Nonparametric item response theory for multilevel test data

Koopman, L.

Publication date
2022

Chapter 3

Standard Errors of Scalability Coefficients in Two-Level Mokken Scale Analysis

Abstract

For the construction of tests and questionnaires that require multiple raters—e.g., a child behavior checklist completed by both parents—a novel ordinal scaling technique is currently being further developed, called two-level Mokken scale analysis. The technique uses within-rater coefficients and between-rater coefficients to assess the scalability of the test. These coefficients are generalizations of Mokken's scalability coefficients. In this chapter, we derived standard errors for the two-level coefficients and for their ratios. The coefficients, the estimates, the estimated standard errors, and the software implementation are discussed and illustrated using a real-data example, and a small-scale simulation study demonstrates the accuracy of the estimates.

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

3.1 Introduction

Mokken scale analysis is a popular nonparametric scaling technique (Mokken, 1971, also, see Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017). It is used for test construction in many areas of the social and behavioral sciences and related fields. Recent examples include clinical psychology (e.g., Freedland et al., 2017; Chou et al., 2017), education (e.g., Joe et al., 2017; Y. Chen et al., 2016), tourism (e.g., Coromina & Camprubí, 2016), health practice (e.g., Swiger et al., 2017), and medicine (e.g., Banas et al., 2017; Ahmadi et al., 2016). Mokken scale analysis consists of several procedures to check the assumptions of the underlying nonparametric item response theory model and an automated item selection procedure to select items from a pool of items. Arguably, the best known aspect of Mokken scale analysis are the scalability coefficients, also known as H coefficients, which are instrumental for defining the degree to which a set of items form a single scale (Mokken, 1971, p. 174). Scalability coefficients are available for item pairs, items, and the entire set of items. In this chapter, we refer to Mokken's orginal scalability coefficients as single-level scalability coefficients. We are currently developing Mokken scale analysis for two-level data based on the ideas of Snijders (2001a, also, see Crişan et al., 2016; Reise et al., 2006), who generalized Mokken's scalability coefficients to two-level data. This study discusses the next step in the development of two-level Mokken scale analysis: deriving standard errors of the two-level scalability coefficients, which are needed for a sound interpretation. Future steps in the development of two-level Mokken scale analysis include the development of an automated item selection procedure and methods to test assumptions of underlying item response theory models.

Mokken scale analysis for two-level data can be applied when subjects are assessed by several raters, for example when measuring the classroom environment using the pupils' ratings on several items of the WIHIC questionnaire (Fraser et al., 1996). Typically, all pupils in a class respond to the questionnaire, and the average test score across pupils is the measured value of the classroom environment. Other examples include child behavior rated by parents, caregivers, or teachers (e.g., Achenbach et al., 2008), teaching behavior rated by students (Maulana et al., 2015), evaluation of university courses rated by participants (e.g., Rampichini et al., 2004), learning environments rated by interns (Boor et al., 2007), ecological settings such as a neighborhood rated by the inhabitants (Raudenbush & Sampson, 1999), and leadership rated by the employees in a work group (Dyer et al., 2005). In these examples, the raters at level 1 (i.e., pupils, students, participants, etc.) are nested within the subjects at level 2 (i.e., classroom, teachers, courses, etc.), but in contrast to most multilevel examples, the interest lies in scaling the subject scores at level 2. Crişan et al. (2016) found that ignoring the two-level structure results in inflated reliability and scalability coefficients. The main problem is that multilevel measurement instruments intend to measure the trait level of the subjects, whereas common item analyses provide information on the raters.

Mokken (1971, pp. 164-169) derived asymptotic standard errors for the single-level

total-scale scalability coefficient for dichotomous items, which could be applied to small numbers of items only, and Van Onna (2004) used several computer intensive methods to compute the sampling distribution of the single-level total-scale scalability coefficient for polytomous items. More recently, under the assumption that the response patterns follow a multinomial distribution, Kuijpers et al. (2013) derived standard errors for all coefficients by means of a marginal modeling framework and the delta method. This method has a smaller computational burden and is therefore applicable to larger data sets, for both dichotomous and polytomous items. Kuijpers et al. (2016) showed that bias of the standard errors was negligible, and that the coverage of the 95% confidence intervals was satisfactory. The structure of two-level data is more complex than the structure of single-level data, so the method of Kuijpers et al. (2013) cannot be applied straightforwardly to two-level data. Three types of problems arise. The number of coefficients is three times larger for two-level data, the distributional assumptions of single-level data do not hold for two-level data, and probabilities should be estimated differently for two-level data. Applying the standard errors derived by Kuijpers et al. (2013) at two-level data is referred to as the *naive approach*. In the present study, these problems were tackled, resulting in corrected standard errors for all two-level scalability coefficients.

The remainder of this chapter is organized as follows. Section 3.2 demonstrates an application of two-level scalability coefficients. Section 3.3 briefly discusses latent variable models for two-level measurement. Section 3.4 discusses the two-level scalability coefficients proposed by Snijders (2001a) in more detail. Section 3.5 describes the mathematical derivation of standard errors and the implementation of the two-level scalability coefficients and their standard errors in software, followed by a discussion in Section 3.6. Throughout the upcoming sections we refer to Appendix A, where a small worked-out data example can be found to enhance understanding of the presented concepts and formulas.

3.2 Applying Two-Level Scalability Coefficients

For two-level Mokken scale analysis, Snijders (2001a) introduced nine different scalability coefficients. He distinguished three classes (our terminology) of scalability coefficients: *within-rater* and *between-rater* coefficients, and the *ratios* of the between-rater and within-rater coefficient. As for single-level coefficients, each class has three types: coefficients for each item pair, coefficients for each item, and a coefficient for the entire scale. All coefficients are denoted by the letter H. The class is indicated by the superscript of H: W for within-rater coefficients, B for between-rater coefficients, and BW for ratios of coefficients (i.e., $H^{BW} = H^B/H^W$). The type is indicated by the subscript of H: ij for item pairs, i for items, and no subscript for the entire set. Indices i and j are item indices, so for specific items, subscripts i and j may be replaced by the corresponding item numbers.

Within-rater scalability coefficients denote the consistency of item scores within raters. Their interpretation is very similar to the interpretation of Mokken's original (single-

level) coefficients, where there is just one rater. Between-rater scalability coefficients denote the consistency of item scores between raters of the same subject. The ratios of the between-rater and within-rater scalability coefficients denote the rater effect: Lower ratios indicate the need for a larger number of raters per subject. Item-pair scalability coefficients consider the item-scores of two items, and an I-item test contains $\binom{I}{2}$ item-pair coefficients H_{ij} for each class. Item scalability coefficients consider the scores of a single item with respect to the scores on all other items, and an I-item test contains I item coefficients H_i for each class. The coefficients for the entire set consider all item scores, and an I-item test contains one scale-coefficient H for each class. Computational details are provided later on.

The within-rater and between-rater scalability coefficients have a maximum value of 1, indicating a perfect correlation between all items. If all variation in item scores is due to random fluctuation, these coefficients have a value of 0. For all classes of two-level scalability coefficients, $\min(H_{ij}) \leq \min(H_i) \leq (H) \leq \max(H_i) \leq \max(H_{ij})$ (Sijtsma & Molenaar, 2002, p. 58). Furthermore, it is expected that $H_{ij}^W \geq H_{ij}^B$, $H_i^W \geq H_i^B$, and $H^W \geq H^B$ (Snijders, 2001a).

For ease of illustration for two-level scalability coefficients, we discuss the coefficients and their standard errors that were estimated on a small real-data set. The sample consisted of 14 upper-level primary-school teachers (the subjects) in the Netherlands. Each teacher was rated by a number of pupils (the raters). The number of pupils per class ranged between 5 and 39 (Mean = 18.50, SD = 10.22), and the total number of pupils was 259. Note that a sample of 14 subjects is not sufficient for test construction but we believe it suffices for this illustration. The pupils rated the teachers using a questionnaire measuring the teacher's autonomy support of pupils. Autonomy support consists of various behaviors such as providing choice, encouraging persistence at difficult activities, and acknowledging feelings (see e.g., Reeve et al., 2004). The data set contains the scores of all 259 pupils on 7 items of the questionnaire (Table 3.1). Each item has five ordered answer categories.

Table 3.1:
Subset of Seven Items Measuring Teachers' Autonomy Supportive Behavior

Item	Content
1	The teacher lets me choose what I am going to do
2	The teacher decides which task I will start with (inversely coded)
3	I get to choose which task I will start with
4	The teacher listens to me when I disagree with something
5	The teacher helps me when I ask for it
6	The teacher accepts me for who I am
7	The teacher helps me when I do not understand a task

Except for the item-pair ratios H_{ij}^{BW}, Table 3.2 shows the estimated two-level scalability coefficients, the naive standard errors, which ignore the nested structure of the data (in brackets), and the corrected standard errors as proposed in this chapter (in parentheses).

All point estimates of the scalability coefficients exceed zero, suggesting a positive relation between the items both within and between raters of the same subject. However, this does not take into account the precision of the estimates. When requiring that the lower bound of the 95% Wald-based confidence interval of the scalability coefficient ($H - 1.96SE$) exceeds zero, 65 of the 87 scalability coefficients exceed zero using the naive estimate, and only 19 exceed zero when using the corrected estimate. Specifically, the between-rater and ratio coefficients are not larger than zero with the corrected approach, thus in the population it is plausible that the items are unrelated on the subject level. Because zero is included in the interval, we cannot conclude that the teachers are consistently ordered based on the ratings of the pupils. This small example shows that the corrected standard errors are necessary to investigate the precision of the coefficient point estimates. This chapter explains how these standard errors can be derived.

3.3 Two-Level Measurement

In two-level test data, S subjects, indexed by s, are rated by a unique set of R_s raters each, indexed by r or q. We use two indices to distinguish between two raters in a pair. Note that each rater scores only one subject. The raters respond to I items, indexed by i or j. Each item has $m + 1$ ordered response categories, scored 0,1,...,m, indexed by x or y. Let X_{sri} denote the item score for subject s by rater r on item i, that is, $X_{sri} = x$ $(x = 0, \ldots, m)$. Subjects are generally scaled by their average score across raters:

$$\overline{X}_{s..} = \frac{1}{IR_s} \sum_{r=1}^{R_s} \sum_{i=1}^{I} X_{sri} \tag{3.1}$$

Several authors have proposed item response theory models for two-level test data. Snijders (2001a) proposed a nonparametric item response theory model that generalizes the Mokken (1971) model for monotone homogeneity to two-level data. Parametric item response theory models for two-level test data with an interest in scaling at level 2 include the ecometric model (Raudenbush & Sampson, 1999), the rater bundle model (Wilson & Hoskens, 2001), and the hierarchical rater model (Patz et al., 2002). For estimating scalability coefficients and deriving their standard errors it is not important which model triggers the item responses. The only assumption we make for estimating the scalability coefficients and deriving their standard errors is that the ordered item scores follow a multinomial distribution with varying multinomial parameters for each subject, which is true under all item response theory models.

Table 3.2:
Estimated Two-Level Scalability Coefficients With Naive Standard Errors in Brackets and Corrected Standard Errors in Parentheses.

	Item-pairs							*Items*		
	1	2	3	4	5	6	7	$\widehat{H}_i^W$	$\widehat{H}_i^B$	$\widehat{H}_i^{BW}$
1		.128*	.139*	.163*	.130*	.148*	.106*	.317**	.137*	.432*
		[.049]	[.053]	[.056]	[.055]	[.068]	[.049]	[.055]	[.044]	[.108]
		(.154)	(.184)	(.166)	(.169)	(.174)	(.155)	(.135)	(.160)	(.356)
2	.281*		.146*	.114	.123	.090	.088	.216*	.113*	.526*
	[.086]		[.070]	[.072]	[.063]	[.095]	[.074]	[.065]	[.057]	[.189]
	(.152)		(.136)	(.156)	(.144)	(.169)	(.154)	(.164)	(.142)	(.317)
3	.316**	.457**		.165*	.112	.184*	.097	.288**	.141*	.491*
	[.072]	[.072]		[.061]	[.063]	[.082]	[.074]	[.058]	[.050]	[.128]
	(.149)	(.178)		(.162)	(.177)	(.178)	(.172)	(.142)	(.157)	(.328)
4	.267*	.120	.193*		.165*	.190*	.132*	.308**	.154*	.500*
	[.082]	[.097]	[.088]		[.066]	[.083]	[.075]	[.060]	[.055]	[.133]
	(.166)	(.170)	(.167)		(.142)	(.163)	(.147)	(.126)	(.147)	(.356)
5	.280*	.114	.337**	.429**		.168*	.107	.357**	.136*	.381*
	[.080]	[.097]	[.087]	[.081]		[.074]	[.073]	[.057]	[.051]	[.121]
	(.173)	(.189)	(.134)	(.140)		(.161)	(.151)	(.122)	(.148)	(.308)
6	.346**	.214*	.231*	.453**	.487**		.128*	.362**	.150*	.416*
	[.082]	[.104]	[.105]	[.083]	[.092]		[.057]	[.055]	[.064]	[.145]
	(.165)	(.240)	(.227)	(.140)	(.143)		(.140)	(.152)	(.152)	(.272)
7	.435**	.183*	.192*	.374**	.461**	.388**		.337**	.111	.330*
	[.080]	[.084]	[.085]	[.076]	[.079]	[.073]		[.052]	[.059]	[.159]
	(.171)	(.222)	(.157)	(.108)	(.120)	(.179)		(.127)	(.144)	(.338)
Total	$\widehat{H}^W = .311^{**}$ [.048] (.130)				$\widehat{H}^B = .135^{*}$ [.048] (.146)			$\widehat{H}^{BW} = .433^{*}$ [.112] (.304)		

Note. Results for the items in Table 3.1. The upper triangle contains H_{ij}^B, the lower triangle H_{ij}^W. * = lower bound of the naive 95% Wald-based confidence interval exceeds zero.
** = lower bound of both the naive and corrected 95% Wald-based confidence intervals exceed zero

3.4 Scalability Coefficients

3.4.1 Within- and Between-Rater Probabilities

Let $\pi_{ij}^{xy(W)}$ be the within-rater bivariate probability $P(X_{sri} = x, X_{srj} = y)$; that is, the probability that rater r scores x on item i and y on item j. In addition, let $\pi_{ij}^{xy(B)}$ be the between-rater bivariate probability $P(X_{sri} = x, X_{sqj} = y)$; that is, the probability that for subject s, one rater (r) scores x on item i and another rater (q) scores y on item j. Furthermore, let π_i^x be the univariate probability $P(X_{sri} = x)$; that is, the probability that for subject s, rater r scores x on item i. Finally, let $\pi_{ij}^{xy(E)} = \pi_i^x \pi_j^y$ denote the expected bivariate probability under marginal independence of the items, that for subject

s, rater r scores x on item i and y on item j.

For I items and K item-pairs, there are $B = K(m+1)^2$ bivariate within-rater probabilities $\pi_{ij}^{xy(W)}$, B bivariate between-rater probabilities $\pi_{ij}^{xy(B)}$, B bivariate expected probabilities $\pi_{ij}^{xy(E)}$, and $U = I(m+1)$ univariate probabilities π_i^x. The population probabilities π are estimated by the sample proportions, p. For two-level data, this amounts to averaging the relative frequencies (Snijders, 2001a; Koopman et al., 2017). Let $\mathbf{1}(X_{sri} = x)$ be an indicator function that takes value 1 if $X_{sri} = x$ and value 0 otherwise. The within-rater bivariate proportion of item-score pattern $(X_{sri} = x, X_{srj} = y)$ is computed as

$$p_{ij}^{xy(W)} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{R_s}\sum_{r=1}^{R_s}\mathbf{1}(X_{sri} = x, X_{srj} = y). \tag{3.2}$$

The between-rater bivariate proportion of item-score pattern $(X_{sri} = x, X_{sqj} = y)$ is computed as

$$p_{ij}^{xy(B)} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{R_s(R_s-1)}\sum_{q\neq r}^{R_s}\mathbf{1}(X_{sri} = x, X_{sqj} = y). \tag{3.3}$$

The univariate proportion of item-score $(X_{sri} = x)$ is computed as

$$p_i^x = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{R_s}\sum_{r=1}^{R_s}\mathbf{1}(X_{sri} = x). \tag{3.4}$$

Finally, the expected bivariate proportion under marginal independence of the items is estimated as

$$p_{ij}^{xy(E)} = p_i^x p_j^y. \tag{3.5}$$

Appendix A.1 illustrates the computation of the bivariate and univariate proportions.

3.4.2 Weighted Guttman Errors

Let X_i denote the item-score of item i. Each item-score X_i has m item steps, denoted Z_{ix} for item i and item-step x ($i = 1, \ldots, I; x = 1, \ldots, m$), taking value 1 if the step has been passed ($Z_{ix} = 1$ if $X_i \geq x$) and 0 if the step has been failed ($Z_{ix} = 0$ if $X_i < x$). Let the popularity of an item-step be the probability of scoring value x or higher on item i, that is, $P(X_i \geq x)$.

Each item-pair has $2m$ item steps, $Z_{i1}, \ldots, Z_{im}, Z_{j1}, \ldots, Z_{jm}$, that need to be sorted in descending order of popularity. In a perfect Guttman scale no further item-steps are passed once an item-step is failed. Therefore, a Guttman error is described as failing a more popular item-step before passing a less popular item-step. As an example, the order of the item-steps for two items with 3 response categories may be

$$Z_{11}, Z_{21}, Z_{12}, Z_{22}. \tag{3.6}$$

Note that item-steps Z_{10} and Z_{20} are omitted, because $P(X_i \geq 0)$ equals 1 by definition. Replacing subscript ix in Equation 3.6 with $(g) = (1), (2), \ldots, (2m)$ results in item-steps $Z_{(1)}, Z_{(2)}, Z_{(3)}, Z_{(4)}$. Each item-step of Equation 3.6 is evaluated for a particular item-score pattern (x, y) as value z_g^{xy} and collected in vector $\mathbf{z}^{xy} = [z_1^{xy}\ z_2^{xy}\ \ldots\ z_{2m}^{xy}]$. For item-score pattern (0, 2), $\mathbf{z}^{02} = [0\ \ 1\ \ 0\ \ 1]$. For this pattern, the second and fourth item-step are passed ($z_2^{02} = z_4^{02} = 1$), whereas the first and third are failed ($z_1^{02} = z_3^{02} = 0$), resulting in a Guttman error. The *weight* of this error indicates the deviation from the perfect Guttman scale, by counting how many item steps are failed before passing a less popular item step (Molenaar, 1991). For pattern (0, 2) the weight is 3, because z_1^{02} is failed before z_2^{02} is passed, and z_1^{02} and z_3^{02} are failed before z_4^{02} is passed. Note that for admissible item-score patterns the weight results in value 0, and for dichotomous items the maximum weight is 1. In general, Guttman weights w_{ij}^{xy} for score x on item i and score y on item j can be computed as

$$w_{ij}^{xy} = \sum_{h=2}^{2m} \left\{ z_h^{xy} \times \left[\sum_{g=1}^{h-1} (1 - z_g^{xy}) \right] \right\}, \tag{3.7}$$

(see e.g., Kuijpers et al., 2013; Koopman et al., 2017). Weights are estimated in a sample as $\widehat{w}_{ij}^{xy}$ by ordering the item-steps according to their estimated item popularity $\widehat{P}(X_i \geq x) = \sum_x^m p_i^x$ (see also Appendix A.2).

3.4.3 Two-Level Scalability Coefficients

Scalability coefficients H compare the weighted sum of observed Guttman errors to the weighted sum of expected Guttman errors under marginal independence of the items (Sijtsma & Molenaar, 2002; Snijders, 2001a; Crişan et al., 2016). Item-pair scalability coefficients reflect the ratio of observed to expected weighted Guttman errors of an item-pair. The within- and between-rater scalability coefficients for item-pairs are defined as

$$H_{ij}^W = 1 - \frac{F_{ij}^W}{F_{ij}^E} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(W)}}{\sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(E)}} \tag{3.8}$$

and

$$H_{ij}^B = 1 - \frac{F_{ij}^B}{F_{ij}^E} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(B)}}{\sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(E)}}, \tag{3.9}$$

respectively. The denominator is equal for the within- and between-rater coefficient because they are based on the same marginal frequencies. Item scalability coefficients sum the weighted Guttman errors across all item-pairs pertaining item i. The within- and between-rater scalability coefficients for items are defined as

$$H_i^W = 1 - \frac{\sum_{j \neq i} F_{ij}^W}{\sum_{j \neq i} F_{ij}^E} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(W)}}{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(E)}} \tag{3.10}$$

and

$$H_i^B = 1 - \frac{\sum_{j \neq i} F_{ij}^B}{\sum_{j \neq i} F_{ij}^E} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(B)}}{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(E)}}, \tag{3.11}$$

respectively. Total scale scalability coefficients sum the weighted Guttman errors across all item-pairs. The within- and between-rater scalability coefficient for the total scale are defined as

$$H^W = 1 - \frac{\sum \sum_{j \neq i} F_{ij}^W}{\sum \sum_{j \neq i} F_{ij}^E} = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(W)}}{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(E)}} \tag{3.12}$$

and

$$H^B = 1 - \frac{\sum \sum_{j \neq i} F_{ij}^B}{\sum \sum_{j \neq i} F_{ij}^E} = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(B)}}{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} \pi_{ij}^{xy(E)}}, \tag{3.13}$$

respectively. The estimated scalability coefficients $\widehat{H}$ are computed by replacing π with p and w with $\widehat{w}$ in Equations 3.8 through 3.13. Appendix A.3 shows an example of estimating scalability coefficients using the proportions and estimated weights from the sample.

3.5 Estimating Standard Errors

We used the following strategy to derive standard errors. First, the scalability coefficients were written as vector functions of the data using a recursive exp-log notation (e.g., Kuijpers et al., 2013; Van der Ark et al., 2008), a technique often used in marginal modeling of categorical data (e.g., Bergsma et al., 2009, pp. 87-92). Second, the matrix of first-order partial derivatives of the vector function was derived. Finally, the delta method was applied (e.g., Agresti, 2012, pp. 577-581).

3.5.1 The Generalized Exp-Log Notation and the Delta Method

The Generalized Exp-Log Notation

The recursive exp-log notation may be used for functions of the data for which the matrices of partial derivatives are not readily obtained. It is a general method to rewrite these functions such that derivation of partial derivatives is easy to implement in software. Let $\mathbf{A}_1, \ldots, \mathbf{A}_c$ be design matrices whose values depend on the function that is written in the recursive exp-log notation. Let $\mathbf{n}$ be a vector of order $L = (m+1)^I$ containing the frequencies of all possible item-score patterns, each pattern taking the form $n_{12\ldots I}^{xx\ldots x}$. The patterns are ordered lexicographically with the last digit changing fastest, such that $\mathbf{n} = [n_{12\ldots I}^{00\ldots0}\ n_{12\ldots I}^{00\ldots1}\ \ldots\ n_{12\ldots I}^{mm\ldots m}]^T$. Let vector $\mathbf{n}_s$ be vector $\mathbf{n}$ for subject s, containing the frequencies of the item-score patterns for subject s. For an example of vector $\mathbf{n}$, see Appendix A.4. Let $\mathbf{g}(\mathbf{n})$ denote a vector function of the data. Finally, let $\exp(\mathbf{x})$ denote the element-

wise exponential of $\mathbf{x}$, and $\log(\mathbf{x})$ the element-wise natural logarithm of $\mathbf{x}$. The recursive exp-log notation writes $\mathbf{g}(\mathbf{n})$ as a series of nested functions $\mathbf{g}_0, \mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_c = \mathbf{g}(\mathbf{n})$; that is,

$$\mathbf{g}(\mathbf{n}) = \exp(\ \mathbf{A}_c\ \log(\ \mathbf{A}_{c-1} \ldots\ \exp(\ \mathbf{A}_2\ \log(\ \mathbf{A}_1 \underbrace{\mathbf{n}}_{\mathbf{g}_0}))))\,. \tag{3.14}$$

$$\underbrace{\underbrace{\underbrace{\underbrace{\mathbf{g}_0}_{\mathbf{g}_1}}_{\mathbf{g}_2}\ \vdots}_{\mathbf{g}_{c-1}}}_{\mathbf{g}_c}$$

Hence,

$$\mathbf{g}_i = \begin{cases} \mathbf{n} & \text{if } i = 0 \\ \log(\mathbf{A}_i\ \mathbf{g}_{i-1}) & \text{if } i \text{ is odd} \\ \exp(\mathbf{A}_i\ \mathbf{g}_{i-1}) & \text{if } i \text{ is even} \end{cases}. \tag{3.15}$$

Deriving the Matrix of First-Order Partial Derivatives

Let the Jacobian of $\mathbf{g}(\mathbf{n})$, which is the matrix of first-order partial derivatives with respect to $\mathbf{n}$, be $\mathbf{G} \equiv \mathbf{G}(\mathbf{n}) = \partial \mathbf{g}(\mathbf{n}) / \partial \mathbf{n}^T$, with $\mathbf{n}^T$ indicating the transpose of vector $\mathbf{n}$. For each $\mathbf{g}_i$ the Jacobian is $\mathbf{G}_i$. Rewriting the scalability coefficients in recursive exp-log notation enables the relatively straightforward computation of the Jacobian, because the chain rule can be applied recurrently. The chain rule is used to differentiate a function of a function, such as $y = \mathrm{g}(\mathrm{h}(x))$ (e.g., Stewart, 2008, p. 197, p. 904). First, substitute $\mathrm{h}(x)$ with u to obtain $y = \mathrm{g}(u)$. Then the derivative of y is

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}. \tag{3.16}$$

Let $\mathrm{Diag}(\mathbf{x})$ be a diagonal matrix with $\mathbf{x}$ on the diagonal, and $\mathrm{Diag}(\mathbf{x})^{-1}$ the inverse of matrix $\mathrm{Diag}(\mathbf{x})$. Applying the chain rule to function $\mathbf{g}_i$ $(i = 0, \ldots, c)$ results in

$$\mathbf{G}_i = \begin{cases} \mathbf{I} & \text{if } i = 0 \\ \mathrm{Diag}(\mathbf{A}_i\ \mathbf{g}_{i-1})^{-1}\ \mathbf{A}_i\ \mathbf{G}_{i-1} & \text{if } i \text{ is an odd number} \\ \mathrm{Diag}\big(\exp(\mathbf{A}_i\ \mathbf{g}_{i-1})\big)\ \mathbf{A}_i\ \mathbf{G}_{i-1} & \text{if } i \text{ is an even number} \end{cases}. \tag{3.17}$$

Applying the Delta Method

The delta method approximates the variance of the transformation of a variable by using a one-step Taylor approximation (e.g., Agresti, 2012, pp. 577-594). Let $\mathbf{V}_\mathbf{n}$ be the variance-covariance matrix of vector $\mathbf{n}$. According to the delta method, the variance-covariance matrix of the transformation of vector $\mathbf{n}$, $\mathbf{V}_{\mathbf{g}(\mathbf{n})}$, is approximated by

$$\mathbf{V}_{\mathbf{g}(\mathbf{n})} \approx \mathbf{G}\ \mathbf{V}_\mathbf{n}\ \mathbf{G}^T. \tag{3.18}$$

The standard errors, collected in $\mathbf{SE}_{\mathbf{g(n)}}$, are retrieved by taking the square root of the diagonal of $\mathbf{V}_{\mathbf{g(n)}}$. The variance-covariance matrix and the standard errors are estimated in the sample as $\widehat{\mathbf{V}}_{\mathbf{g(n)}}$ and $\widehat{\mathbf{SE}}_{\mathbf{g(n)}}$, respectively.

A Simple Example

A simple example of the recursive exp-log notation is provided to enhance understanding of the method, before moving on to rewriting the scalability coefficients. In this example we derive the standard errors of the sample proportions, p_a and p_b, for dichotomous items X_a and X_b, respectively. Let n_{ij}^{xy} denote the frequency of respondents scoring x on item i and y on item j. The item-score frequencies of items X_a and X_b are lexicographically stored in vector $\mathbf{n} = [n_{ab}^{00}\ n_{ab}^{01}\ n_{ab}^{10}\ n_{ab}^{11}]^T$. For item X_i, a simple calculation results in the sample proportion $p_i = n_{ij}^{1+}/N = (n_{ij}^{10} + n_{ij}^{11})/N$, with N the total number of observations. The proportions can be computed using the recursive exp-log notation. Let

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \tag{3.19}$$

and $[p_a\ p_b]^T = \mathbf{g}(\mathbf{n}) = \exp\left(\mathbf{A}_2 \log\left(\mathbf{A}_1\mathbf{n}\right)\right)$ the transformation of $\mathbf{n}$. First, $\mathbf{g}_0 = \mathbf{n}$. Then, following Equation 3.15,

$$\mathbf{g}_1 = \log(\mathbf{A}_1\mathbf{g}_0) = \log\left(\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} n_{ab}^{00} \\ n_{ab}^{01} \\ n_{ab}^{10} \\ n_{ab}^{11} \end{pmatrix}\right) = \log\begin{pmatrix} n_{ab}^{1+} \\ n_{ab}^{+1} \\ N \end{pmatrix}. \tag{3.20}$$

and

$$\mathbf{g}(\mathbf{n}) = \mathbf{g}_2 = \exp(\mathbf{A}_2\mathbf{g}_1) = \exp\left(\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}\log\begin{pmatrix} n_{ab}^{1+} \\ n_{ab}^{+1} \\ N \end{pmatrix}\right) = \begin{pmatrix} p_a \\ p_b \end{pmatrix}. \tag{3.21}$$

Following Equation 3.17, $\mathbf{G}_0 = \mathbf{I}$,

$$\begin{aligned} \mathbf{G}_1 &= \mathrm{Diag}(\mathbf{A}_1\mathbf{g}_0)^{-1}\mathbf{A}_1\mathbf{G}_0 \\ &= \mathrm{Diag}(\mathbf{A}_1\mathbf{g}_0)^{-1}\mathbf{A}_1\ \mathbf{I} \\ &= \mathrm{Diag}(\mathbf{A}_1\mathbf{g}_0)^{-1}\mathbf{A}_1 \\ &= \begin{pmatrix} 1/n_{ab}^{1+} & 0 & 0 \\ 0 & 1/n_{ab}^{+1} & 0 \\ 0 & 0 & 1/N \end{pmatrix}\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 1/n_{ab}^{1+} & 1/n_{ab}^{1+} \\ 0 & 1/n_{ab}^{+1} & 0 & 1/n_{ab}^{+1} \\ 1/N & 1/N & 1/N & 1/N \end{pmatrix}, \end{aligned} \tag{3.22}$$

and

$$\begin{aligned}
\mathbf{G} &= \mathbf{G}_2 = \text{Diag}(\exp(\mathbf{A}_2\mathbf{g}_1))\mathbf{A}_2\mathbf{G}_1 \\
&= \text{Diag}\left(\exp\left(\mathbf{A}_2\ \mathbf{g}_1\right)\right)\mathbf{A}_2\ \text{Diag}\left(\mathbf{A}_1\ \mathbf{g}_0\right)^{-1}\mathbf{A}_1 \\
&= \text{Diag}\left(\mathbf{g}_2\right)\mathbf{A}_2\ \text{Diag}\left(\mathbf{A}_1\ \mathbf{n}\right)^{-1}\mathbf{A}_1 \\
&= \begin{pmatrix} p_a & 0 \\ 0 & p_b \end{pmatrix}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}\begin{pmatrix} 0 & 0 & 1/n_{ab}^{1+} & 1/n_{ab}^{1+} \\ 0 & 1/n_{ab}^{+1} & 0 & 1/n_{ab}^{+1} \\ 1/N & 1/N & 1/N & 1/N \end{pmatrix} \\
&= N^{-1}\begin{pmatrix} -p_a & -p_a & 1-p_a & 1-p_a \\ -p_b & 1-p_b & -p_b & 1-p_b \end{pmatrix}.
\end{aligned} \tag{3.23}$$

Vector $\mathbf{n}$ is assumed to follow a multinomial distribution with parameters N and $\mathbf{p} = \mathbf{n}/N = [p_{ab}^{00}\ p_{ab}^{01}\ p_{ab}^{10}\ p_{ab}^{11}]^T$, resulting in the estimated variance-covariance matrix

$$\begin{aligned}
\mathbf{V_n} &= N\left[\text{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T\right] \\
&= N\begin{pmatrix} p_{ab}^{00}(1-p_{ab}^{00}) & -p_{ab}^{00}p_{ab}^{01} & -p_{ab}^{00}p_{ab}^{10} & -p_{ab}^{00}p_{ab}^{11} \\ -p_{ab}^{01}p_{ab}^{00} & p_{ab}^{01}(1-p_{ab}^{01}) & -p_{ab}^{01}p_{ab}^{10} & -p_{ab}^{01}p_{ab}^{11} \\ -p_{ab}^{10}p_{ab}^{00} & -p_{ab}^{10}p_{ab}^{01} & p_{ab}^{10}(1-p_{ab}^{10}) & -p_{ab}^{10}p_{ab}^{11} \\ -p_{ab}^{11}p_{ab}^{00} & -p_{ab}^{11}p_{ab}^{01} & -p_{ab}^{11}p_{ab}^{10} & p_{ab}^{11}(1-p_{ab}^{11}) \end{pmatrix}.
\end{aligned} \tag{3.24}$$

Using Equation 3.18 to estimate the variance-covariance matrix of $\mathbf{g}(\mathbf{n})$, it may be verified that

$$\begin{aligned}
\widehat{\mathbf{V}}_{\mathbf{g(n)}} &= \mathbf{G}\ \mathbf{V_n}\ \mathbf{G}^T \\
&= N^{-1}\begin{pmatrix} p_a(1-p_a) & -p_ap_b \\ -p_bp_a & p_b(1-p_b) \end{pmatrix}.
\end{aligned} \tag{3.25}$$

The variance of the sampling distribution of p_a and p_b are the diagonal elements of $\widehat{\mathbf{V}}_{\mathbf{g(n)}}$ (Equation 3.25) and equal the well-known asymptotic variance estimator of the multinomial sampling distribution $p_a(1-p_a)/N$ and $p_b(1-p_b)/N$, respectively, with the standard errors being its square root.

3.5.2 Standard Errors of Two-Level Scalability Coefficients

The two main challenges of applying the exp-log notation and the delta method are the construction of design matrices $\mathbf{A}_1$ to $\mathbf{A}_c$ for all 9 two-level scalability coefficients, and the specification of an appropriate distribution for vector $\mathbf{n}$ with the derivation of its variance-covariance matrix. This section demonstrates the construction of the design matrices for the item-pair, item, and total scale coefficients, respectively, for all classes of coefficients, followed by a section on deriving the variance-covariance matrix of vector $\mathbf{n}$.

Let $\mathbf{H}_{ij}^B = [H_{12}^B\ H_{13}^B \ldots H_{(I-1,I)}^B]^T, \mathbf{H}_{ij}^W = [H_{12}^W\ H_{13}^W \ldots H_{(I-1,I)}^W]^T$, and $\mathbf{H}_{ij}^{BW} = [H_{12}^{BW}\ H_{13}^{BW} \ldots H_{(I-1,I)}^{BW}]^T$ be vectors of size K, containing the between-rater item-pair coefficients, the within-rater item-pair coefficients, and the ratios of item-pair coefficients, respectively. Let $\mathbf{H}_{ij} = \mathbf{g}(\mathbf{n}) = [\mathbf{H}_{ij}^{B^T}\ \mathbf{H}_{ij}^{W^T}\ \mathbf{H}_{ij}^{BW^T}]^T$ be a vector of size $3K$ containing all item-pair coefficients. Similarly, let $\mathbf{H}_i = \mathbf{g}^\dagger(\mathbf{n}) = [\mathbf{H}_i^{B^T}\ \mathbf{H}_i^{W^T}\ \mathbf{H}_i^{BW^T}]^T$ be a vector of size $3I$ containing all item coefficients, and let $\mathbf{H} = \mathbf{g}^\ddagger(\mathbf{n}) = [H^B\ H^W\ H^{BW}]^T$ be a vector of

size 3 containing the three total-scale coefficients.

Item-Pair Scalability Coefficients in Exp-Log Notation

The recursive exp-log notation to compute the two-level item-pair scalability coefficients is

$$\mathbf{H}_{ij} = \mathbf{g}(\mathbf{n}) = \exp(\mathbf{A}_6 \log(\mathbf{A}_5 \exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n})))))). \tag{3.26}$$

The $(2B + U) \times L$ matrix $\mathbf{A}_1$ contains submatrices $\mathbf{B}^B$, $\mathbf{B}^W$, and $\mathbf{U}$,

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{B}^B \\ \mathbf{B}^W \\ \mathbf{U} \end{pmatrix}. \tag{3.27}$$

Matrices $\mathbf{B}^B$ and $\mathbf{B}^W$ link the observed item-score frequencies to the bivariate between- and within-rater proportions, respectively and matrix $\mathbf{U}$ to the univariate proportions. Let $\mathbf{p}_s^B = \mathbf{n}_s/(SR_s(R_s-1))$ be a vector containing item-score proportions for the between-rater proportions per subject and let $\mathbf{p} = \sum_{s=1}^{S} \mathbf{n}_s/(SR_s)$ be a vector containing the sample proportions of the item-score patterns in vector $\mathbf{n}$. Let subscript (l) $(l = 1, 2, \ldots L)$ represent the l-th element of a vector. Also, let $\mathbf{1}(X_{i(l)} = x)$ denote an indicator function of score x on item i on the l-th item-score pattern of vector $\mathbf{n}$. Finally, n_{sj}^y denotes the frequency of raters scoring y on item j for subject s.

For the b-th bivariate proportion (x, y) and the l-th item-score pattern, entry (b, l) of the $B \times L$ submatrix $\mathbf{B}^B$ takes value $\mathbf{1}(X_{i(l)} = x)[\sum_{s=1}^{S}(n_{sj}^y - \mathbf{1}(X_{j(l)} = y))p_{s(l)}^B]/n_{(l)}$. In the $B \times L$ submatrix $\mathbf{B}^W$ entry (b, l) takes value $\mathbf{1}(X_{i(l)} = x, X_{j(l)} = y)p_{(l)}/n_{(l)}$ for the b-th bivariate proportion and the l-th item-score pattern. Element (u, l) of the $U \times L$ submatrix $\mathbf{U}$ takes value $\mathbf{1}(X_{i(l)} = x)p_{(l)}/n_{(l)}$ for the u-th univariate proportion and the l-th item-score pattern. For a small-scale example of matrix $\mathbf{A}_1$ see Appendix A.5, Table 5.

Multiplying matrix $\mathbf{A}_1$ with vector $\mathbf{n}$ results in a vector containing the bivariate between-rater proportions ($\mathbf{p}_{ij}^B = [p_{12}^{00(B)}\ p_{12}^{01(B)}\ \ldots\ p_{(I-1),I}^{mm(B)}]^T$, Equation 3.3), within-rater proportions ($\mathbf{p}_{ij}^W = [p_{12}^{00(W)}\ p_{12}^{01(W)}\ \ldots\ p_{(I-1),I}^{mm(W)}]^T$, Equation 3.2), and univariate proportions ($\mathbf{p}_i = [p_1^0\ p_1^1\ \ldots\ p_I^m]^T$, Equation 3.4). Hence, function $\mathbf{g}_1$ equals

$$\mathbf{g}_1 = \log(\mathbf{A}_1\ \mathbf{n}) = \log \begin{pmatrix} \mathbf{p}_{ij}^B \\ \mathbf{p}_{ij}^W \\ \mathbf{p}_i \end{pmatrix}. \tag{3.28}$$

Design matrices $\mathbf{A}_2$ to $\mathbf{A}_5$ are adjusted versions of matrices $\mathbf{A}_2$ to $\mathbf{A}_5$ in Kuijpers et al. (2013, pp. 61 - 63). Let $\mathbf{1}_{(v)}$ and $\mathbf{0}_{(v)}$ denote a unit vector and zero vector, respectively, of length v, let $\mathbf{I}_{(v)}$ denote the $v \times v$ identity matrix, and let $\mathbf{0}$ reflect a zero-matrix or vector, which order depends on the order of its neighbouring matrices. Let $\mathbf{P}$ be a $B \times U$ indicator matrix where entry (b, u) takes value 1 if the u-th univariate proportion contributes to the b-th expected bivariate proportion $p_{ij}^{xy(E)}$ (Equation 3.5), and 0 otherwise. The $3B \times$

$(2B + U)$ matrix $\mathbf{A}_2$ equals

$$\mathbf{A}_2 = \begin{pmatrix} \mathbf{I}_{(2B)} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} \end{pmatrix}. \tag{3.29}$$

Let $(\mathbf{p}_{ij}^E = [p_{12}^{00(E)}\ p_{12}^{01(E)}\ \ldots\ p_{(I-1),I}^{mm(E)}]^T)$ be the vector containing the expected bivariate proportions under marginal independence of the items. Using the result in Equation 3.28 for $\mathbf{g}_1$, function $\mathbf{g}_2$ equals

$$\mathbf{g}_2 = \exp(\ \mathbf{A}_2\ \mathbf{g}_1\ ) = \begin{pmatrix} \mathbf{p}_{ij}^B \\ \mathbf{p}_{ij}^W \\ \mathbf{p}_{ij}^E \end{pmatrix}. \tag{3.30}$$

Let $\oplus$ indicate the direct sum. Vector $\mathbf{w}_{ij} = [w_{ij}^{00}\ w_{ij}^{01}\ \ldots\ w_{ij}^{mm}]^T$ contains the $(m+1)^2$ weights for item-pair (i, j) (Equation 3.7). The $K \times B$ block-diagonal matrix $\mathbf{W}$ contains the weights for all K pairs of items; that is,

$$\mathbf{W} = \bigoplus_{i<j}^{I} \mathbf{w}_{ij}^T = \begin{pmatrix} \mathbf{w}_{12}^T & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_{13}^T & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w}_{14}^T & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{w}_{I-1,I}^T \end{pmatrix}. \tag{3.31}$$

Let vector $\mathbf{c}$ be a copy of the first row of $\mathbf{W}$, necessary to construct scalar 1 in Equations 3.8 and 3.9, and $\mathbf{0}_{(B)}$ be a zero-vector of length B. Let $\otimes$ denote the Kronecker product. Then, the $(3K + 1) \times 3B$ matrix $\mathbf{A}_3$ is given by

$$\mathbf{A}_3 = \begin{pmatrix} \mathbf{c}^T & \mathbf{0}_{(2B)}^T \\ \mathbf{I}_{(3)} \otimes \mathbf{W} \end{pmatrix}. \tag{3.32}$$

Let $\mathbf{F}_{ij} = [F_{12}\ F_{13}\ \ldots\ F_{I-1,I}]^T$ be the vector containing the weighted sum of Guttman errors, using superscript B, W and E for the observed between-rater, observed within-rater, and expected under marginal independence variant, respectively (Equations 3.8 and 3.9). Using the result in Equation 3.30 for $\mathbf{g}_2$, $\mathbf{g}_3$ equals

$$\mathbf{g}_3 = \log(\mathbf{A}_3\ \mathbf{g}_2) = \log \begin{pmatrix} \mathbf{w}_{12}^T \mathbf{p}_{12}^B \\ \mathbf{W}\ \mathbf{p}_{ij}^B \\ \mathbf{W}\ \mathbf{p}_{ij}^W \\ \mathbf{W}\ \mathbf{p}_{ij}^E \end{pmatrix} = \log \begin{pmatrix} F_{12}^B \\ \mathbf{F}_{ij}^B \\ \mathbf{F}_{ij}^W \\ \mathbf{F}_{ij}^E \end{pmatrix}. \tag{3.33}$$

The $(2K + 1) \times (3K + 1)$ matrix $\mathbf{A}_4$ is given by

$$\mathbf{A}_4 = \begin{pmatrix} 1 & -1\ \mathbf{0}_{(2K-1)}^T & \mathbf{0}_{(K)}^T \\ \mathbf{0}_{(2K)} & \mathbf{I}_{(2K)} & -\mathbf{1}_{(2)} \otimes \mathbf{I}_{(K)} \end{pmatrix}. \tag{3.34}$$

Using Equation 3.33 for $\mathbf{g}_3, \mathbf{g}_4$ results in

$$\mathbf{g}_4 = \exp\left(\ \mathbf{A}_4\ \mathbf{g}_3\ \right) = \begin{pmatrix} 1 \\ \mathbf{F}^B_{ij}/\mathbf{F}^E_{ij} \\ \mathbf{F}^W_{ij}/\mathbf{F}^E_{ij} \end{pmatrix}. \tag{3.35}$$

The $2K \times (2K+1)$ matrix $\mathbf{A}_5$ is given by

$$\mathbf{A}_5 = \begin{pmatrix} \mathbf{1}_{(2K)} & -\mathbf{I}_{(2K)} \end{pmatrix}, \tag{3.36}$$

and $\mathbf{g}_5$ given by

$$\mathbf{g}_5 = \log\left(\mathbf{A}_5\ \mathbf{g}_4\right) = \log\begin{pmatrix} 1 - \mathbf{F}^B_{ij}/\mathbf{F}^E_{ij} \\ 1 - \mathbf{F}^W_{ij}/\mathbf{F}^E_{ij} \end{pmatrix} = \log\begin{pmatrix} \mathbf{H}^B_{ij} \\ \mathbf{H}^W_{ij} \end{pmatrix}. \tag{3.37}$$

Finally, the $3K \times 4K$ matrix $\mathbf{A}_6$ is given by

$$\mathbf{A}_6 = \begin{pmatrix} \mathbf{I}_{(3K)} & (0\ 0\ -1)^T \otimes \mathbf{I}_{(K)} \end{pmatrix}, \tag{3.38}$$

which gives

$$\mathbf{g}(\mathbf{n}) = \exp\begin{pmatrix}\ \mathbf{A}_6\ \mathbf{g}_5\end{pmatrix} = \begin{pmatrix} \mathbf{H}^B_{ij} \\ \mathbf{H}^W_{ij} \\ \mathbf{H}^{BW}_{ij} \end{pmatrix}, \tag{3.39}$$

the vector containing all item-pair scalability coefficients.

Item Scalability Coefficients in Exp-Log Notation

The recursive exp-log notation for the two-level item scalability coefficients is

$$\mathbf{H}_i = \mathbf{g}^\dagger(\mathbf{n}) = \exp(\mathbf{A}^\dagger_6 \log(\mathbf{A}^\dagger_5 \exp(\mathbf{A}^\dagger_4 \log(\mathbf{A}^\dagger_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{n})))))). \tag{3.40}$$

Design matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are used again in the computation of the item scalability coefficients. Design matrices $\mathbf{A}^\dagger_3$, $\mathbf{A}^\dagger_4$, $\mathbf{A}^\dagger_5$, and $\mathbf{A}^\dagger_6$ differ slightly from $\mathbf{A}_3$, $\mathbf{A}_4$, $\mathbf{A}_5$, and $\mathbf{A}_6$. The difference between item coefficients compared to the item-pair coefficients is that the weighted Guttman errors need to be summed over the item-pairs for each item i (Equation 3.10 and 3.11). Therefore, the steps up to computation of the weighted Guttman errors are identical.

Row i of the $I \times K(m+1)^2$ matrix $\mathbf{W}^\dagger$ pertains to item i. Each item-pair has $(m+1)^2$ columns, which contains vector $\mathbf{w}^T_{ij}$ if $j \neq i$ in row i, and a zero-vector for the columns belonging to the remaining item-pairs. Hence, matrix $\mathbf{W}^\dagger$ is

$$\mathbf{W}^\dagger = \begin{pmatrix} \mathbf{w}^T_{12} & \mathbf{w}^T_{13} & \dots & \mathbf{w}^T_{1I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{w}^T_{12} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{w}^T_{23} & \dots & \mathbf{w}^T_{2I} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{w}^T_{1I} & \mathbf{0} & \dots & \mathbf{w}^T_{2I} & \mathbf{0} & \dots & \mathbf{w}^T_{I-1,I} \end{pmatrix}. \tag{3.41}$$

Let vector $\mathbf{c}^\dagger$ be a copy of the first row of matrix $\mathbf{W}^\dagger$. Replacing $\mathbf{c}$ with $\mathbf{c}^\dagger$ and $\mathbf{W}$ with $\mathbf{W}^\dagger$ in matrix $\mathbf{A}_3$ (Equation 3.31) results in matrix $\mathbf{A}_3^\dagger$. Using the result in Equation 3.30 for $\mathbf{g}_2, \mathbf{g}_3^\dagger$ equals

$$\mathbf{g}_3^\dagger = \log\left(\mathbf{A}_3^\dagger\ \mathbf{g}_2\right) = \log\begin{pmatrix} \sum_{j\neq 1} F_{1j}^B \\ \sum_{j\neq i} \mathbf{F}_{ij}^B \\ \sum_{j\neq i} \mathbf{F}_{ij}^W \\ \sum_{j\neq i} \mathbf{F}_{ij}^E \end{pmatrix}. \tag{3.42}$$

Matrices $\mathbf{A}_4^\dagger$, $\mathbf{A}_5^\dagger$, and $\mathbf{A}_6^\dagger$ are obtained by changing K to I in the order of the submatrices and subvectors of $\mathbf{A}_4$ (Equation 3.35), $\mathbf{A}_5$ (Equation 3.37), and $\mathbf{A}_6$ (Equation 3.39), respectively; that is,

$$\mathbf{A}_4^\dagger = \begin{pmatrix} 1 & -1\ \mathbf{0}_{(2I-1)}^T & \mathbf{0}_{(I)}^T \\ \mathbf{0}_{(2I)} & \mathbf{I}_{(2I)} & -\mathbf{1}_{(2)} \otimes\ \mathbf{I}_{(I)} \end{pmatrix}, \tag{3.43}$$

$$\mathbf{A}_5^\dagger = \begin{pmatrix} \mathbf{1}_{(2I)} & -\mathbf{I}_{(2I)} \end{pmatrix}, \tag{3.44}$$

and

$$\mathbf{A}_6^\dagger = \begin{pmatrix} \mathbf{I}_{(3I)} & (0\ 0\ -1)^T \otimes \mathbf{I}_{(I)} \end{pmatrix}. \tag{3.45}$$

Using the result in Equation 3.42 for $\mathbf{g}_3^\dagger, \mathbf{g}^\dagger(\mathbf{n})$ equals

$$\mathbf{g}^\dagger(\mathbf{n}) = \exp(\mathbf{A}_6^\dagger \log(\mathbf{A}_5^\dagger \exp(\mathbf{A}_4^\dagger \mathbf{g}_3^\dagger))) = \begin{pmatrix} \mathbf{H}_i^B \\ \mathbf{H}_i^W \\ \mathbf{H}_i^{BW} \end{pmatrix}. \tag{3.46}$$

Total Scale Scalability Coefficients in Exp-Log Notation

The recursive exp-log notation for the two-level total scale scalability coefficients is

$$\mathbf{H} = \mathbf{g}^\ddagger(\mathbf{n}) = \exp(\mathbf{A}_6^\ddagger \log(\mathbf{A}_5^\ddagger \exp(\mathbf{A}_4^\ddagger \log(\mathbf{A}_3^\ddagger \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n})))))). \tag{3.47}$$

Similar to the changes for the item scalability coefficients (Section 3.5.2), the differences in the function to compute the total-scale coefficients only affect matrices $\mathbf{A}_3$ to $\mathbf{A}_5$. Submatrix $\mathbf{W}$ is reduced to a vector of order B containing all Guttman weights, $\mathbf{w}^\ddagger = [\mathbf{w}_{12}\ \mathbf{w}_{12}\ \ldots\ \mathbf{w}_{I-1,I}]^T$. Replacing both $\mathbf{c}$ and $\mathbf{W}$ with $\mathbf{w}^\ddagger$ in matrix $\mathbf{A}_3$ (Equation 3.31) results in matrix $\mathbf{A}_3^\ddagger$. Subsequently, $\mathbf{g}_3^\ddagger$ equals

$$\mathbf{g}_3^\ddagger = \log\left(\mathbf{A}_3^\ddagger\ \mathbf{g}_2\right) = \log\begin{pmatrix} \sum\sum_{j\neq i} F_{ij}^B \\ \sum\sum_{j\neq i} F_{ij}^B \\ \sum\sum_{j\neq i} F_{ij}^W \\ \sum\sum_{j\neq i} F_{ij}^E \end{pmatrix}. \tag{3.48}$$

Matrices $\mathbf{A}_4^\ddagger$, $\mathbf{A}_5^\ddagger$, and $\mathbf{A}_6^\ddagger$ are obtained by changing K to 1 in the order of the submatrices and subvectors of $\mathbf{A}_4$ (Equation 3.35), $\mathbf{A}_5$ (Equation 3.37), and $\mathbf{A}_6$ (Equation

3.39), respectively; that is,

$$\mathbf{A}_4^{\ddagger} = \begin{pmatrix} 1 & -1\ 0 & 0 \\ \mathbf{0}_{(2)} & \mathbf{I}_{(2)} & -\mathbf{1}_{(2)} \end{pmatrix}, \tag{3.49}$$

$$\mathbf{A}_5^{\ddagger} = \begin{pmatrix} \mathbf{1}_{(2)} & -\mathbf{I}_{(2)} \end{pmatrix}, \tag{3.50}$$

and

$$\mathbf{A}_6^{\ddagger} = \begin{pmatrix} \mathbf{I}_{(3)} & (0\ 0\ -1)^T \end{pmatrix}. \tag{3.51}$$

Finally, $\mathbf{g}^{\ddagger}(\mathbf{n})$ equals

$$\mathbf{g}^{\ddagger}(\mathbf{n}) = \exp(\mathbf{A}_6^{\ddagger} \log(\mathbf{A}_5^{\ddagger} \exp(\mathbf{A}_4^{\ddagger} \mathbf{g}_3^{\ddagger}))) = \begin{pmatrix} H^B \\ H^W \\ H^{BW} \end{pmatrix}. \tag{3.52}$$

Deriving the Variance-Covariance Matrix of n

In single-level data, vector $\mathbf{n}$ is assumed to follow a multinomial distribution with probability vector $\boldsymbol{\pi}$. If multiple ratings of the same subject are present, the variance in the data will be larger than expected under a multinomial distribution, because two sources of variation are present: the random fluctuation of the multinomial parameters across subjects and the variation of the raters within a subject (Vágó et al., 2011; Agresti, 2012, p. 7). If in two-level data a multinomial distribution is assumed for $\mathbf{n}$, this overdisperion is ignored, which results in too small standard errors (the naive standard errors in Table 3.2).

Suppose that for each subject $R_1 = R_2 = \ldots = R_S = R$, and probability vector $\boldsymbol{\pi}_s$ exists for subject s, with expectation $E(\boldsymbol{\pi}_s) = \boldsymbol{\pi}$. Then, for a given single subject, the conditional distribution of the vector with item-score patterns is multinomial with expectation $E(\mathbf{n}|\boldsymbol{\pi}) = R\boldsymbol{\pi}$ and variance-covariance matrix $V(\mathbf{n}|\boldsymbol{\pi}) = R(\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)$. The variance of the marginal distribution of $\mathbf{n}$ for a randomly selected subject is $E(V(\mathbf{n}|\boldsymbol{\pi})) + V(E(\mathbf{n}|\boldsymbol{\pi}))$ (Rice, 2006, p. 151, Theorem B). Because the subjects are assumed to be independent, the variance-covariance matrix of $\mathbf{n}$ for S subjects is defined as

$$\begin{aligned} \mathbf{V}_{\mathbf{n}} &= \textstyle\sum_{s=1}^{S} \left[E(V(\mathbf{n}|\boldsymbol{\pi})) + V(E(\mathbf{n}|\boldsymbol{\pi}))\right] \\ &= \textstyle\sum_{s=1}^{S} \left[E[R(\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)] + V(R\boldsymbol{\pi})\right] \\ &= SR[\text{Diag}(E(\boldsymbol{\pi})) - E(\boldsymbol{\pi}\boldsymbol{\pi}^T)] + SR^2[E(\boldsymbol{\pi}\boldsymbol{\pi}^T) - E(\boldsymbol{\pi})E(\boldsymbol{\pi})^T] \\ &= SR[\text{Diag}(E(\boldsymbol{\pi})) - E(\boldsymbol{\pi})E(\boldsymbol{\pi})^T] + SR(R-1)[E(\boldsymbol{\pi}\boldsymbol{\pi}^T) - E(\boldsymbol{\pi})E(\boldsymbol{\pi})^T] \end{aligned} \tag{3.53}$$

(see, e.g., Vágó et al., 2011; Rice, 2006, p. 140 Corollary B). If the number of raters R_s varies per subject, value R in Equation 3.53 can be replaced by the harmonic mean $\overline{R_s} = S/\sum_{s=1}^{R_s} R_s^{-1}$. For single-level scalability coefficients, there is only one replication per subject ($R = 1$), and the right-hand side of Equation 3.53 reduces to $S[\text{Diag}(E(\boldsymbol{\pi})) - E(\boldsymbol{\pi})E(\boldsymbol{\pi})^T]$, the well-known covariance matrix of the multinomial distribution with

parameters S and $E(\boldsymbol{\pi})$.

Estimating the Standard Errors

Applying the rules from Equation 3.17 to functions $\mathbf{g}(\mathbf{n}), \mathbf{g}^{\dagger}(\mathbf{n})$, and $\mathbf{g}^{\ddagger}(\mathbf{n})$ results in the Jacobian matrices $\mathbf{G}$, $\mathbf{G}^{\dagger}$, and $\mathbf{G}^{\ddagger}$, respectively. Because of its complexity and size, the Jacobian is not printed. The variance-covariance matrices of the coefficients are approximated by means of the delta method (Equation 3.18) as

$$\begin{array}{lcl} \mathbf{V}(\mathbf{H}_{ij}) & \approx & \mathbf{G}\ \ \mathbf{V_n}\ \mathbf{G}^{T} \\ \mathbf{V}(\mathbf{H}_{i}) & \approx & \mathbf{G}^{\dagger}\ \mathbf{V_n}\ \mathbf{G}^{\dagger T} \\ \mathbf{V}(\mathbf{H}) & \approx & \mathbf{G}^{\ddagger}\ \mathbf{V_n}\ \mathbf{G}^{\ddagger T}. \end{array} \tag{3.54}$$

The standard errors are retrieved by taking the square root of the diagonal of the variance-covariance matrices in Equation 3.54.

3.5.3 Asymptotic Distribution of Two-Level Scalability Coefficients

The distribution of the single-level coefficients for dichotomous items is asymptotically normal (Mokken, 1971, pp. 166-167, Theorem 3.3.1). The proof is based on the fact that the asymptotic distribution of linear functions of the vector with item-score pattern frequencies $\mathbf{n}$, in which the frequencies are considered as random variables, is normal (Rao, 1973, p. 383 (ii)). This proof is also valid for two-level case, because vector $\mathbf{n}$ is constructed from S independent subjects, each with finite expectation and variance. Therefore the multivariate central limit theorem applies (Rao, 1973, p. 128 (iv)), although it is necessary that the variance-covariance matrix is adjusted to account for overdispersion, which has been done in Section 3.5.2. If $\widehat{\mathbf{H}}_{(S)}$ is the vector of estimated two-level scalability coefficients for a random sample of S independent subjects, with expectation $\mathbf{H}$ and estimated variance-covariance matrix $\mathbf{V}(\widehat{\mathbf{H}}_{(S)})$, then for $S \to \infty$, $(\widehat{\mathbf{H}}_{(S)} - \mathbf{H}) \to N[0, \mathbf{V}(\widehat{\mathbf{H}}_{(S)})]$.

3.5.4 Performance for Simulated Data

In a small-scale simulation study, we investigated the sampling distribution of the two-level scalability coefficients and the coverage of the Wald-based confidence intervals. A normally distributed sampling distribution and a 95% coverage rate indicate that the standard errors are unbiased and accurate. The population was based on the real-data example and consisted of 100,000 subjects, each scored on seven 5-category items by 18 raters. The scores were generated by the hierarchical rater model (Patz et al., 2002). Model parameters were chosen such that the total-scale coefficients were similar to the values in the small real-data example. An overview of the data simulation method is provided in Appendix B. Because the asymptotic results are based on the number of subjects $S \to \infty$, it is expected that the results deteriorate as S decreases. We investigated

two levels of S that are relatively small: $S = 14$ (as in the real-data example) and $S = 50$. Both levels represent a relatively poor condition for obtaining unbiased and accurate standard error estimates. For both levels of S, 1,000 data sets were sampled from the population; for each sample, the two-level scalability coefficients and their standard errors were estimated. Due to limited space, the remaining variables were fixed.

Figure 3.1 shows the results. The sampling distribution of all coefficients was close to normal. For $S = 14$ subjects, on average H^W was slightly overestimated in the samples, H^B was correctly estimated, and H^{BW} was underestimated. For all three coefficients the standard errors were slightly smaller than the standard deviation of the sampling distribution. In addition, the coverage was slightly too low, with .95 falling outside the 95% confidence interval. For $S = 50$ subjects, the estimated coefficients and standard errors were close to the true values, and the 95% confidence intervals of the estimated coverages included .95. This simulation example demonstrates that even for limited sample sizes, the sampling distribution of the two-level scalability coefficients is close to normal and Wald-based intervals quickly give satisfactory coverage rates.

3.5.5 Computational Strategy

Computing the design matrices and the matrices of partial derivatives can be quite demanding, as more items are being used and more subjects are being scaled. For example, with 10 five-category items matrix $\mathbf{A}_1$ is of order $2,301 \times 9,765,625$. Two adjustments can be applied to reduce the computational burden substantially; using only nonzero frequencies in vector $\mathbf{n}$ and directly computing $\mathbf{g}_3$ and $\mathbf{G}_3$ from the data.

The length of vector $\mathbf{n}$ and the number of columns in the design matrix $\mathbf{A}_1$ and Jacobian matrices $\mathbf{G}_i$ is L, the number of all possible item-score patterns. Value L increases exponentially with the number of items. However, only observed patterns contribute to the computation of the scalability coefficients and the standard errors, and unobserved patterns may be removed from the vectors and matrices (see Kuijpers et al., 2013, p. 55 for proof). As a result, the number of observed item-score patterns L^* is at most the lesser of $(m + 1)^I$ and the number of subject-rater combinations $\sum_{s=1}^{S} R_s$.

Tedious but straightforward algebra shows that the result of $\mathbf{g}_3$ and its matrix of partial derivatives $\mathbf{G}_3$ can be computed directly from the data. This is convenient, because the order of matrix $\mathbf{A}_1$ to $\mathbf{A}_3$ and of the rows of $\mathbf{G}_1$ and $\mathbf{G}_2$ is a multiple of the number of bivariate response patterns B, which grows rapidly when more items or answer categories are used ($B = 1,125$ for 10 five-category items). The order of the remaining matrices does not exceed a multiple of the number of item-pairs K ($K = 45$ for 10 items), although the number of columns of the matrices $\mathbf{G}_i$ will always equal L^*. See Appendix C for direct computation of $\mathbf{g}_3$ and $\mathbf{G}_3$ from the data.

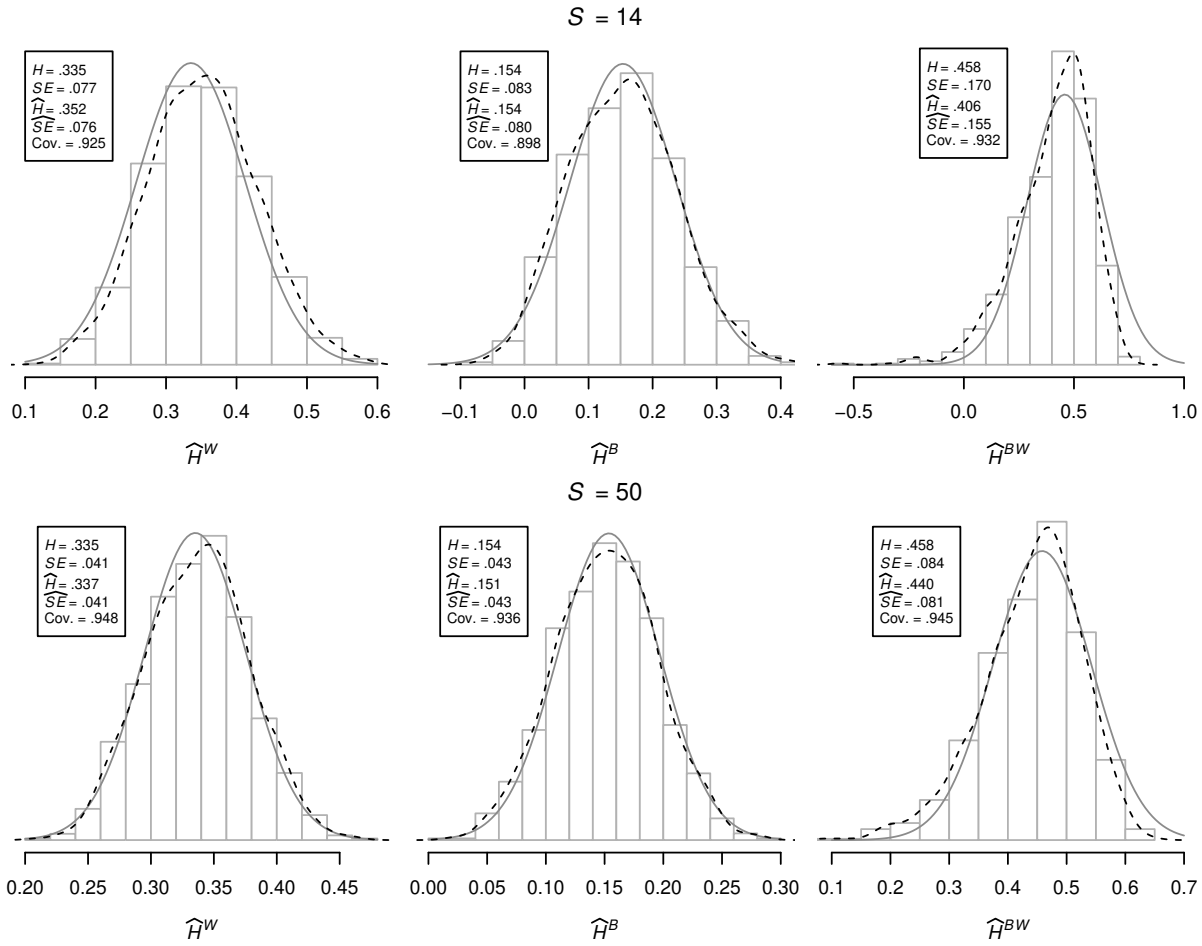

Figure 3.1: Plot of the sampling distribution of the two-level scalability coefficients for $S = 14$ (upper panel) and $S = 50$ (lower panel) subjects, based on 1,000 simulated data sets. The dashed black line is the kernel density of the sampling distribution and the solid gray line is the density of the normal distribution with population value H as mean and the standard deviation (SE) of the sampling distribution. Value $\widehat{H}$ is the average estimated coefficient and $\widehat{SE}$ the average estimated standard error across the simulated data sets. Coverage (Cov.) is the proportion of times the population H falls inside the 95% Wald-based confidence interval of the sample estimate.

3.5.6 Implementation in R

The estimation of the two-level scalability coefficients and their standard errors are available as function `MLcoefH()` in R (R Core Team, 2020) in the package `mokken` (Van der Ark, 2007, 2012). The argument of `MLcoefH()` is a data matrix with one subject column and a column per item. The function returns a list with three matrices, one for the item-pair, one for the item, and one for the total scale coefficients. These matrices contain the within, between, and ratio coefficients with their standard errors. The autonomy-support data example from this chapter can be obtained in R by the following command lines.

```
> # Load mokken package
> library(mokken)
> # Read data
> data(autonomySupport)
> # Scalability coefficients and standard errors
> MLcoefH(autonomySupport)
```

3.6 Discussion

We derived standard errors for two-level scalability coefficients (Snijders, 2001a; Crişan et al., 2016). As a result, the precision of estimated scalability coefficients can be determined, leading to more information with respect to the scalability of the items in the data. Estimation of both the two-level scalability coefficients and their standard errors is implemented as R-function `MLcoefH()` in the `mokken`-package. The computational shortcut has reduced the computation time considerably, but estimating standard errors can still be time consuming if the number of items and subjects is large.

The main reason to compute standard errors is confidence interval construction. We chose to use the Wald-based confidence interval, as the distribution of the two-level scalability coefficients is asymptotically normal. The simulation example demonstrated that even for a small number of subjects, the standard error estimates and coverage levels were close to the desired values. In addition, the sampling distribution of the two-level scalability coefficients was close to normal. Future research should focus on the bias and coverage of the two-level coefficients in a wider range of conditions, such as unequal group sizes and other values of the scalability coefficients. There may be situations where alternative intervals are preferred, such as bootstrap or profile likelihood confidence intervals.

With the derivation of the standard errors, the development of two-level Mokken scale analysis can continue. We intend to develop methods to determine how well the model fits the data. Also, generalization of the scalability coefficients and standard errors is required for situations where raters score multiple subjects. In addition, we plan to generalize the automated item selection procedure to accommodate two-level test data as well.