|\theta_s) = 1$ by definition. Probabilities $P(\xi_{si} = x|\theta_s) = P(\xi_{si} \geq x|\theta_s) - P(\xi_{si} \geq x+1|\theta_s)$ for $x = 0, \ldots, 4$ were used to sample the ideal ratings from a multinomial distribution.

Level 1: Signal detection model. Rater severity δ_{sr} was sampled from a normal distribution with mean 0 and standard deviation 0.75. The probability of observed score $X_{sri} = x$ given severity δ_{sr} and ideal rating ξ_{si}, $P(X_{sri} = x|\xi_{si}, \delta_{sr})$ $(x = 0, \ldots, 4)$, was obtained from a discrete signal detection model. The probabilities are normally distributed in x with mean $\xi_{si} + \delta_{sr}$ and standard deviation $\tau = 0.5$,

$$P(X_{sri} = x|\xi_{si}, \delta_{sr}) \propto \exp\left\{-\frac{\left[x - (\xi_{si} + \delta_{sr})\right]^2}{2\tau^2}\right\}. \tag{B11}$$

Probabilities $P(X_{sri} = x|\xi_{si}, \delta_{sr})$ for $x = 0, \ldots, 4$ were normalized to sum to 1, and used to sample the observed scores from a multinomial distribution.

Appendix C Computing $\mathbf{g}_3$ and $\mathbf{G}_3$ directly from the data

To reduce the computational burden when estimating standard errors of the scalability coefficients it is possible to compute $\mathbf{g}_3$ (Equation 3.33) and its Jacobian $\mathbf{G}_3$ directly from the data. Vector $\mathbf{g}_3$ contains the natural logarithm of the observed and expected weighted sum of Guttman errors ($\mathbf{F}_{ij}$) with a copy of the first element (F_{12}^B), and can be easily computed using Equations 3.8 and 3.9,

$$\mathbf{g}_3 = \log \begin{pmatrix} F_{12}^B \\ \mathbf{F}_{ij}^B \\ \mathbf{F}_{ij}^W \\ \mathbf{F}_{ij}^E \end{pmatrix}. \tag{C12}$$

The Jacobian then is a $(3K+1) \times L^*$ matrix

$$\mathbf{G}_3 = \frac{\partial \mathbf{g}_3}{\partial \mathbf{n}^T} = \begin{pmatrix} \mathbf{c}^T \\ \frac{\partial \mathbf{F}_{ij}^B}{\partial \mathbf{n}^T} \\ \frac{\partial \mathbf{F}_{ij}^W}{\partial \mathbf{n}^T} \\ \frac{\partial \mathbf{F}_{ij}^E}{\partial \mathbf{n}^T} \end{pmatrix}, \tag{C13}$$

where $\mathbf{c}$ is a vector that equals the first row of $\frac{\partial \mathbf{F}_{ij}^B}{\partial \mathbf{n}^T}$. Writing the numerator of the last term of Equation 3.9 in matrix notation, it follows that

$$\mathbf{F}_{ij}^B = \mathbf{W} \exp(\mathbf{I}_{(B)} \log(\mathbf{B}^B \mathbf{n})); \tag{C14}$$

writing the numerator of the last term of Equation 3.8 in matrix notation, it follows that

$$\mathbf{F}_{ij}^W = \mathbf{W} \exp(\mathbf{I}_{(B)} \log(\mathbf{B}^W \mathbf{n})); \tag{C15}$$

and writing the denominator of the last term of Equation 3.8 in matrix notation, it follows that

$$\mathbf{F}_{ij}^E = \mathbf{W} \exp(\mathbf{P} \log(\mathbf{U}\ \mathbf{n})). \tag{C16}$$

The result in Equations C14, C15, and C16 can be used to compute $\mathbf{g}_3$. Applying Equation 3.17 to $\mathbf{F}_{ij}^B$ (Equation C14), $\mathbf{F}_{ij}^W$ (Equation C15), and $\mathbf{F}_{ij}^E$ (Equation C16) provides three $K \times L^*$ matrices, for which the rows pertain to item-pairs $1, \ldots, K$ and the columns to item-score pattern $1, \ldots, L^*$. For $\mathbf{F}_{ij}^B$, the partial derivative then equals

$$\frac{\partial \mathbf{F}_{ij}^B}{\partial \mathbf{n}^T} = \text{Diag}(\mathbf{F}_{ij}^B)^{-1} \mathbf{W}\ \mathbf{B}^B, \tag{C17}$$

where the resulting element (k, l) equals the dot product of the k-th row of $\mathbf{W}$ and the l-th column of $\mathbf{B}^B$, divided by the k-th element of $\mathbf{F}_{ij}^B$. For $\mathbf{F}_{ij}^W$, the partial derivative

equals

$$\frac{\partial \mathbf{F}_{ij}^{W}}{\partial \mathbf{n}^{T}} = \text{Diag}(\mathbf{F}_{ij}^{W})^{-1}\mathbf{W}\ \mathbf{B}^{W}, \tag{C18}$$

where the resulting element (k, l) equals the dot product of the k-th row of $\mathbf{W}$ and the l-th column of $\mathbf{B}^{W}$, divided by the k-th element of $\mathbf{F}_{ij}^{W}$. For $\mathbf{F}_{ij}^{E}$, the partial derivative equals

$$\frac{\partial \mathbf{F}_{ij}^{E}}{\partial \mathbf{n}^{T}} = \text{Diag}(\mathbf{F}_{ij}^{E})^{-1}\mathbf{W}\ \text{Diag}(\exp(\mathbf{P}\ \log(\mathbf{p}_i)))\mathbf{P}\ \text{Diag}(\mathbf{p}_i)^{-1}\ \mathbf{U}, \tag{C19}$$

where the resulting element (k, l) equals $\sum_x \sum_y w_{ij}^{xy}(u_{i(l)}^{x} p_j^{y} + u_{j(l)}^{y} p_i^{x})/F_{ij}^{E}$. The result in Equations C17, C18, and C19 can be used to compute $\mathbf{G}_3$ (Equation C13). Jacobians $\mathbf{G}_3^{\dagger}$ and $\mathbf{G}_3^{\ddagger}$ may be obtained using $\mathbf{G}_3$. Let $\mathbf{D}^{T}$ be an $I \times K$ matrix for which the rows pertain to items $1, \ldots, I$, and the columns pertain to item-pairs $1, \ldots, K$. Element (i, k) of $\mathbf{D}^{\dagger}$ equals 1 if item i is in item-pair k, and 0 otherwise. It follows that,

$$\mathbf{G}_3^{\dagger} = \frac{\partial \mathbf{g}_3^{\dagger}}{\partial \mathbf{n}^{T}} = \begin{pmatrix} \mathbf{c}^{\dagger} \\ \mathbf{D}^{\dagger}\frac{\partial \mathbf{F}_{ij}^{B}}{\partial \mathbf{n}^{T}} \\ \mathbf{D}^{\dagger}\frac{\partial \mathbf{F}_{ij}^{W}}{\partial \mathbf{n}^{T}} \\ \mathbf{D}^{\dagger}\frac{\partial \mathbf{F}_{ij}^{E}}{\partial \mathbf{n}^{T}} \end{pmatrix}, \tag{C20}$$

and

$$\mathbf{G}_3^{\ddagger} = \frac{\partial \mathbf{g}_3^{\ddagger}}{\partial \mathbf{n}^{T}} = \begin{pmatrix} \mathbf{c}^{\ddagger} \\ \mathbf{1}_{(K)}^{T}\frac{\partial \mathbf{F}_{ij}^{B}}{\partial \mathbf{n}^{T}} \\ \mathbf{1}_{(K)}^{T}\frac{\partial \mathbf{F}_{ij}^{W}}{\partial \mathbf{n}^{T}} \\ \mathbf{1}_{(K)}^{T}\frac{\partial \mathbf{F}_{ij}^{E}}{\partial \mathbf{n}^{T}} \end{pmatrix}, \tag{C21}$$

where $\mathbf{c}^{\dagger}$ is a copy of the first row of $\mathbf{D}^{\dagger}\frac{\partial \mathbf{F}_{ij}^{B}}{\partial \mathbf{n}^{T}}$, and $\mathbf{c}^{\ddagger}$ equals $\mathbf{1}_{K}^{T}\frac{\partial \mathbf{F}_{ij}^{B}}{\partial \mathbf{n}^{T}}$.

Appendix D Illustrative example

Table D6 shows two small constructed data examples, each with two subjects and five raters per subject on two three-category items. The same item scores are present in both data sets, but rater 4 of subject 1 and rater 5 of subject 2 are exchanged in the second data set.

Table D6:

Two Small Constructed Multi-Rater Data Examples, One with a Large Rater Effect and One with a Small Rater Effect. Both Data Sets Have Two Subjects (s), Each Rated by a Unique Set of Five Raters (r) on Two Three-Category Items (X_i and X_j).

	Data set 1: Large rater effect										Data set 2: Small rater effect									
	$s=1$					$s=2$					$s=1$					$s=2$				
r	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
X_i	2	2	2	1	1	0	0	1	1	2	2	2	2	2	1	0	0	1	1	1
X_j	1	2	2	0	1	0	1	0	1	1	1	2	2	1	1	0	1	0	1	0
	$\overline{X}_{1\cdot\cdot}=1.4$					$\overline{X}_{2\cdot\cdot}=0.7$					$\overline{X}_{2\cdot\cdot}=1.6$					$\overline{X}_{2\cdot\cdot}=0.5$				
	$H^W=.762$			$H^B=.167$		$H^{BW}=.219$					$H^W=.762$			$H^B=.702$		$H^{BW}=.922$				
CI	$(0.343, 1.181)$			$(-0.231, 0.565)$		$(-0.288, 0.726)$					$(0.349, 1.175)$			$(0.435, 0.970)$		$(0.441, 1.402)$				

Note. CI is the 95% Wald-based confidence interval.

For both data sets in Table D6, the item-step ordering is $Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2}$. Therefore, consistent item-score patterns are $(0,0)$, $(1,0)$, $(1,1)$, $(2,1)$, and $(2,2)$, whereas patterns $(0,1)$, $(0,2)$, $(1,2)$, and $(2,0)$ are Guttman errors. Within raters, Guttman error $(0,1)$ occurs once in each data set (rater 2 of subject 2). In the first data set, there are five between-rater Guttman errors $(0,1)$ (for subject 2, rater 1 scored 0 on X_i, whereas rater 2, 4, and 5 scored 1 on X_j, and rater 2 scored 0 on X_i, whereas rater 4 and 5 scored 1 on X_j), four between-rater Guttman errors $(1,2)$, and five between-rater Guttman errors $(2,0)$, summing up to 14 between-rater Guttman errors. In the second data set there are only three $(0,1)$ and two $(1,2)$ between-rater Guttman errors, summing up to five.

Because there are relatively many between-rater Guttman errors in the first data-set, there is little consistency between raters of the same subject and H^B is low compared to H^W, as is reflected in ratio $H^{BW} = .219$. Although scalability coefficients H^W and H^B exceed the criteria presented by Snijders (2001a), the ratio coefficient is below .3 and the 95% confidence interval of H^B and H^{BW} includes zero. This indicates that the item responses are mainly determined by the raters and it is doubtful whether it makes sense to scale subjects on θ using the test score on this set of items. In the second data set there is almost as much consistency between raters as there is within raters, reflected by a ratio coefficient of $H^{BW} = .922$. All coefficients are above the criteria of Snijders and

the confidence intervals exceed zero. This indicates that the item responses are mainly determined by the subject, and subjects can be scaled on θ using these items.

The data example demonstrates that high values for two-level coefficients do not require perfect agreement among raters of the same subject. For H^{BW} to be high it is of importance that the probability of a between-rater Guttman error pattern is close to the probability of a within-rater Guttman error pattern.

Appendix E Lemmas to Chapter 8

Lemma E1. *UN-1 implies UN-2.*

Proof. If θ is unidimensional, its expectation $E(\theta)$ is also unidimensional. Variable θ is unidimensional by UN-1. As group-level variable $\gamma = E_\delta(\theta)$ (Equation 8.1), variable γ is also unidimensional. □

Lemma E2. *LI-1 implies LI-2.*

Proof. The conditional probability of item-score pattern $\mathbf{X}_s = \mathbf{x}_s$ given γ is

$$P(\mathbf{X}_s = \mathbf{x}_s|\gamma) \quad = \quad P(\textstyle\sum_{r=1}^{R} \mathbf{X}_{sr} = \mathbf{x}_s|\gamma) \tag{E22}$$

By the law of total expectation, Equation E22 equals

$$= \quad E_\delta[P(\textstyle\sum_{r=1}^{R} \mathbf{X}_{sr} = \mathbf{x}_s|\gamma, \delta)] \tag{E23}$$

By LI-1, individual item scores are independent given θ. Hence, Equation E23 equals

$$= \quad E_\delta[\textstyle\prod_{i=1}^{I} P(\sum_{r=1}^{R} X_{sri} = x_{si}|\gamma, \delta)] \tag{E24}$$

For independent variables: The expectation of their product is the product of their expectations (e.g., Rice, 2006, p. 124, Corollary A). Hence, Equation E24 equals

$$\begin{array}{lll} & = & \prod_{i=1}^{I} E_\delta[P(\sum_{r=1}^{R} X_{sri} = x_{si}|\gamma, \delta)] \\ & = & \prod_{i=1}^{I} P(\sum_{r=1}^{R} X_{sri} = x_{si}|\gamma) \\ & = & \prod_{i=1}^{I} P(X_{si} = x_{si}|\gamma) \end{array} \tag{E25}$$

The last term in Equation E25 equals the definition of LI-2. □

Lemma E3. *MO-1 implies MO-2.*

Proof. Let $P(\delta)$ denote the probability density function of the distribution of δ. The group-level item-response function is,

$$\begin{array}{rcl} E_i(\gamma) & = & E_\delta[E_i(\theta)] \text{ (Equation 8.6)} \\ & = & \int E_i(\theta) P(\delta|\gamma) d\delta \end{array} \tag{E26}$$

As δ and γ are independent, $P(\delta|\gamma) = P(\delta)$, and the last term of Equation E26 reduces to

$$\int E_i(\theta) P(\delta) d\delta \tag{E27}$$

By MO-1, $E_i(\theta)$ is nondecreasing in θ. Hence, Equation E26 is nondecreasing in γ, which equals the definition of MO-2. □

Lemma E4. *IIO-1 implies IIO-2.*

Proof. By IIO-1

$$\begin{array}{lll} E_i(\theta) & \leq & E_j(\theta) \\ \Leftrightarrow \int E_i(\theta)P(\delta|\gamma)d\delta & \leq & \int E_j(\theta)P(\delta|\gamma)d\delta \\ \Leftrightarrow E_\delta[E_i(\theta)] & \leq & E_\delta[E_j(\theta)] \\ \Leftrightarrow E_i(\gamma) & \leq & E_j(\gamma) \end{array} \tag{E28}$$

by 8.6. The final result in Equation E28 equals the definition of IIO-2. □

Lemma E5. *MHM-1 implies that $E(g(\mathbf{X}_{sr})|\theta)$ is nondecreasing in θ for any bounded, nondecreasing function $g(\cdot)$.*

Proof. By LI-1, scores X_{sri} within $\mathbf{X}_{sr}$ are independent given θ. By MO-1, X_{sri} is stochastically ordered in θ; that is, for $\theta_a < \theta_b$, $E_i(\theta_a) \leq E_i(\theta_b)$ for all i. For a set of independent variables the stochastic ordering is preserved under convolutions, for any bounded, nondecreasing function $g(\cdot)$ (Shaked & Shanthikumar e.g., 2007, Theorem 1.A.3(b); see also Ahmed et al. 1981, Lemma 3.3; Holland & Rosenbaum 1986, Lemma 2). Hence

$$E[g(\mathbf{X}_{sr})|\theta_a] \leq E[g(\mathbf{X}_{sr})|\theta_b]. \tag{E29}$$

□

Lemma E6. *MHM-2 implies $E(g(\mathbf{X}_s)|\gamma)$ is nondecreasing for any bounded, nondecreasing function $g(\cdot)$ and all γ.*

Proof. The proof mimics the proof of Lemma E5. By LI-2, scores X_{si} within $\mathbf{X}_s$ are independent given γ. By MO-2, X_{si} is stochastically ordered in γ; that is, for $\gamma_a < \gamma_b$, $E_i(\gamma_a) \leq E_i(\gamma_b)$ for all i. For a set of independent variables the stochastic ordering is preserved under convolutions for any bounded, nondecreasing function $g(\cdot)$. Hence

$$E[g(\mathbf{X}_s)|\gamma_a] \leq E[g(\mathbf{X}_s)|\gamma_b]. \tag{E30}$$

□

References

Achenbach, T. M., Becker, A., Döpfner, M., Heiervang, E., Roessner, V., Steinhausen, H. C., & Rothenberger, A. (2008). Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: Research findings, applications, and future directions. *Journal of Child Psychology and Psychiatry*, *49*(3), 251–275. http://doi.org/10.1111/j.1469-7610.2007.01867.x

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Wiley.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. http://doi.org/10.1080/00031305.1998.10480550

Ahmadi, K., Reidpath, D. D., Allotey, P., & Hassali, M. A. A. (2016). A latent trait approach to measuring HIV/AIDS related stigma in healthcare professionals: Application of Mokken scaling technique. *BMC Medical Education*, *16*(1), 155–164. http://doi.org/10.1186/s12909-016-0676-3

Ahmed, A.-H. N., Leon, R., & Proschan, F. (1981). Generalized association, with applications in multivariate statistics. *The Annals of Statistics*, *9*(1), 168–176. http://doi.org/10.1214/aos/1176345343

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. http://doi.org/10.1007/BF02293814

Atkinson, M., Zibin, S., & Chuang, H. (1997). Characterizing quality of life among patients with chronic mental illness: A critical examination of the self-report methodology. *American Journal of Psychiatry*, *154*(1), 99–105. http://doi.org/10.1176/ajp.154.1.99

Banas, K., Lyimo, R. A., Hospers, H. J., Van der Ven, A., & De Bruin, M. (2017). Predicting adherence to combination antiretroviral therapy for HIV in Tanzania: A test of an extended theory of planned behaviour model. *Psychology & Health*, *32*(10), 1249–1265. http://doi.org/10.1080/08870446.2017.1283037

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). Wiley.

Bergsma, W. P., Croon, M. A., & Hagenaars, J. A. (2009). *Marginal models for dependent, clustered, and longitudinal categorical data*. Springer.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick. *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

Boor, K., Scheele, F., Van der Vleuten, C. P. M., Scherpbier, A. J. J. A., Teunissen, P. W., & Sijtsma, K. (2007). Psychometric properties of an instrument to measure the clinical learning environment. *Medical Education*, *41*(1), 92–99. http://doi.org/10.1111/j.1365-2929.2006.02651.x

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061–1071. http://doi.org/10.1037/0033-295X.111.4.1061

Brusco, M. J., Köhn, H.-F., & Steinley, D. (2015). An exact method for partitioning dichotomous items within the framework of the monotone homogeneity model. *Psychometrika*, *80*(4), 949–967. http://doi.org/10.1007/s11336-015-9459-8

Bull, S. B., Darlington, G. A., Greenwood, C. M. T., & Shin, J. (2001). Design considerations for association studies of candidate genes in families. *Genetic Epidemiology*, *20*(2), 149–174. http://doi.org/10.1002/1098-2272(200102)20:2⟨149::AID-GEPI1⟩3.0.CO;2-A

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. http://doi.org/10.18637/jss.v048.i06

Chang, H.-H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika*, *59*(3), 391–404. http://doi.org/10.1007/BF02296132

Chen, S.-K., Hwang, F.-M., & Lin, S. S. (2013). Satisfaction ratings of QOLPAV: Psychometric properties based on the graded response model. *Social Indicators Research*, *110*(1), 367–383. http://doi.org/10.1007/s11205-011-9935-1

Chen, Y., Watson, R., & Hilton, A. (2016). An exploration of the structure of mentors' behavior in nursing education using exploratory factor analysis and Mokken scale analysis. *Nurse Education Today*, *40*(1), 161–167. http://doi.org/10.1016/j.nedt.2016.03.001

Cheng, G., Yu, Z., & Huang, J. Z. (2013). The cluster bootstrap consistency in generalized estimating equations. *Journal of Multivariate Analysis*, *115*(1), 33–47. http://doi.org/10.1016/j.jmva.2012.09.003

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). Wiley.

Chou, Y. H., Lee, C. P., Liu, C. Y., & Hung, C. I. (2017). Construct validity of the depression and somatic symptoms scale: Evaluation by Mokken scale analysis. *Neuropsychiatric Disease and Treatment*, *13*(1), 205–211. http://doi.org/10.2147/NDT.S118825

Conijn, J. M., Smits, N., & Hartman, E. E. (2020). Determining at what age children provide sound self-reports: An illustration of the validity-index approach. *Assessment*, *27*(7), 1604–1618. http://doi.org/10.1177/1073191119832655

Coromina, L., & Camprubí, R. (2016). Analysis of tourism information sources using a Mokken scale perspective. *Tourism Management*, *56*(1), 75–84. http://doi.org/10.1016/j.tourman.2016.03.025

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2020). On the practical consequences of misfit in Mokken scaling. *Applied Psychological Measurement*, *44*(6), 482–496. http://doi.org/10.1177/0146621620920925

Crişan, D. R., Van de Pol, J. E., & Van der Ark, L. A. (2016). Scalability coefficients for two-level polytomous item scores: An introduction and an application. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research: The 80th annual meeting of the Psychometric Society, Beijing, 2015* (pp. 139–153). Springer. http://doi.org/10.1007/978-3-319-38759-8_11

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333–356. http://doi.org/10.1111/j.1745-3984.2011.00143.x

Deen, M., & De Rooij, M. (2020). ClusterBootstrap: An R package for the analysis of hierarchical data using generalized linear models with the cluster bootstrap. *Behavior Research Methods*, *52*(2), 572–590. http://doi.org/10.3758/s13428-019-01252-y

De Rooij, M., & Worku, H. M. (2012). A warning concerning the estimation of multinomial logistic models with correlated responses in SAS. *Computer Methods and Programs in Biomedicine*, *107*(2), 341–346. http://doi.org/10.1016/j.cmpb.2012.01.008

Dorai-Raj, S. (2014). *binom: Binomial confidence intervals for several parameterizations. R-package, version 1.1-1.* [Computer software]. Retrieved from https://CRAN.R-project.org/package=binom/

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*(1), 149–167. http://doi.org/10.1016/j.leaqua.2004.09.009

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (1st ed.). Chapman & Hall.

Elley, C. R., Kerse, N., Chondros, P., & Robinson, E. (2005). Intraclass correlation coefficients from three cluster randomised controlled trials in primary and residential health care. *Australian and New Zealand Journal of Public Health*, *29*(5), 461–467. http://doi.org/10.1111/j.1467-842X.2005.tb00227.x

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika*, *79*(2), 303–316. http://doi.org/10.1007/s11336-013-9341-5

Elsman, E. B. M., Van Nispen, R. M. A., & Van Rens, G. H. M. B. (2020). Psychometric evaluation of the Participation and Activity Inventory for Children and Youth (PAI-CY) 0–2 years with visual impairment. *Quality of Life Research*, *29*(3), 775–781. http://doi.org/10.1007/s11136-019-02343-1

Embretson, S. E., & Reise, P. E. (2000). *Item response theory for psychologists*. Erlbaum.

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*(1), 105–120. http://doi.org/10.1037/1082-989X.12.1.105

Esary, J. D., Proschan, F., & Walkup, D. W. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics*, *38*(5), 1466–1474. http://doi.org/10.1214/aoms/1177698701

Fan, W., Williams, C. M., & Corkin, D. M. (2011). A multilevel analysis of student perceptions of school climate: The effect of social and academic risk factors. *Psychology in the Schools*, *48*(6), 632–647. http://doi.org/10.1002/pits.20579

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. Academic Press.

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390. http://doi.org/10.1111/j.1467-9868.2007.00593.x

Fisher, K. J., & Li, F. (2004). A community-based walking trial to improve neighborhood quality of life in older adults: a multilevel analysis. *Annals of Behavioral Medicine*, *28*(3), 186–194. http://doi.org/10.1207/s15324796abm2803_7

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Fraser, B., McRobbie, C., & Fisher, D. (1996). Development, validation and use of personal and class forms of a new classroom environment questionnaire. *Proceedings*

Western Australian Institute for Educational Research Forum 1996. http://www.waier.org.au/forums/1996/fraser.html

Freedland, K. E., Lemos, M., Doyle, F., Steinmeyer, B. C., Csik, I., & Carney, R. M. (2017). The techniques for overcoming depression questionnaire: Mokken scale analysis, reliability, and concurrent validity in depressed cardiac patients. *Cognitive Therapy and Research*, *41*(1), 117–129. http://doi.org/10.1007/s10608-016-9797-6

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*(3), 383–392. http://doi.org/10.1007/BF02294219

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, *149*(9), 876–883. http://doi.org/10.1093/oxfordjournals.aje.a009904

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.

Harden, J. J. (2011). A bootstrap method for conducting statistical inference with clustered data. *State Politics & Policy Quarterly*, *11*(2), 223–246. http://doi.org/10.1177/1532440011406233

Hemker, B. T., Sijtsma, K., Molenaar, I., & Junker, B. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*(4), 679–693. http://doi.org/10.1007/BF02294042

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional itembank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*(4), 337–352. http://doi.org/10.1177/014662169501900404

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*(3), 331–347. http://doi.org/10.1007/BF02294555

Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*(4), 487–506. http://doi.org/10.1007/BF02296191

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*(4), 1523–1543. http://doi.org/10.1214/aos/1176350174

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Routledge.

Huber, A., Oldridge, N., Benzer, W., Saner, H., & Höfer, S. (2020). Validation of the German HeartQoL: A short health-related quality of life questionnaire for cardiac patients. *Quality of Life Research*, *29*(4), 1093–1105. http://doi.org/10.1007/s11136-019-02384-6

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent bernoulli random variables. *Psychometrika*, *59*(1), 77–79. http://doi.org/10.1007/BF02294266

Joe, H.-K., Hiver, P., & Al-Hoorie, A. H. (2017). Classroom social climate, self-determined motivation, willingness to communicate, and achievement: A study of structural relationships in instructed second language settings. *Learning and Individual Differences*, *53*, 133–144. http://doi.org/10.1016/j.lindif.2016.11.005

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, *21*(3), 1359–1378. http://doi.org/10.1214/aos/1176349262

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics*, *25*(3), 1327–1343. http://doi.org/10.1214/aos/1069362751

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*(1), 65–81. http://doi.org/10.1177/01466216000241004

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298. http://doi.org/10.1207/S15324818AME1604_2

Koopman, L., Zijlstra, B. J. H., De Rooij, M., & Van der Ark, L. A. (2020). Bias of two-level scalability coefficients and their standard errors. *Applied Psychological Measurement*, *44*(3), 197–214. http://doi.org/10.1177/0146621619843821

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2017). Weighted Guttman errors: Handling ties and two-level data. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 138–190). Springer. http://doi.org/10.1007/978-3-319-56294-0_17

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2020). Standard errors of two-level scalability coefficients. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 213–236. http://doi.org/10.1111/bmsp.12174

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2021). Range-preserving confidence intervals and significance tests for scalability coefficients in Mokken scale analysis. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 85st annual meeting of the Psychometric Society, virtual* (pp. 175–185). Springer. http://doi.org/10.1007/978-3-030-74772-5_16

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2022). A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. *Quality of Life Research*, *31*(1), 25–36. http://doi.org/10.1007/s11136-021-02840-2

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*(1), 42–69. http://doi.org/10.1177/0081175013481958

Kuijpers, R. E., Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2016). Bias in point estimates and standard errors of Mokken's scalability coefficients. *Applied Psychological Measurement*, *40*(5), 331–345. http://doi.org/10.1177/0146621616638500

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). Wiley.

Lewnard, J. A., Givon-Lavi, N., Huppert, A., Pettigrew, M. M., Regev-Yochay, G., Dagan, R., & Weinberger, D. M. (2015). Epidemiological markers for interactions among Streptococcus pneumoniae, Haemophilus influenzae, and Staphylococcus aureus in upper respiratory tract carriage. *The Journal of Infectious Diseases*, *213*(10), 1596–1605. http://doi.org/10.1093/infdis/jiv761

Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika*, *76*(2), 200–216. http://doi.org/10.1007/S11336-010-9199-8

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*(4), 578–595. http://doi.org/10.1177/0013164409355697

Luijten, M. A. J., Van Litsenburg, R. R. L., Terwee, C. B., Grootenhuis, M. A., & Haverman, L. (2021). Psychometric properties of the Patient-Reported Outcomes Measurement Information System (PROMIS®) pediatric item bank peer relationships in the Dutch general population. *Quality of Life Research*, *30*(7), 2061–2070. http://doi.org/10.1007/s11136-021-02781-w

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. http://doi.org/10.1027/1614-2241.1.3.86

Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics*, *32*(3), 287–314. http://doi.org/10.3102/1076998606298033

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. http://doi.org/10.1007/BF02296272

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, *26*(2), 169–194. http://doi.org/10.1080/09243453.2014.939198

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311–314. http://doi.org/10.1177/014662169401800402

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*(4), 341–350. http://doi.org/10.1016/S0895-4356(03)00007-6

Mok, M. M. C. (2002). Determinants of students' quality of school life: A path model. *Learning Environments Research*, *5*(3), 275–300. http://doi.org/10.1023/A:1021924322950

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

Molenaar, I. W. (1983). *Item steps (Heymans Bulletin 83-630-EX)*. University of Groningen, Groningen, The Netherlands.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, *12*(37), 97-117. https://www.vvsor.nl/wp-content/uploads/2020/06/KM1991037007.pdf

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

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for windows*. iec ProGAMMA.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11). http://doi.org/10.1002/sim.8086

Ng, E. S.-W., Grieve, R., & Carpenter, J. R. (2013). Two-stage non-parametric bootstrap sampling with shrinkage correction for clustered data. *The Stata Journal*, *13*(1), 141–164. https://www.stata-journal.com/sjpdf.html?articlenum=st0288

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384. http://doi.org/10.3102/10769986027004341

Peetsma, T. T. D., Wagenaar, E., & De Kat, E. (2001). School motivation, future time perspective and well-being of high school students in segregated and integrated schools in the Netherlands and the role of ethnic self-description. In J. Koppen, I. Lunt, & C. Wulf (Eds.), *Education in Europe. Cultures, values, institutions in transition* (pp. 54–74). Waxmann.

R Core Team. (2020). *R: A language and environment for statistical computing*. [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org/

Rampichini, C., Grilli, L., & Petrucci, A. (2004). Analysis of university course evaluations: From descriptive measures to multilevel models. *Statistical Methods and Applications*, *13*(3), 357–373. http://doi.org/10.1007/s10260-004-0087-1

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). Wiley.

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

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, *29*(1), 1–41. http://doi.org/10.1111/0081-1750.00059

Ravens-Sieberer, U., Herdman, M., Devine, J., Otto, C., Bullinger, M., Rose, M., & Klasen, F. (2014). The European KIDSCREEN approach to measure quality of life and well-being in children: Development, current application, and future advances. *Quality of Life Research*, *23*(3), 791–803. http://doi.org/10.1007/s11136-013-0428-3

Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*(2), 147–169. http://doi.org/10.1023/B:MOEM.0000032312.95499.6f

Reise, S. P., Meijer, R. R., Ainsworth, A. T., Morales, L. S., & Hays, R. D. (2006). Application of group-level item response models in the evaluation of costumer reports about health plan quality. *Multivariate Behavioral Research*, *41*(1), 85–102. http://doi.org/10.1207/s15327906mbr4101_6

Rice, J. A. (2006). *Mathematical statistics and data analysis* (3rd ed.). Thomson Brooks/Cole.

Rinkel, W. D., Aziz, M. H., Van Neck, J. W., Cabezas, M. C., Van der Ark, L. A., & Coert, J. H. (2019). Development of grading scales of pedal sensory loss using Mokken scale analysis on the Rotterdam Diabetic Foot Study Test Battery data. *Muscle & Nerve*, *60*(5), 520–527. http://doi.org/10.1002/mus.26628

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*(3), 425–435. http://doi.org/10.1007/BF02306030

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349–359. http://doi.org/10.1007/BF02294217

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometrika monograph supplement No. 17). Psychometric Society. http://www.psychometrika.org/journal/online/MN17.pdf

Sen, P. K., & Singer, J. M. (1993). *Large sample methods in statistics: An introduction with applications*. Chapman & Hall.

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. Springer.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Sherman, M., & Le Cessie, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation*, *26*(3), 901–925. http://doi.org/10.1080/03610919708813417

Sijtsma, K., Emons, W. H. M., Bouwmeester, S., Nyklíček, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the world health organization quality-of-life scale (WHOQOL-Bref). *Quality of Life Research*, *17*(2), 275–290. http://doi.org/10.1007/s11136-007-9281-6

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*(2), 183–200. http://doi.org/10.1007/BF02294774

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, *25*(4), 391–415. http://doi.org/10.3102/10769986025004391

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*(1), 79–105. http://doi.org/10.1111/j.2044-8317.1996.tb01076.x

Sijtsma, K., Meijer, R. R., & Van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*(1), 31–37. http://doi.org/10.1016/j.paid.2010.08.016

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

Sijtsma, K., & Molenaar, I. W. (2016). Mokken models. In W. J. Van der Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 303–321). CRC Press.

Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158. http://doi.org/10.1111/bmsp.12078

Sijtsma, K., & Van der Ark, L. A. (2020). *Measurement models for psychological attributes.* Chapman and Hall/CRC Press.

Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken Scale Analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement*, *36*(6), 516–539. http://doi.org/10.1177/0146621612451050

Smits, N., Paap, M., & Böhnke, J. R. (2018). Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Quality of Life Research*, *27*(4), 1055–1063. http://doi.org/10.1007/s11136-018-1821-8

Snijders, T. A. B. (2001a). Two-level non-parametric 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). Springer. http://doi.org/10.1007/978-1-4613-0169-1_17

Snijders, T. A. B. (2001b). *TWOMOK.* [Computer software]. Retrieved from https://www.stats.ox.ac.uk/~snijders/

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. http://doi.org/10.3102/1076998616646200

Stewart, J. (2008). *Calculus: Early transcendentals* (6th ed.). Thompson Brooks/Cole.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*(1), 75–99. http://doi.org/10.1007/s00357-013-9122-y

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement*, *74*(5), 809–822. http://doi.org/10.1177/0013164414529793

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology*, *12*(4), 117–123. http://doi.org/10.1027/1614-2241/a000115

Swiger, P. A., Raju, D., Breckenridge-Sproat, S., & Patrician, P. A. (2017). Adaptation of the Practice Environment Scale for military nurses: A psychometric analysis. *Journal of Advanced Nursing*, *73*(9), 2219–2236. http://doi.org/10.1111/jan.13276

Thoonen, E. E. J., Sleegers, P. J. C., Peetsma, T. T. D., & Oort, F. J. (2011). Can teachers motivate students to learn? *Educational Studies*, *37*(3), 345–360. http://doi.org/10.1080/03055698.2010.507008

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55. http://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 179–187. http://doi.org/10.1348/000711007X173391

Vágó, E., Kemény, S., & Láng, Z. (2011). Overdispersion at the binomial and multinomial distribution. *Periodica Polytechnica Chemical Engineering*, *55*(1), 17–20. http://doi.org/10.3311/pp.ch.2011-1.03

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*(1), 3–24. http://doi.org/10.1177/0146621603259277

Van Abswoude, A. A. H., Vermunt, J. K., Hemker, B. T., & Van der Ark, L. A. (2004). Mokken scale analysis using hierarchical clustering procedures. *Applied Psychological Measurement*, *28*(5), 332–354. http://doi.org/10.1177/0146621604265510

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*(3), 273–282. https://doi.org/10.1177/01466210122032073

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19. http://doi.org/10.18637/jss.v020.i11

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. http://doi.org/10.18637/jss.v048.i05

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*(2), 272–279. http://doi.org/10.1007/s11336-010-9147-7

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika*, *73*(2), 183–208. http://doi.org/10.1007/s11336-007-9034-z

Van der Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational research*, *49*(2), 127–152. http://doi.org/10.1080/00131880701369651

Van der Linden, W. J. (2016). *Handbook of item response theory. Volume 1: Models*. CRC Press.

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer.

Van der Veen, I., & Peetsma, T. (2009). The development in self-regulated learning behaviour of first-year students in the lowest level of secondary school in the Netherlands. *Learning and Individual differences*, *19*(1), 34–46. http://doi.org/10.1016/j.lindif.2008.03.001

Van Onna, M. J. H. (2004). Estimates of the sampling distribution of scalability coefficient H. *Applied Psychological Measurement*, *28*(6), 427–449. http://doi.org/10.1177/0146621604268735

Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139–163. https://doi.org/10.1093/pan/mpg002

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). Springer. http://doi.org/10.1007/978-1-4613-0169-1_5

Watt, G., McConnachie, A., Upton, M., Emslie, C., & Hunt, K. (2000). How accurately do adult sons and daughters report and perceive parental deaths from coronary disease? *Journal of Epidemiology & Community Health*, *54*(11), 859–863. http://doi.org/10.1136/jech.54.11.859

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306. http://doi.org/10.3102/10769986026003283

Wind, S. A. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice*, *36*(2), 50–66. http://doi.org/10.1111/emip.12153

Zijlmans, E. A. O., Tijmstra, J., van der Ark, L. A., & Sijtsma, K. (2018). Item-score reliability in empirical-data sets and its relationship with other item indices. *Educational and Psychological Measurement*, *78*(6), 998–1020. http://doi.org/10.1177/0013164417728358

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research*, *42*, 531–555. http://doi.org/10.1080/00273170701384340

Zijsling, D., Keuning, J., Keizer-Mittelhaëuser, M.-A., Naaijer, H., & Timmermans, A. (2017). *Cohortonderzoek COOL5-18: Technisch rapport meting VO-3 in 2014*. GION Onderwijs/Onderzoek. https://www.rug.nl/research/portal/files/41740853/Cool_afn1314_vo3_techrapport.pdf

Summary

Nonparametric Item Response Theory for Multilevel Test Data

Psychological measurement aims at measuring a psychological construct (e.g., an attribute, skill, or ability) by means of observable variables such as items on a test or questionnaire (e.g., Sijtsma & Van der Ark, 2020, pp. 5–6). To relate the item scores to the psychological construct, item response theory (IRT) models are used. A subset of these models are nonparametric IRT models, which fit test data relatively well compared to other IRT models (Mokken, 1971; Sijtsma & Molenaar, 2002). A fitting nonparametric IRT model implies that the sum score across items may be used to order respondents and, for some models, that the mean item score may be used to order items. IRT models and their features are only valid if they fit test data well (Sijtsma & Van der Ark, 2020, p. 3): Validating a test or questionnaire using IRT models is therefore a vital step before using it for measurement (see, also, Borsboom et al., 2004). Various methods exist to evaluate the fit of nonparametric IRT models, also known as Mokken scale analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017).

The research in this thesis focuses on generalizing nonparametric IRT and MSA to multilevel test data. In multilevel test data respondents are nested in groups, such as students in classrooms or employees in companies. The focus may be on the respondents themselves (i.e., students or employees), or on a group aspect scored by respondents (e.g., experienced teacher support in the classroom or work climate at the company). A multilevel data structure can severely affect statistical analyses (e.g., Maas & Hox, 2005; Hox, 2010), possibly leading to the inclusion of items to the test that do not contribute to (or possibly negatively affect) accurate measurement, or the quality of the test may be overestimated. In addition, most methods in MSA provide results on the respondent level only, not on the group level. Hence, for multilevel test data, currently available MSA methods are only of limited value.

For multilevel test data, there exists a two-level nonparametric IRT model for scaling groups scored by multiple respondents, with two-level scalability coefficients for determining the degree to which a set of items form a single scale at the respondent level and the group level (Snijders, 2001a; similar to Mokken's scalability coefficients, see also, Mokken, 1971, p. 174). Scalability coefficients exist for item pairs, items, and the total test. Chapter 2 provides an introduction to weighted Guttman errors, necessary for computing scalability coefficients, and discusses solutions to two problems in computing weighted

Guttman errors that are currently not handled correctly by all software: Computing weighted Guttman errors if two items have the same estimated popularity, and computing weighted Guttman errors when the data have a two-level structure.

In Chapter 3 the standard errors for all two-level coefficients are derived. As a result, the precision of estimated scalability coefficients can be determined, leading to more information with respect to the scalability of the item-pairs, items, and the total test. Chapter 4 presents a simulation study in which the bias of the estimated two-level scalability coefficients, the bias of their estimated standard errors (with two estimation methods), and the coverage of the confidence intervals was investigated, under various testing conditions. Results showed negligible bias of the point estimates and standard errors and good coverage of the confidence intervals. Hence, the standard errors can be confidently used in practice. Chapter 5 provides a transformation for the all types of scalability coefficients and their standard errors, useful for very strong scales.

Chapter 6 derives point estimates, standard errors, and test statistics for Mokken's scalability coefficients in clustered data; that is, multilevel data with the focus on respondents rather than on groups. In addition, a simulation study compares traditional estimation methods to these derived estimation methods, and Wald-based methods to range-preserving methods from Chapter 5 for scalability coefficients in clustered data. Chapter 7 uses the results from Chapter 5 and 6 to incorporate significance tests in an automated item selection algorithm, to take sample fluctuations into account. The result was integrated into a two-step, test-guided MSA for scale construction, to guide the analysis for nonclustered and clustered data.

Chapter 8 introduces four two-level nonparametric IRT models, their assumptions, the implied stochastic ordering properties, and the implied observable data properties that are useful for model fit investigation. Relations between models and properties are presented. Finally, Chapter 9 presents a general discussion of the results from the previous chapters, guidelines for practical use, and suggestions for future research. The developments throughout this thesis were illustrated using real-data examples and implemented in R (R Core Team, 2020) in the package `mokken` (Van der Ark, 2007, 2012).

Samenvatting/Summary in Dutch

Non-parametrische Item-responstheorie voor Multilevel Toetsdata

Psychologisch meten richt zich op het meten van een psychologisch construct (zoals een eigenschap, vaardigheid of vermogen) door middel van observeerbare variabelen zoals de items (vragen) van een toets of vragenlijst (bijv., Sijtsma & Van der Ark, 2020, pp. 5–6). Om de itemscores te relateren aan het psychologische construct, worden item-responstheorie (IRT) modellen gebruikt. Een deelgroep van deze modellen zijn non-parametrische IRT modellen, die relatief goed passen op toetsdata in vergelijking met andere IRT modellen (Mokken, 1971; Sijtsma & Molenaar, 2002). Een passend non-parametrisch IRT model impliceert dat de somscore over alle items kan worden gebruikt om respondenten te rangschikken en, voor sommige modellen, dat de gemiddelde itemscore over respondenten kan worden gebruikt om de items te rangschikken. IRT modellen en hun eigenschappen zijn alleen valide als ze goed passen bij de toetsdata (Sijtsma & Van der Ark, 2020, p. 3): Het valideren van een toets of vragenlijst met behulp van IRT modellen is daarom een essentiële stap voordat deze voor meting wordt gebruikt (zie ook, Borsboom et al., 2004). Er bestaan verschillende methoden om de passendheid van non-parametrische IRT modellen te evalueren, ook wel bekend als Mokkenschaalanalyse (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017).

Het onderzoek in dit proefschrift richt zich op het generaliseren van non-parametrische IRT en MSA naar multilevel toetsdata. Multilevel toetsdata zijn data waarbij de respondenten genest zijn in groepen, zoals studenten in klassen of werknemers in bedrijven. De nadruk kan liggen op de respondenten zelf (de studenten of werknemers), of op een groepsaspect gescoord door de respondenten (zoals de steun die studenten ervaren van de docent of het werkklimaat in het bedrijf). Een multilevel datastructuur kan statistische analyses ernstig beïnvloeden (zie, bijv., Maas & Hox, 2005; Hox, 2010), waardoor mogelijk items in een toets worden opgenomen die niet bijdragen tot (of mogelijk een negatieve invloed hebben op) nauwkeurig meten, of de kwaliteit van de toets kan worden overschat. Bovendien leveren de meeste methoden in MSA alleen resultaten op respondentniveau, niet op groepsniveau. Voor multilevel toetsdata zijn de huidige MSA-methoden dus slechts van beperkte waarde.

Voor multilevel toetsdata bestaat er een twee-level non-parametrisch IRT model voor het schalen van groepen op basis van itemscores van meerdere respondenten, met twee-level schaalbaarheidscoëfficiënten voor het bepalen van de mate waarin de items één schaal

vormen, op respondentniveau en op groepsniveau (Snijders, 2001a; vergelijkbaar met Mokken's schaalbaarheidscoëfficiënten, zie ook, Mokken, 1971, p.174). Schaalbaarheidscoëfficiënten bestaan voor item-paren, items, en de totale toets. Hoofdstuk 2 introduceert gewogen Guttmanfouten, nodig voor het bereken van de schaalbaarheidscoëfficiënten, en bespreekt oplossingen voor twee problemen bij het berekenen van Guttmanfouten die momenteel niet door alle software correct worden behandeld: Het berekenen van gewogen Guttmanfouten wanneer twee items dezelfde geschatte populariteit hebben en het berekenen van gewogen Guttmanfouten wanneer de data een multilevel structuur hebben.

In hoofdstuk 3 worden standaardfouten voor alle tweelevel schaalbaarheidscoëfficiënten afgeleid. Als gevolg hiervan kan de precisie van geschatte schaalbaarheidscoëfficiënten worden bepaald, wat leidt tot meer informatie met betrekking tot de schaalbaarheid van de item-paren, items, en de totale toets. Hoofdstuk 4 presenteert een simulatiestudie waarin de bias van de geschatte tweelevel schaalbaarheidscoëfficiënten, de bias van hun geschatte standaardfouten (met twee schattingsmethoden) en de dekking van de betrouwbaarheidsintervallen werd onderzocht, onder verschillende condities. De resultaten toonden een verwaarloosbare bias van de puntschattingen en standaardfouten en een goede dekking van de betrouwbaarheidsintervallen. De standaardfouten kunnen dus goed in de praktijk worden gebruikt. Hoofdstuk 5 geeft een transformatie voor alle soorten schaalbaarheidscoëfficiënten en hun standaardfouten, vooral bruikbaar voor zeer sterke schalen.

Hoofdstuk 6 leidt puntschattingen, standaardfouten en toetsstatistieken af voor Mokken's schaalbaarheidscoëfficiënten in geclusterde toetsdata, dat is, multilevel toetsdata met de nadruk op respondenten in plaats van op groepen. Daarnaast wordt een simulatiestudie gepresenteerd die traditionele schattingsmethoden vergelijkt met deze afgeleide schattingsmethoden. Hoofdstuk 7 gebruikt de resultaten uit hoofdstuk 5 en hoofdstuk 6 om significantietoetsen te integreren in een geautomatiseerd itemselectiealgoritme, waardoor er rekening gehouden wordt met steekproeffluctuatie. Het resultaat werd geïntegreerd in een twee-staps, toets-geleide MSA voor schaalconstructie, als leidraad voor het analyseren van zowel niet-geclusterde en geclusterde toetsdata.

Hoofdstuk 8 introduceert vier twee-level non-parametrische IRT modellen, hun aannames en geïmpliceerde eigenschappen die nuttig zijn om de passendheid van de modellen te evalueren. Relaties tussen modellen en eigenschappen worden gepresenteerd. Tenslotte bevat Hoofdstuk 9 een algemene discussie van de resultaten uit de voorgaande hoofdstukken, richtlijnen voor praktisch gebruik en suggesties voor toekomstig onderzoek. De ontwikkelingen in dit proefschrift zijn geïllustreerd aan de hand van datavoorbeelden en geïmplementeerd in R (R Core Team, 2020) in het pakket `mokken` (Van der Ark, 2007, 2012).

Publications

Chapter 2 is published as:

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2017). Weighted Guttman errors: Handling ties and two-level data. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 183–190). Springer. doi: 10.1007/978-3-319-56294-0_17

LK programmed the developments, wrote the draft of the article, revised the article, and prepared it for submission. BJHZ and LAA provided feedback throughout the writing process and revised parts of the article.

Chapter 3 is published as:

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2020). Standard errors of two-level scalability coefficients. *British Journal of Mathematical and Statistical Psychology*, 73(2), 213–236. doi: 10.1111/bmsp.12174

LK derived the methods, programmed the developments, wrote the draft of the article, revised the article, and prepared it for submission. BJHZ and LAA provided feedback throughout the writing process and revised parts of the article.

Chapter 4 is published as:

Koopman, L., Zijlstra, B. J. H., De Rooij, M., & Van der Ark, L. A. (2020). Bias of two-level scalability coefficients and their standard errors. *Applied Psychological Measurement*, 44(3), 213–236. doi: 10.1177/0146621619843821

LK programmed the simulation study, wrote the draft of the article, revised the article, and prepared it for submission. BJHZ, MR, and LAA provided feedback throughout the writing process and revised parts of the article.

Chapter 5 is published as:

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2021). Range-preserving confidence intervals and significance tests for scalability coefficients in Mokken scale analysis. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 85th Annual Meeting of the Psychometric Society, Virtual* (pp. 175–185). Springer. doi: 10.1007/978-3-030-74772-5_16

LK and LAA thought of two transformations for scalability coefficients, LK programmed the computational developments and simulation study, wrote the draft of the article and prepared it for submission. BJHZ and LAA provided feedback throughout the writing process and revised parts of the article.

Chapter 6 is published as Supplementary file 1 and Chapter 7 as the main article of:

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2022). A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. *Quality of Life Research*, 31(1), 25–36. Springer. doi: 10.1007/s11136-021-02840-2

LK programmed the simulation study and computational developments, wrote the draft of the article, revised the article, and prepared it for submission. BJHZ and LAA provided feedback throughout the writing process and revised parts of the article.

Chapter 8 is submitted for publication as:

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (submitted for publication). *Assumptions and properties of two-level nonparametric item response theory models.*

LK wrote the draft of, and revised, the article. BJHZ and LAA provided feedback throughout the writing process and revised parts of the article.

Acknowledgements/Dankwoord

Allereerst gaat mijn eeuwige dank uit naar mijn promotor Andries en copromotor Bonne (in mijn persoonlijke aantekeningen ook wel terug te vinden als Bondries). Jullie zijn een geweldig team gebleken voor het begeleiden op een betrokken, constructieve, inhoudelijke, motiverende, ondersteunende, vakoverstijgende, bereikbare, carrière-in-het-vizier-houdende, persoonlijke-leven-niet-vergetende manier. Zonder jullie had ik dit hoge niveau niet kunnen bereiken (en waren mijn proofs ongetwijfeld niet zo nauwkeurig nagekeken). Ik realiseer mij nu pas hoezeer ik een student van jullie ben; geneigd naar conservatieve en robuuste methoden, met oog voor praktische relevantie en bruikbaarheid. Ik prijs me gelukkig dat ik door kan en wil gaan op dit pad. Commissieleden, bedankt dat jullie bereid waren een oordeel te vellen over dit werk, jullie zijn een grote inspiratie geweest voor veel onderdelen in dit proefschrift.

Daarnaast wil ik graag mijn directe medepromovendi bedanken die de broodnodige gezelligheid, ondersteuning en inspiratie hebben geboden: Laura Kolbe, Debby ten Hove en Hannelies de Jonge, alsmede mijn andere collega's uit D7.24, de Methoden en Techniekengroep, POW en de UvA. Ook dank aan IOPS voor het bieden van een stimulerend netwerk buiten de UvA, met name tijdens de vele conferenties, cursussen en borrels die ik heb bijgewoond. Bedankt NWO voor het vertrouwen in dit project en het toekennen van de Onderzoekstalentbeurs.

Heel veel dank aan mijn familie, schoonfamilie en vrienden: Jullie zijn stuk voor stuk toppers. Maar uiteraard de allergrootste dank aan Max, omdat je me wel en niet serieus neemt, je grappig bent, je zo goed bent in dingen waar ik slecht in ben, we zo'n rijk leven hebben opgebouwd en je minstens zo enthousiast bent als ik om tijdelijk in Oslo te gaan wonen. En natuurlijk Valentina en Flip, omdat jullie zo lief en vrolijk zijn, en het leven naast (en soms tijdens) werk zo gezellig (en druk) maken.