Bias in Point Estimates and Standard Errors of Mokken's Scalability Coefficients

Kuijpers, R.E.; van der Ark, L.A.; Croon, M.A.; Sijtsma, K.

DOI
10.1177/0146621616638500

Publication date
2016

Document Version
Final published version

Published in
Applied Psychological Measurement

Article

Bias in Point Estimates and Standard Errors of Mokken's Scalability Coefficients

Applied Psychological Measurement
2016, Vol. 40(5) 331–345

DOI: 10.1177/0146621616638500
apm.sagepub.com
SAGE

Renske E. Kuijpers[1], L. Andries van der Ark[2], Marcel A. Croon[3], and Klaas Sijtsma[3]

Abstract

Mokken scale analysis uses three types of scalability coefficients to assess the quality of (a) pairs of items, (b) individual items, and (c) an entire scale. Both the point estimates and the standard errors of the scalability coefficients assume that the sample ordering of the item steps is identical to the population ordering, but due to sampling error, the sample ordering may be incorrect and, consequently, the estimates and the standard errors may be biased. Two simulation studies were used to investigate the bias of the estimates and the standard errors of the scalability coefficients, as well as the coverage of the 95% confidence intervals. Distance between item steps was the most important design factor. In addition, sample size, number of items, number of answer categories, and item discrimination were included in the design. Bias of the standard errors was negligible. Bias of the estimates was largest when all item steps were identical in the population, especially for small sample sizes. Furthermore, bias of the estimates decreased as number of answer categories increased and as item discrimination decreased. Coverage of the 95% confidence intervals was close to .950, but for small sample size coverage deteriorated. Coverage also became poorer as number of items increased, in particular for dichotomous items.

Keywords

categorical analysis, Guttman scale, nonparametric item response theory, polytomous items, scale construction, standard errors, marginal models

Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002) is used to construct tests and questionnaires. Among other model assessment methods, Mokken scale analysis uses an automated item selection procedure to partition a set of items into one or more scales, such that the items in a particular scale measure a common trait using a reasonable level of discrimination power to be controlled by the researcher (Sijtsma & Molenaar, 2002, p. 68). The item selection

[1]Leiden University, The Netherlands
[2]University of Amsterdam, The Netherlands
[3]Tilburg University, The Netherlands

Corresponding Author:
Renske E. Kuijpers, Department of Methodology and Statistics, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands.
Email: r.e.kuijpers@fsw.leidenuniv.nl

is based on a nonparametric item response theory (NIRT) model known as the monotone homogeneity model (Sijtsma & Molenaar, 2002, Chapter 2). Mokken scale analysis is used to construct tests in various research areas such as psychology, for assessing psychological distress and well-being (Watson, Wang, Thompson, & Meijer, 2014), depression and anxiety (Bech, Bille, Moller, Hellström, & Ostergaard, 2014), disability in activities of daily living (Kingston et al., 2012), learning disability (Murray & McKenzie, 2013), and sexual sadism (Nitschke, Osterheider, & Mokros, 2009).

Mokken scale analysis uses three types of scalability coefficients for assessing the quality of (a) item pairs, (b) individual items, and (c) a set of items. The item selection procedure uses the scalability coefficients as criteria for item set partitioning and as diagnostics for the strength of the scales. To compute a scalability coefficient, the sample ordering of the item steps (Molenaar, 1991) is needed. Because of sampling fluctuation, the sample ordering may be different from the population ordering, thus biasing the estimates of the scalability coefficients. The distortion may be more serious when distance between adjacent population item steps is small and sample size is small. Hence, scalability coefficients may either underestimate or overestimate their parameter values. For dichotomous items, based on statistical reasoning involving all the 2×2 tables, Sijtsma and Molenaar (2002, p. 56) suggested that bias is almost negligible for $N > 200$ when incidental pairs of item steps are close together, say, less than .02 units, and for $N > 400$ when many item steps are close together. From their discussion, it is clear that additional research may be needed to support accurate recommendations. Kuijpers, Van der Ark, and Croon (2013) analytically derived standard errors for each of the three scalability coefficients. The standard errors are based on the sample item step ordering, and a sample ordering different from the population ordering may produce positively or negatively biased standard error estimates. Consequently, confidence intervals may have an incorrect coverage.

Simulation studies were used to assess the magnitude of the bias in the scalability coefficient estimates, the standard error estimates, and the coverage of the confidence intervals. Because it is expected that a smaller distance between adjacent population item steps produces more reversals in the sample item step ordering, the authors investigated the effect of differences between sample and population item step orderings on the estimates of the scalability coefficients and their standard errors. Bias of the estimates and the standard errors, and the coverage of the confidence intervals were assessed under several conditions. The most important design factor was distance between population item steps; a smaller distance was expected to increase the probability that the sample and population item step orderings differ from each other. Other design factors were sample size, number of items, number of answer categories, and item discrimination.

This article is organized as follows. First, the authors discuss Mokken scale analysis and the scalability coefficients. Second, they explain the computation of the standard errors by means of marginal modeling. Third, they discuss Simulation Study 1 and fourth, a follow-up Simulation Study 2 that investigates surprising results from Study 1. Finally, they provide recommendations on how to use the standard errors.

Mokken Scale Analysis

The Monotone Homogeneity Model

Mokken scale analysis is based on the monotone homogeneity model (Mokken, 1971, Chapter 4; Sijtsma & Molenaar, 2002, pp. 22-23), which is an NIRT model for measuring respondents on an ordinal scale. Let θ denote the latent variable that underlies performance on the J items in the test. For dichotomous items, the monotone homogeneity model implies the stochastic ordering of θ by means of total score X_+, which is the sum of the J item scores, denoted X_j with

$j = 1, \ldots, J$, so that $X_+ = \sum_{j=1}^{J} X_j$. For polytomous items, the monotone homogeneity model implies a weaker stochastic ordering property; for details, see Van der Ark and Bergsma (2010). If the monotone homogeneity model fits the data well, the stochastic ordering properties can be used for ordering respondents on latent variable θ by means of total score X_+.

Fit of the monotone homogeneity model to the data is investigated by checking whether several of the model's manifest properties are satisfied in the data. For example, the model implies that all interitem covariances are nonnegative in the population; hence, for a set of items to constitute a scale, the interitem covariances must be nonnegative. If not, the monotone homogeneity model is not the model that generated the data and must be rejected as an explanatory model. Nonnegativity of interitem covariances is investigated by evaluating whether the sample values of the $J(J-1)/2$ item pair scalability coefficients H_{ij} are nonnegative. The automated item selection procedure rejects item pairs having negative interitem covariances as candidates for admittance to a scale. When a researcher assembles his own scale(s), due to sampling fluctuation, sample H_{ij} values may be negative, and coefficients' standard errors should be used to avoid the wrong conclusion.

Scalability Coefficients

Item steps and weighted Guttman errors. The scalability coefficients are based on the common item step ordering in each pair of items and the weighted sum of Guttman errors that is based on the item step ordering (Molenaar, 1991; also see Kuijpers et al., 2013). A single item j having $z + 1$ ordered answer categories has z ordered item steps: $X_j \geq 1, X_j \geq 2, \ldots, X_j \geq z$. It is assumed that this ordering is the same for each respondent. A score x on item j can be considered to be the result of passing the most popular item steps, $X_j \geq 1, \ldots, X_j \geq x$, and failing the next, less popular item steps, $X_j \geq x+1, \ldots, X_j \geq z$. Let Y_j^x be a binary variable, with value 1 if the respondent has passed item step $X_j \geq x$ and value 0 if the respondent failed item step $X_j < x$; then, $X_j = \sum_{u=1}^{x} Y_j^u$.

Two items i and j together have $2z$ item steps; the ordering of these $2z$ item steps is needed for estimating item pair coefficient H_{ij}. To order the $2z$ item steps, one uses the z probabilities that a randomly chosen respondent passes an item step of item i, $P(X_i \geq x)$, and similarly the z probabilities for item j, $P(X_j \geq x)$. For item score $X_j \geq 0$, by definition we have $P(X_j \geq 0) = 1$, and this option is ignored. If in a particular item a less popular step is passed, by definition the more popular step is also passed.

Different respondents may pass and fail item steps in an order that is inconsistent with the common item step ordering for the two items, so that some individuals pass a less popular item step while failing a more popular item step. This incidence is referred to as a Guttman error (Guttman, 1950; Molenaar, 1991). Table 1 shows an example of the joint probabilities of having a score x on item a and a score y on item b, that is, $P(X_a = x, X_b = y)$ with $x, y = 0, 1, 2, 3$. The marginal probabilities are defined by $P(X_a = x)$ and $P(X_b = y)$, and the cumulative probabilities by $P(X_a \geq x)$ and $P(X_b \geq y)$. For this example, the cumulative probabilities order the item steps by descending popularity as

$$X_b \geq 1, X_a \geq 1, X_b \geq 2, X_b \geq 3, X_a \geq 2, X_a \geq 3. \quad (1)$$

Let index h enumerate the number of most popular item steps passed. Item-score patterns (0,0), (0,1), (1,1), (1,2), (1,3), (2,3), and (3,3) (in Table 1, corresponding probabilities are printed in boldface) are consistent with the Guttman (1950) model, because the h most popular item steps in Equation 1 were passed and the remaining $2z-h$ less popular steps were failed. The remaining item-score patterns are inconsistent with the Guttman model, and to arrive at any of these

Table 1. Cross-Tabulation of Probability of Obtaining Particular Item-Score Patterns.

	X_b					
X_a	0	1	2	3	$P(X_a = x)$	$P(X_a \geq x)$
0	**.044** (0)	**.013** (0)	.019 (1)	.025 (2)	.101	1.000
1	.023 (1)	**.060** (0)	**.106** (0)	**.267** (0)	.456	.899
2	.011 (4)	.028 (2)	.193 (1)	**.145** (0)	.377	.443
3	.002 (7)	.012 (4)	.042 (2)	**.010** (0)	.066	.066
$P(X_b = y)$	.080	.113	.360	.447	1.000	
$P(X_b \geq y)$	1.000	.920	.807	.447		

Note. Probabilities of item-score patterns that are in agreement with the Guttman model are printed in boldface. Guttman weights are shown within parentheses.

patterns, one or more Guttman errors are made. For example, someone who obtained item-score pattern (0,3), failed the more popular item step $X_a \geq 1$ but passed the less popular item steps $X_b \geq 2$ and $X_b \geq 3$.

Molenaar (1991) proposed weighing the sample frequencies of the Guttman errors (in Table 1, weights are shown within parentheses) depending on the degree to which the item step ordering was violated according to the Guttman model. The weight for a particular item-score pattern ($X_i = x$, $X_j = y$), denoted w_{ij}^{xy}, is equal to the number of item step pairs for which the less popular step is passed and the more popular step is failed; see Ligtvoet, Van der Ark, Te Marvelde, and Sijtsma (2010) and Kuijpers et al. (2013) for the computation of the weights. Because the weights play a crucial role in the potential bias in the scalability coefficients and the standard errors, the computation is reiterated here.

Consider indicator vector $\mathbf{q}_{ij}^{xy} = (q_{ij(1)}^{xy}, q_{ij(2)}^{xy}, \ldots, q_{ij(2z)}^{xy})$, whose elements correspond to the $2z$ ordered item steps of item pair (i, j) and assume elements to have value 1 if an item step was passed to obtain item-score pattern ($X_i = x$, $X_j = y$), and value 0 otherwise. Following Equation 1, the $2z$ item steps are ordered by descending popularity. Then, weight w_{ij}^{xy} equals

$$w_{ij}^{xy} = \sum_{u=2}^{2z} q_{ij(u)}^{xy} \left(\sum_{v=1}^{u-1} |1 - q_{ij(v)}^{xy}| \right). \tag{2}$$

For each pair of 0s and 1s, Equation 2 counts how often a score 0 precedes a score 1 in vector $\mathbf{q}_{ij}^{xy}$. For example, for item-score pattern (1,2), the first three item steps in Equation 1 were passed. These are the three most popular steps, implying $\mathbf{q}_{ab}^{12}$ = (1,1,1,0,0,0), and because 0 scores do not precede 1 scores, weight $w_{ab}^{12} = 0$. For item-score pattern (0,3), item steps $X_b \geq 1$, $X_b \geq 2$, and $X_b \geq 3$ were passed, so that vector $\mathbf{q}_{ab}^{03}$ = (1,0,1,1,0,0). In this example, a 0 score precedes a 1 score twice, and thus weight $w_{ab}^{03} = 2$.

Different random samples produce item step orderings different from the population ordering, resulting in sample weights different from population weights. For example, for two different random samples containing 200 observations each, drawn from the population values in Table 1, Table 2 shows the joint frequencies for the two samples. In the first sample (Table 2, upper panel), the estimated item step ordering is identical to the population item step ordering (Table 1). As the estimated ordering is identical to the population ordering, the sample weights equal the population weights. In the second sample (Table 2, lower panel), the estimated item step ordering and the corresponding weights are different from the population values. Using weights different from population weights may result in biased estimates and standard errors.

Table 2. Frequency Tables for Two Samples (N = 200) Drawn From the Distribution in Table 1.

	X_b					
X_a	0	1	2	3	Frequency	$\hat{P}(X_a \geq x)$
0	**13** (0)	**1** (0)	2 (1)	4 (2)	20	1.000
1	2 (1)	**10** (0)	**20** (0)	**64** (0)	96	.900
2	2 (4)	2 (2)	40 (1)	**30** (0)	74	.420
3	0 (7)	3 (4)	6 (2)	**1** (0)	10	.050
Frequency	17	16	68	99	200	
$\hat{P}(X_b \geq y)$	1.000	.915	.835	.495		
0	**8** (0)	**1** (0)	6 (1)	4 (3)	19	1.000
1	6 (1)	**12** (0)	**24** (0)	51 (1)	93	.905
2	3 (3)	7 (1)	**44** (0)	**26** (0)	80	.440
3	0 (6)	2 (3)	5 (1)	**1** (0)	8	.040
Frequency	17	22	79	82	200	
$\hat{P}(X_b \geq y)$	1.000	.915	.805	.410		

Note. In Sample 1 (upper panel), the estimated item step ordering is identical to the population item step ordering. In Sample 2 (lower panel), the estimated item step ordering is different from the population ordering, resulting in different Guttman weights. Probabilities of item-score patterns that are in agreement with the Guttman model are printed in boldface. Guttman weights are shown within parentheses.

Molenaar (1991) showed that when two item steps have equal popularities, scalability coefficients have the same value irrespective of the sample ordering of the two item step popularities. This implies that whenever the item step ordering contains ties, the scalability coefficient has the same value irrespective of the item step popularity that occurs first in the ordering.

Scalability coefficients and their standard errors. For item pair (i, j), scalability coefficient H_{ij} expresses the strength of the association between items i and j corrected for the marginal distributions of their item scores (Van der Ark, Croon, & Sijtsma, 2008a, 2008b). Coefficient H_{ij} compares the sum of weighted observed Guttman errors for item pair (i, j), denoted by F_{ij}, with the sum of weighted Guttman errors expected given marginal independence, denoted by E_{ij}, and subtracts the ratio F_{ij}/E_{ij} from 1:

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} n_{ij}^{xy}}{\sum_x \sum_y w_{ij}^{xy} e_{ij}^{xy}}. \quad (3)$$

Here, n_{ij}^{xy} denotes the observed frequency of Guttman errors, and e_{ij}^{xy} denotes the corresponding expected bivariate frequency given marginal independence. Given the monotone homogeneity model, in the population $0 \leq H_{ij} \leq 1$ (Mokken, 1971, pp. 148–153; Sijtsma & Molenaar, 2002, p. 59).

Item scalability coefficient H_j expresses the strength of the association between item j and the other items in a test (Sijtsma & Molenaar, 2002, p. 36) by combining the information from the $J-1$ H_{ij}s ($i \neq j$) in which item j is involved. For item j, coefficient H_j compares the sum of weighted observed Guttman errors with the sum of weighted expected Guttman errors:

$$H_j = 1 - \frac{\sum_{i \neq j} F_{ij}}{\sum_{i \neq j} E_{ij}} = 1 - \frac{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} n_{ij}^{xy}}{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} e_{ij}^{xy}}. \quad (4)$$

The monotone homogeneity model implies that $0 \leq H_j \leq 1$. Because H_j values close to 0 imply that item j is weakly associated with the other items and contributes little to a reliable person ordering, Mokken (1971, p. 184) proposed that for an item to be selected in a scale, $H_j \geq c > 0$. By default, $c = .3$, but the researcher may choose positive lower bound c so as to control the quality of the scale (Sijtsma & Molenaar, 2002, p. 60). Consequently, items with $H_j < c$ are left out of the scale (Sijtsma & Molenaar, 2002, p. 36).

Total scale scalability coefficient H expresses the degree to which respondents can be ordered by means of a set of items (Sijtsma & Molenaar, 2002, p. 39), and is a weighted average of the J H_j coefficients (Mokken & Lewis, 1982). Coefficient H compares the sum of weighted observed Guttman errors with the sum of weighted expected Guttman errors, and is defined as

$$H = 1 - \frac{\sum \sum_{i \neq j} F_{ij}}{\sum \sum_{i \neq j} E_{ij}} = 1 - \frac{\sum \sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} n_{ij}^{xy}}{\sum \sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} e_{ij}^{xy}}. \quad (5)$$

The monotone homogeneity model implies that $0 \leq H \leq 1$. Mokken (1971, p. 185) proposed that for a sufficiently reliable person ordering, $.3 \leq H \leq 1$. Hence, an item set for which $H < .3$ does not define a scale. Furthermore, a scale is defined to be weak if $.3 \leq H < .4$, moderate if $.4 \leq H < .5$, and strong if $H \geq .5$. In the absence of Guttman errors, H_{ij}, H_j, and H equal 1, and their values decrease as the number of Guttman errors increases.

Biased H_{ij}, H_j, and H coefficients may influence the composition of a Mokken scale. When H_{ij} or H_j is underestimated, a sufficiently discriminating item may incorrectly be left out of a scale, and when the coefficients are overestimated, weakly discriminating items may incorrectly be included in a scale. A biased H provides an incorrect assessment of the strength of a scale. Hence, biased estimates and biased standard errors should be avoided.

Kuijpers et al. (2013) used a two-step method based on categorical marginal models to derive asymptotic standard errors for each of the three scalability coefficients. First, data were collected in a frequency vector $\mathbf{n}$, in which the number of elements is equal to the number of item-score patterns in the data. Under the nonrestrictive assumption that $\mathbf{n}$ follows a multinomial distribution, the variance–covariance matrix of $\mathbf{n}$, denoted $\mathbf{V_n}$, is well known (e.g., Agresti, 2013). Second, each of the three scalability coefficients was written as a vector function of $\mathbf{n}$, denoted $\mathbf{g(n)}$. Let $\mathbf{G(n)}$ be the matrix of first partial derivatives of $\mathbf{g(n)}$ to $\mathbf{n}$, then according to the delta method, the variance–covariance matrix of the scalability coefficients, denoted $\mathbf{V_{g(n)}}$, is approximated by $\mathbf{G(n)V_nG(n)}^T$. The standard errors of the scalability coefficients are obtained by taking the square root of the diagonal elements of $\mathbf{G(n)V_nG(n)}^T$. The derivation of $\mathbf{g(n)}$ and $\mathbf{G(n)}$ is cumbersome (for more details, see Kuijpers et al., 2013).

Simulation Study I

Scalability coefficients were computed using the sample item step ordering. However, due to sampling fluctuation, the ordering may be different from the population ordering, thus affecting the estimates and the standard errors. For small sample size and small distance between item steps, more reversals of item step pairs are expected to occur. A simulation study was conducted

Table 3. Location Parameters δ_{jx} in Equation 6.

			Distance between item steps							
z + 1	J	j	None		Small		Moderate		Large	
2	2	1	0.000		−0.113		−0.227		−0.343	
		2	0.000		0.113		0.227		0.343	
	3	1	0.000		−0.227		−0.460		−0.706	
		2	0.000		0.000		0.000		0.000	
		3	0.000		0.227		0.460		0.706	
3	2	1	−0.250	0.250	−0.343	0.113	−0.706	0.227	−1.119	0.343
		2	−0.250	0.250	−0.113	0.343	−0.227	0.706	−0.343	1.119
	3	1	−0.250	0.250	−0.581	0.113	−1.278	0.227	−2.563	0.343
		2	−0.250	0.250	−0.343	0.343	−0.706	0.706	−1.119	1.119
		3	−0.250	0.250	−0.113	0.581	−0.227	1.278	−0.343	2.563

Note. For each distance between consecutive item steps and for items, the table shows δ_{j1} (upper panel), and δ_{j1} and δ_{j2} (lower panel). $z + 1 =$ number of answer categories; J = number of items; j = item index.

to investigate the effects of different factors on the bias of $\hat{H}$ and the bias of its standard errors, and the coverage of the 95% confidence intervals.

Method

Simulation model. We simulated data using the graded response model (Samejima, 1969, 1972). This model is a parametric version and hence a special case of the monotone homogeneity model (Hemker, Sijtsma, Molenaar, & Junker, 1996). The graded response model defines the probabilities of scoring at least x, $x = 0, 1, \ldots, z$, on item j by means of a logistic function with a discrimination parameter α_j and z location parameters δ_{jx}. For one item, the location parameters are ordered such that $\delta_{jx} < \delta_{j,x+1}$. In the simulation model, discrimination parameters within an item were kept constant, and the probability of a score of at least x on item j equals

$$P(X_j \geq x|\theta) = \frac{\exp\left[\alpha\left(\theta - \delta_{jx}\right)\right]}{1 + \exp\left[\alpha\left(\theta - \delta_{jx}\right)\right]}. \quad (6)$$

By definition, $P(X_j \geq 0|\theta) = 1$.

The values of discrimination parameter α were varied such that, in combination with $\theta \sim N(0, 1)$ and known location parameters, population values of H either had acceptable values $H \geq c = .3$ or unacceptable values $H < c = .3$. Location parameters δ_{jx} (Table 3) varied across design cells depending on the level of design factor ''Distance between item steps.'' For each sample of size N, N θ values were randomly drawn from a normal distribution. For each set of θ values, and for each value of α, a data set was generated using Equation 6 in which the δ_{jx} values (Table 3) were inserted.

Design. The design factors were varied as follows.

Discrimination parameter (α). Discrimination parameters equaled 1, 1.5, or 2. Keeping all other factors constant, item discrimination has a positive effect on the scalability coefficients (e.g., Sijtsma, 1988, Chapter 3). The effect of item discrimination on the bias in point estimates and standard errors was unknown.

Number of items (J). Number of items equaled 2 or 3; J was small so as to keep the simulation study manageable. Small J does not limit the results, because H is a weighted mean of the

Table 4. Theoretical Cumulative Item Step Probabilities $P(X_j \geq x)$ in Equation 7.

			Distance between item steps							
$z + 1$	J	j	None		Small		Moderate		Large	
2	2	1	.500		.530		.560		.590	
		2	.500		.470		.440		.410	
	3	1	.500		.560		.620		.680	
		2	.500		.500		.500		.500	
		3	.500		.440		.380		.320	
3	2	1	.566	.434	.590	.470	.680	.440	.770	.410
		2	.566	.434	.530	.410	.560	.320	.590	.230
	3	1	.566	.434	.650	.470	.800	.440	.950	.410
		2	.566	.434	.590	.410	.680	.320	.770	.230
		3	.566	.434	.530	.350	.560	.200	.590	.050

Note. For each distance between consecutive item steps and for items, the table shows $P(X_j \geq 1)$ (upper panel), and $P(X_j \geq 1)$ and $P(X_j \geq 2)$ (lower panel). $z + 1$ = number of answer categories; J = number of items; j = item index.

pairwise $J(J-1)/2$ H_{ij} coefficients, and therefore, we expected bias in the estimates and the standard errors and the coverage of the 95% confidence intervals to stay equal irrespective of the number of items.

Number of answer categories ($z + 1$). Items were dichotomous ($z + 1 = 2$) or polytomous ($z + 1 = 3$). Polytomous items are expected to produce more errors in the sample ordering and thus to produce more bias in the estimates of the scalability coefficients and their standard errors, and a poorer coverage of the 95% confidence interval.

Sample size (N). Sample size was small ($N = 50$), medium ($N = 200$), large ($N = 500$), or very large ($N = 1{,}500$). As N grows smaller, additional observations in the error cells have more influence on the sample item step ordering, and more likely produce stronger bias in the point estimates and the standard errors, and more strongly deteriorate the coverage of the 95% confidence interval.

Distance between item steps. The greater the distance between two adjacent item steps, the more likely the sample item step ordering is correct. Distance between item steps had four levels, labeled *None, Small, Moderate*, and *Large*. Distance was varied by manipulating the location parameters δ_{jx} of the graded response model. The ordering of the item steps was fixed to $P(X_1 \geq 1) > P(X_2 \geq 1) > \cdots > P(X_J \geq 1) > P(X_1 \geq 2) > \cdots > P(X_J \geq 2) > \cdots > P(X_1 \geq z) > \cdots > P(X_J \geq z)$. For this ordering, the distance between two consecutive item step probabilities is denoted by Δ, which equaled 0 (None), .06 (Small), .12 (Moderate), and .18 (Large). Table 4 shows the resulting cumulative item step probabilities. Once item step probabilities were fixed, we determined the corresponding location parameters δ_{jx}, such that

$$P(X_j \geq x) = \int P(X_j \geq x|\theta)\, \mathrm{d}G(\theta) \tag{7}$$

equaled the desired values in Table 4 for cumulative distribution $G(\theta)$. Because a smaller Δ value produces a smaller distance between population item step popularities, more reversals of the item step ordering are expected to occur in the sample. Consequently, we expected greater bias in the estimates and the standard errors of the scalability coefficients and a poorer coverage of the 95% confidence interval.

Table 5. Simulation Study 1: Population Values for Coefficient H, for All α and All Distances Between Item Steps.

		α = 1				α = 1.5				α = 2			
z + 1	*J*	N	S	M	L	N	S	M	L	N	S	M	L
2	2	.174	.190	.207	.225	.293	.329	.366	.404	.394	.449	.504	.558
	3	.174	.195	.217	.240	.293	.340	.386	.431	.394	.465	.531	.592
3	2	.190	.197	.220	.240	.327	.344	.388	.425	.444	.470	.531	.581
	3	.190	.207	.236	.258	.327	.363	.415	.451	.444	.496	.566	.612

Note. α = discrimination parameter; z + 1 = number of answer categories; *J* = number of items; N = no distance between item steps (Δ = 0); S = small distance between item steps (Δ = .06); M = moderate distance between item steps (Δ = .12), L = large distance between item steps (Δ = .18).

Outcome variables. The outcome variables were bias of the estimates of scalability coefficient *H*, bias of the standard errors of $\hat{H}$, and the coverage of the 95% confidence interval. The number of replications, *Q*, for each design cell was 10,000.

*Bias of the estimates (*bias*).* Let $\hat{H}_q$ denote the sample value of *H* in the *q*th replication (*q* = 1, . . . , *Q*), and let *H* denote the parameter, which was computed directly from the item step probabilities using linear programming. Bias, based on *Q* replications, was

$$bias = \frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{H}_q - H\right). \tag{8}$$

*Bias of the standard errors (*bias.se*).* The authors first computed the standard deviation of the estimates of *H*, denoted $sd(\hat{H})$, across the *Q* replications. Let $\bar{H} = [1/(Q-1)]\sum_{q=1}^{Q}\hat{H}_q$, then

$$sd\left(\hat{H}\right) = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}\left(\hat{H}_q - \bar{H}\right)^2}. \tag{9}$$

Standard deviation $sd(\hat{H})$ estimates the variability of $\hat{H}$ across replications and serves as a gold standard for the standard error. Let $se(\hat{H}_q)$ denotes the estimated standard error of the *q*th estimate of *H*. Then, the bias of the standard errors equals

$$bias.se = \frac{1}{Q}\sum_{q=1}^{Q}[se\left(\hat{H}_q\right) - sd\left(\hat{H}\right)]. \tag{10}$$

Coverage of the 95% confidence interval. The authors first constructed a confidence interval for each *q*th replication, using $\hat{H}_q \pm 1.96 \times se(\hat{H}_q)$. Then, the coverage was defined by the proportion of replications for which the 95% confidence interval contains the population value of *H*.

Table 5 shows parameter *H*, which was varied across design cells. Sample size does not affect parameter *H*. The simulation study was programmed in R (R Core Team, 2014), using the R-package *mokken* (Van der Ark, 2007, 2012) to compute $\hat{H}$ and the standard error of $\hat{H}$ for each sample across the 10,000 replications.

Results

The bias of $\hat{H}$ was less than .05 in all conditions (Figure 1). Compared with the other item step distances, for Δ = 0, the bias of $\hat{H}$ was slightly larger for both *J* = 2 (left panel) and *J* = 3 (right

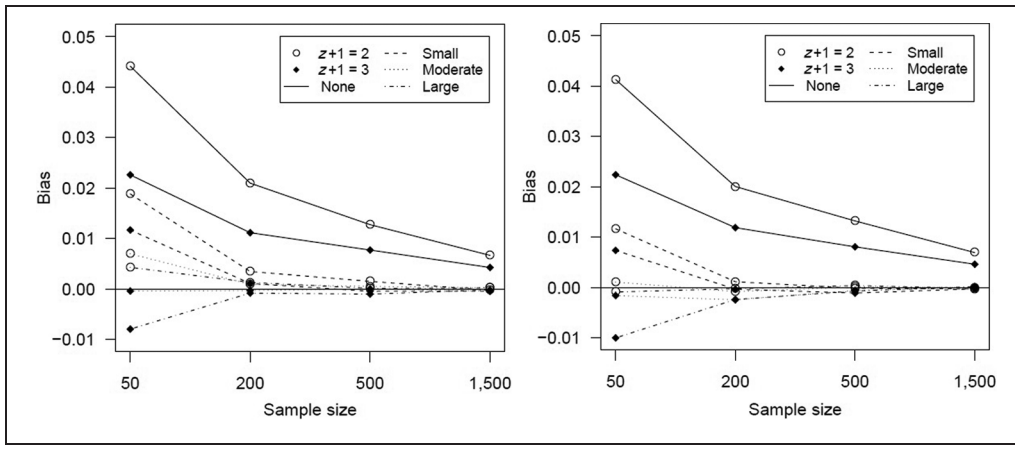

Figure 1. Bias in sample $\hat{H}$ for $J = 2$ (left panel) and $J = 3$ (right panel), $\alpha = 1.5$ for the four distances between item steps.

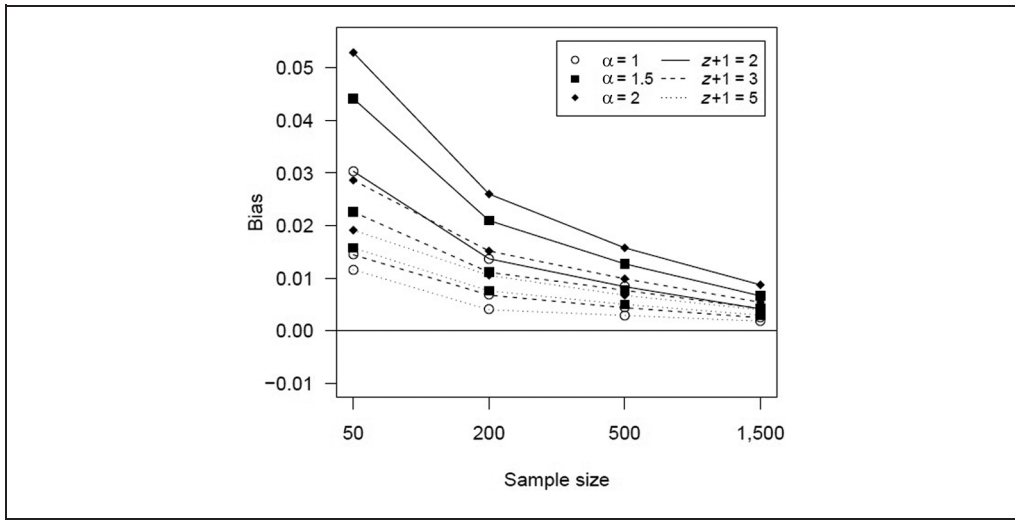

Figure 2. Bias in sample $\hat{H}$ for $J = 2$ for varying α and varying number of answer categories.

panel). As expected, an increase of number of items J did not affect bias of $\hat{H}$. For all four distances between item steps Δ, Figure 1 shows that the bias of $\hat{H}$ decreased as sample size N increased. For $\Delta = 0$ (None) and $\Delta = .6$ (Small), bias was considerably larger for $N = 50$ than for the other sample sizes. Also, an increase of item discrimination led to greater bias (Figure 2). Inconsistent with the expectation that bias increases as number of answer categories $z + 1$ increases, bias of $\hat{H}$ was larger for $z + 1 = 2$ than for $z + 1 = 3$.

For most conditions, the bias of $\hat{H}$ and the bias of the standard errors of $\hat{H}$ equaled 0 or nearly 0. The conditions for $\Delta = 0$ showed the largest bias and the poorest coverage; for $\Delta = 1.5$, Table 6 shows the bias of $\hat{H}$ and of its standard errors and the coverage of the 95% confidence intervals. We predicted bias of the standard errors to increase as the number of

Table 6. Results of Simulation Study 1 for $\Delta = 0$ and $\Delta = 1.5$.

		$z + 1 = 2$			$z + 1 = 3$		
J	*N*	*bias*	*bias.se*	Cov.	*bias*	*bias.se*	Cov.
2	50	.044	.000	**.929**	.023	−.003	**.924**
	200	.021	.002	.947	.011	.000	**.945**
	500	.013	.002	.948	.008	.000	**.943**
	1,500	.007	.001	.952	.004	.000	.947
3	50	.041	−.002	**.927**	.022	−.001	**.932**
	200	.020	.001	**.943**	.012	.000	**.941**
	500	.013	.001	**.939**	.008	.000	**.941**
	1,500	.007	.001	**.943**	.005	.000	**.942**

Note. Coverage values outside the 95% Agresti–Coull interval [.946, .954] are printed in boldface. *J* = number of items; *N* = sample size; *z* + 1 = number of answer categories; *bias* = bias of estimates of *H*; *bias.se* = bias of standard errors of $\hat{H}$; Cov. = coverage of 95% confidence interval.

Table 7. Simulation Study 2: Population Values for Coefficient *H*.

	α		
z + 1	1	1.5	2
2	.174	.293	.394
3	.190	.327	.444
5	.212	.369	.502

Note. *z* + 1 = number of answer categories; α = discrimination parameter.

answer categories increased and sample size decreased, but in all design cells, for all values of α, bias was 0 or close to 0.

Coverage of the 95% confidence intervals was almost equal to .950 in all conditions. To accurately interpret the values of the coverage, a 95% Agresti–Coull confidence interval was derived (Agresti & Coull, 1998). The interval was equal to [.946, .954]. In some conditions, coverage was just below the Agresti–Coull interval, but we consider these deviations negligible. Only for $N = 50$, irrespective of the value of discrimination parameter α, coverage was substantively smaller than expected.

Simulation Study 2

Study 1 showed that, compared with $J = 2$, bias was unaffected for $J = 3$, but these small test lengths seemed insufficient for ruling out bias effects for larger sets of items. Study 1 also showed that, inconsistent with the expectation, bias of $\hat{H}$ decreased as number of answer categories increased. Thus, for larger number of items ($J = 10$) and larger number of answer categories ($z + 1 = 5$), the authors investigated the bias of $\hat{H}$ and $sd(\hat{H})$, and the coverage of the 95% confidence interval.

Study 1 showed that for $\Delta > 0$, bias of $\hat{H}$ and the standard errors was negligible, and coverage was close to .950. Hence, Study 2 was done only for $\Delta = 0$. The design was similar to that of Study 1, but to keep the study manageable, design factors were not fully crossed. Table 7 shows the *H* parameters; note that parameter values were unaffected by number of items.

Table 8. Results of Simulation Study 2 for $\Delta = 0$; for $J = 10$ (Left Panel) and $z + 1 = 5$ (Right Panel).

		$J = 10$					$z + 1 = 5$		
$z + 1$	N	*bias*	*bias.se*	Cov.	J	N	*bias*	*bias.se*	Cov.
2	50	.040	−.001	**.901**	2	50	.016	−.003	**.920**
	200	.020	.000	**.902**		200	.008	.000	**.939**
	500	.013	.000	**.893**		500	.005	.000	**.944**
	1,500	.007	.000	**.896**		1,500	.003	.000	**.944**
3	50	.023	−.002	**.921**	3	50	.013	−.003	**.928**
	200	.012	.000	**.927**		200	.007	.000	**.944**
	500	.008	.000	**.929**		500	.005	.000	.946
	1,500	.005	.000	**.933**		1,500	.003	.000	**.944**
					10	50	.016	−.001	**.931**
						200	.007	.000	**.937**
						500	.005	.000	**.937**
						1,500	.003	.000	**.941**

Note. Coverage values outside the 95% Agresti–Coull interval [.946, .954] are printed in boldface. J = number of items; $z + 1$ = number of answer categories; N = sample size; *bias* = bias of estimates of H; *bias.se* = bias of standard errors of $\hat{H}$; Cov. = coverage of 95% confidence interval.

Table 8 shows results for $\alpha = 1.5$. Independent of α, bias of $\hat{H}$ and bias of the standard errors of $\hat{H}$ were unaffected as number of items J increased from three to 10; bias found in Study 2 was comparable with bias found in Study 1. When number of answer categories increased from three to five, bias of $\hat{H}$ decreased, especially for $N = 50$. This outcome again contradicted the expectation that bias increases as number of answer categories increases, and was consistent for each α value.

Figure 3 shows the coverage for $J = 10$ for varying α. Coverage of the 95% confidence interval was substantially worse for $J = 10$ than for $J = 3$. None of the values lay in the Agresti–Coull interval. These results contradict the expectation that the coverage remains the same as number of items increases. Compared with polytomous items, coverage of the 95% confidence interval was worse for dichotomous items. Hence, contrary to the expectation, coverage improved as number of answer categories increased. Consistent with Study 1, coverage was considerably better for larger N than for $N = 50$. However, in contrast to Study 1, in Study 2 item discrimination α did affect coverage; as α decreased, coverage improved.

Discussion

The estimates and the standard errors of Mokken's scalability coefficients are based on the assumption that the sample item step ordering is identical to the population ordering. A violation of this assumption may bias the estimates and standard errors of scalability coefficients and coverage of the corresponding confidence intervals may be incorrect. The two simulation studies showed that bias of $\hat{H}$ was negligible, suggesting that the heuristic guidelines the authors discussed in the introduction (Sijtsma & Molenaar, 2002, p. 56) may have been too strict. Only if item steps are identical or sample size is small ($N < 200$), one may expect a small positive bias. Straat, Van der Ark, and Sijtsma (2014) recommended that for item selection, samples should at least have a size between 250 and 500 when item quality is high, and between 1,250 and 1,750 when item quality is low. For these sample sizes, the authors found that bias of $\hat{H}$ was negligible; hence, the marginal modeling approach may be accurate.

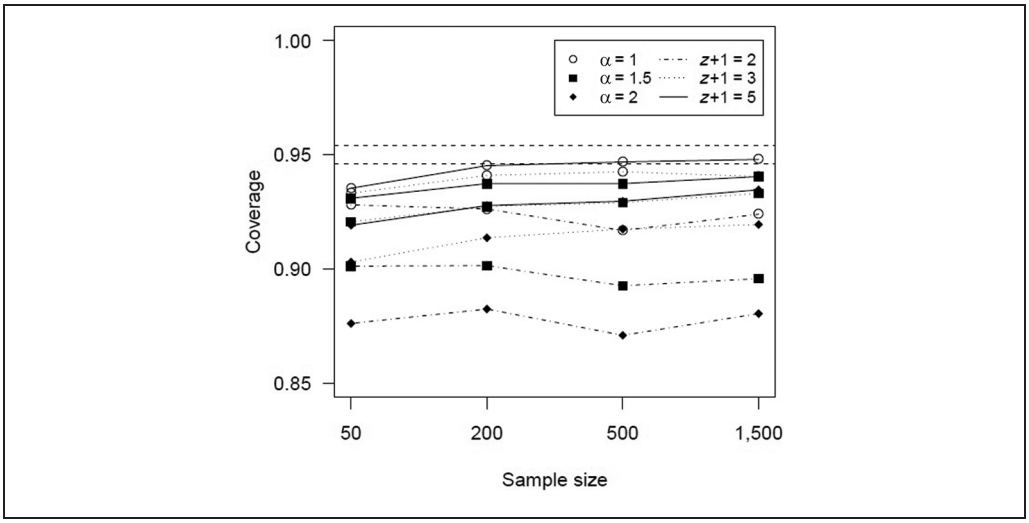

Figure 3. Coverage of 95% confidence intervals for $J = 10$ for varying number of answer categories.

Inconsistent with the expectations, bias of $\hat{H}$ decreased as number of answer categories increased. The decrease of bias was persistent when number of answer categories was raised to five (Study 2). Possibly, compared with a small number of item steps, a larger number of answer categories and item steps causes a reversal of adjacent item steps to have a smaller influence on bias. For all other conditions in Study 2, bias results were comparable with those in Study 1. Bias of the standard errors of $\hat{H}$ was negligible; hence, categorical marginal modeling is accurate for deriving standard errors of scalability coefficients. The availability of the standard errors in the R-package *mokken* renders them readily accessible.

For most conditions, coverage of the 95% confidence intervals was slightly under .950. For small N, large J, and high item discrimination α, coverage was slightly poorer. For dichotomous items, coverage dropped under 90% for large J, especially when α was high. Hence, for large J, point estimates and standard errors were unbiased, but coverage of the confidence intervals based on the point estimates and standard errors was poor. Because for correct coverage Wald-based confidence intervals require a symmetric distribution, this unexpected result may have been caused by a skewed $\hat{H}$ distribution, for which some evidence was found in a post hoc analysis. For the design cell producing the worst coverage, the distribution of $\hat{H}$ was positively skewed; skewness was computed using the R-package *e1071* (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2014) and equaled .144. Skewness was approximately 0 for design cells that resulted in a correct coverage. The Kolmogorov–Smirnov test was significant in all cases, suggesting for large J that the distribution of $\hat{H}$ deviates from standard normality, possibly affecting coverage of the 95% confidence intervals.

The coverage of 95% confidence intervals, even if not perfect, may be adequate for practical use, but may be improved if asymmetric confidence intervals are used. The Wald-based 95% confidence interval used in this study (i.e., $\hat{H}_q \pm 1.96 \times se(\hat{H}_q)$) is symmetric by definition, whereas confidence intervals such as likelihood profile confidence intervals or score confidence intervals (e.g., Lang, 2008), or bootstrap confidence intervals (e.g., Efron & Tibshirani, 1993) can be asymmetric and may improve coverage. This is a topic for further research.

The automated item selection procedure (Sijtsma & Molenaar, 2002, Chapter 4) in Mokken scale analysis only uses the sample scalability coefficients for selecting items into scales.

However, ignoring standard errors of sample coefficients may be a source of selection error (Kuijpers et al., 2013). Future research may systematically investigate the influence of standard errors on the automatic item selection procedure in Mokken scale analysis. A next step would be to implement the standard errors in the automatic item selection procedure.

Acknowledgment

The authors thank Rudy Ligtvoet and Iris A. M. Smits for their comments on an earlier draft of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research of Renske E. Kuijpers was funded by the Netherlands Organization of Scientific Research (NWO), Grant 406-12-013.

References

Agresti, A. (2013). *Categorical data analysis*. Hoboken, NJ: Wiley.

Agresti, A., & Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *American Statistician*, *52*, 119-126.

Bech, P., Bille, J., Moller, S. B., Hellström, L. C., & Ostergaard, S. D. (2014). Psychometric validation of the Hopkins Symptom Checklist (SCL-90) subscales for depression, anxiety, and interpersonal sensitivity. *Journal of Affective Disorders*, *160*, 98-103.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

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, studies in social psychology in World War II* (Vol. 4, pp. 60-90). Princeton, NJ: Princeton University Press.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*, 679-693.

Kingston, A., Collerton, J., Davies, K., Bond, J., Robinson, L., & Jagger, C. (2012). Losing the ability in activities of daily living in the oldest old: A hierarchic disability scale from the Newcastle 85+ study. *PLoS ONE*, *7*(2), e31665.

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

Lang, J. B. (2008). Score and profile likelihood confidence intervals for contingency table parameters. *Statistics in Medicine*, *27*, 5975-5990.

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

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2014). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien (R package version 1.6-3) [Computer software]. Retrieved from http://CRAN.R-project.org/package=e1071

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

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417-430.

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

Murray, A. L., & McKenzie, K. (2013). Estimating the severity of intellectual disability in adults: A Mokken scaling analysis of the Learning Disability Screening Questionnaire. *Psychological Assessment*, *25*, 1002-1006.

Nitschke, J., Osterheider, M., & Mokros, A. (2009). A cumulative scale of severe sexual sadism. *Sexual Abuse: A Journal of Research and Treatment*, *21*, 262-278.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Samejima, F. (1969). *Estimation of latent trait ability using a response pattern of graded scores* (Psychometrika Monograph, No. 17). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Samejima, F. (1972). *A general model for free-response data* (Psychometrika Monograph No. 18). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN18.pdf

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

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

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement*, *74*, 809-822.

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

Van der Ark, L. A. (2012). New developments in Mokken scale analysis. *Journal of Statistical Software*, *48*(5), 1-27.

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

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008a). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika*, *73*, 183-208.

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008b). Possibilities and challenges in Mokken scale analysis using marginal models. In K. Shigemasu, A. Okada, T. Imaizuma, & T. Hodhina (Eds.), *New trends in psychometrics* (pp. 525-532). Tokyo, Japan: Universal Academic Press.

Watson, R., Wang, W., Thompson, D. R., & Meijer, R. R. (2014). Investigating invariant item ordering in the Mental Health Inventory: An illustration of the use of different methods. *Personality and Individual Differences*, *66*, 74-78.