Bias of two-level scalability coefficients and their standard errors

Koopman, L.; Zijlstra, B.J.H.; de Rooij, M.; van der Ark, L.A.

DOI
10.1177/0146621619843821

Publication date
2020

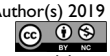
Document Version
Final published version

Published in
Applied Psychological Measurement

Article

Bias of Two-Level Scalability Coefficients and Their Standard Errors

Applied Psychological Measurement
2020, Vol. 44(3) 197–214



DOI: 10.1177/0146621619843821
journals.sagepub.com/home/apm


Letty Koopman[1], Bonne J. H. Zijlstra[1], Mark de Rooij[2] and L. Andries van der Ark[1]

Abstract

Two-level Mokken scale analysis is a generalization of Mokken scale analysis for multi-rater data. The bias of estimated scalability coefficients for two-level Mokken scale analysis, the bias of their estimated standard errors, and the coverage of the confidence intervals has been investigated, under various testing conditions. It was found that the estimated scalability coefficients were unbiased in all tested conditions. For estimating standard errors, the delta method and the cluster bootstrap were compared. The cluster bootstrap structurally underestimated the standard errors of the scalability coefficients, with low coverage values. Except for unequal numbers of raters across subjects and small sets of items, the delta method standard error estimates had negligible bias and good coverage. Post hoc simulations showed that the cluster bootstrap does not correctly reproduce the sampling distribution of the scalability coefficients, and an adapted procedure was suggested. In addition, the delta method standard errors can be slightly improved if the harmonic mean is used for unequal numbers of raters per subject rather than the arithmetic mean.

Keywords

cluster bootstrap, delta method, Mokken scale analysis, rater effects, standard errors, two-level scalability coefficients

In multi-rater assessments, multiple raters evaluate or score the attribute of subjects on a standardized questionnaire. For example, several assessors may assess teachers' teaching skills using a set of rubrics (e.g., Maulana, Helms-Lorenz, & Van de Grift, 2015; Van der Grift, 2007), both parents may rate their child's behavior using a health-related quality of life questionnaire (e.g., Ravens-Sieberer et al., 2014), and policy holders may evaluate the quality of health-care plans using several survey items (e.g., Reise, Meijer, Ainsworth, Morales, & Hays, 2006). In multi-rater assessments, raters (assessors, parents, policy holders) are nested within subjects (teachers, children, health-care plans). From this two-level data, measuring the

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

Corresponding Author:
Letty Koopman, Research Institute of Child Development and Education, University of Amsterdam, P. O. Box 15776, 1001 NG Amsterdam, The Netherlands.
Email: V.E.C.Koopman@UvA.nl

attribute (teaching skills, behavior, quality) of the subjects at Level 2 is of most interest. Because raters are the respondents, they may have a large effect on the responses to the items, which can interfere with measuring the subjects' attribute.

For dichotomous items, Snijders (2001) proposed two-level scalability coefficients to investigate the scalability of the items used in multi-rater assessments. These coefficients are generalizations of Mokken's (1971) single-level scalability coefficients (or H coefficients), which are useful as measures to assess whether ''the items have enough in common for the data to be explained by one underlying latent trait . . . in such a way that ordering the subject by the total score is meaningful'' (Sijtsma & Molenaar, 2002, p. 60). Mokken introduced scalability coefficients for each item-pair (H_{ij}), each item (H_i), and the total set of items (H). For multi-rater data, Snijders proposed extending the H_{ij}, H_i, and H coefficients to within-rater scalability coefficients (denoted by the superscript W), between-rater scalability coefficients (denoted by the superscript B), and the ratio of the between to within coefficients (denoted by the superscript BW).

The scalability coefficients are related to measurement models, in which subject and rater effects are jointly modeled (Snijders, 2001). A more detailed description of the measurement models and the two-level coefficients is provided below. Crisan, Van de Pol, and Van der Ark (2016) generalized the two-level scalability coefficients for dichotomous items to polytomous items, and Koopman, Zijlstra, and Van der Ark (in press) derived standard errors for the estimated two-level scalability coefficients using the delta method (e.g., Agresti, 2012, pp. 577-581; Sen & Singer, 1993, pp. 131-152). Alternatively, a cluster bootstrap may be used to estimate standard errors. The cluster bootstrap (Sherman & Le Cessie, 1997; see also Cheng, Yu, & Huang, 2013; Deen & De Rooij, in press; Field & Welsh, 2007; Harden, 2011) has not been applied to two-level scalability coefficients, but it has been applied in similar data structures—for example, children within county (Sherman & Le Cessie, 1997), siblings or genetic profiles within families (Bull, Darlington, Greenwood, & Shin, 2001; Watt, McConnachie, Upton, Emslie, & Hunt, 2000), repeated measurements of homeless people their housing status (De Rooij & Worku, 2012), or of children's microbial carriage (Lewnard et al., 2015).

For the two-level scalability coefficients, the problem at hand is that neither the bias of the point estimates nor the bias and accuracy of the standard errors have been thoroughly investigated. For the single-level scalability coefficients, the point estimates were mostly unbiased (Kuijpers, Van der Ark, Croon, & Sijtsma, 2016) and for both the analytically derived standard errors using the delta method (Kuijpers et al., 2016) and the bootstrap standard errors (Van Onna, 2004), the levels of bias and accuracy were satisfactory. However, these results cannot be generalized to two-level scalability coefficients because single-level coefficients do not take into account between-rater scalability, nor the dependency in the data due to the nesting of raters within subjects. The goal of this article is to investigate the bias of the point estimates and the standard errors of the two-level scalability coefficients. The remainder of this article first discusses two-level nonparametric item response theory (IRT) models, two-level scalability coefficients, and the two standard error estimation methods. Then, the article discusses the simulation study to investigate bias and coverage, and its results.

Nonparametric IRT Models for Two-Level Data

In multi-rater data, an attribute of subject s ($s = 1, \ldots, S$) is scored by R_s raters using I items. Raters are indexed by r or p ($r, p = 1, \ldots, R_s; r \neq p$), and items are indexed by i or j ($i, j = 1, \ldots, I; i \neq j$). Each item has $m + 1$ ordered response categories, indexed by x or y ($x, y = 0, 1, \ldots, m$). Let X_{sri} denote the score of subject s by rater r on item i. Typically, the

mean item score across raters, $\overline{X}_{s\cdot\cdot} = (IR_s)^{-1} \sum_{r=1}^{R_s} \sum_{i=1}^{I} X_{sri}$, is used as a measurement for the attribute of subject s.

In 2001, Snijders proposed a two-level nonparametric IRT model for two-level data, based on the monotone homogeneity model (Mokken, 1971; Sijtsma & Molenaar, 2002). Let θ_s be the value of subject s on a unidimensional latent trait θ that represents the attribute being measured, and δ_{sr} a deviation that consists of the effect of rater r and the interaction effect of rater r and subject s. Hence, $\theta_s + \delta_{sr}$ is the value of subject s on the latent trait according to rater r. It is assumed that, on average, the rater deviation for subject s equals zero ($E(\delta_{sr}) = 0$). In Snijders's model, the responses to the different items and subjects are assumed stochastically independent given the latent values θ_s and δ_{sr}. The probability that subject s obtains at least score x on item i when assessed by rater r, $P(X_{sri} \geq x|\theta_s, \delta_{sr})$, is monotone nondecreasing in $\theta_s + \delta_{sr}$. Because $E(\delta_{sr}) = 0$, the monotonicity assumption implies a nondecreasing item-step response function $P(X_{sri} \geq x|\theta_s)$, which is the expectation of $P(X_{sri} \geq x|\theta_s, \delta_{sr})$ with respect to the distribution of δ_{sr}.

An alternative generalization of the monotone homogeneity model for two-level data is the nonparametric hierarchical rater model. The hierarchical rater model (DeCarlo, Kim, & Johnson, 2011; Mariano & Junker, 2007; Patz, Junker, Johnson, & Mariano, 2002) is a two-stage model for multi-rater assessments in which a single performance is rated. Similar to Snijders's model, latent values θ_s and δ_{sr} are the subject's latent trait level and the rater's deviation, respectively. The hierarchical rater model assumes an unobserved ideal rating of the performance of subject s on each item i, denoted by ξ_{si}. The ideal ratings may vary across performances and are solely based on the subject's latent trait value. The ideal ratings to the different items are assumed stochastically independent given θ_s, and the item-step response function $P(\xi_{si} \geq x|\theta_s)$ is nondecreasing in θ_s. The observed item score X_{sri} is the rater's evaluation of ideal rating ξ_{si} (i.e., of the performance). For raters with negative δ_{sr}, the probability increases that X_{sri} is smaller than ξ_{si}, and for raters with positive δ_{sr}, the probability increases that X_{sri} is larger than ξ_{si}. Observed ratings X_{sri} are stochastically independent given ξ_{si} and δ_{sr} and the item-step response function $P(X_{sri} \geq x|\xi_{si}, \delta_{sr})$ is nondecreasing in $\xi_{si} + \delta_{sr}$.

Scalability Coefficients for Two-Level Data

Scalability coefficients evaluate the ordering of observed item responses. They are a function of the weighted item probabilities. These weights are explained briefly here (for more details, see Koopman, Zijlstra, & Van der Ark, 2017; Kuijpers, Van der Ark, & Croon, 2013), and illustrated in the appendix using a small data example. Let $P(X_{sri} = x, X_{srj} = y)$ denote the bivariate probability that rater r of subject s scores x on item i and y on item j. Let $P(X_{sri} = x, X_{spj} = y)$ ($p \neq r$) denote the bivariate probability that rater r of subject s scores x on item i and another rater (p) of the same subject scores y on item j. Let $P(X_i = x)$ be the probability that a certain rater scores x on item i for a certain subject.

Let $\mathbf{1}(\cdot)$ denote an indicator function, which takes value 1 if its argument is true and value 0 otherwise. Each item-score X_i has m item steps $Z_{ix} = \mathbf{1}(X_i \geq x)$ ($i = 1, 2, \ldots, I$; $x = 1, 2, \ldots, m$). An item step is passed if $Z_{ix} = 1$, and an item step is failed if $Z_{ix} = 0$. $P(X_i \geq x)$ is the popularity of item step Z_{ix}. Item steps of each item-pair are sorted in descending order of popularity. A Guttman error is defined as passing a less popular item step after a more popular item step has been failed. For instance, if for item-pair X_i, X_j the order of item steps is $Z_{i1}, Z_{j1}, Z_{j2}, Z_{i2}, Z_{i3}, Z_{j3}$ (i.e., $P(X_i \geq 1) \geq P(X_j \geq 1) \geq P(X_j \geq 2) \geq P(X_i \geq 2) \geq P(X_i \geq 3) \geq P(X_j \geq 3)$), then item-score pattern ($X_i = 0, X_j = 1$) is a Guttman error, because this item-score pattern requires that the second ordered item step $Z_{j1} = 1$ must be passed, whereas the first, easier step $Z_{i1} = 0$, is failed. Patterns that are not a Guttman error are referred to as consistent patterns. If a Guttman error is

observed within the same rater (i.e., $(X_{sri} = 0, X_{srj} = 1)$), this is referred to as a within-rater error. If a Guttman error is observed across two different raters of the same subject (i.e., $(X_{sri} = 0, X_{spj} = 1)$), this is referred to as a between-rater error. A Guttman error is considered more severe if more ordered steps have been failed before a less popular item step has been passed (e.g., $X_i = 0, X_j = 3$ is worse than $X_i = 0, X_j = 1$). The severity of the Guttman error for item-score pattern $(x, y) = (X_i = x, X_j = y)$ is indicated by weight w_{ij}^{xy}, which denotes the number of failed item steps preceding passed item steps (Molenaar, 1991). Let $z_h^{xy} \in \{0, 1\}$ denote the evaluation of the h-th ($1 \leq h \leq 2m$) ordered item step with respect to item-score pattern (x, y), then weight w_{ij}^{xy} is computed as

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

For consistent item-score patterns value w_{ij}^{xy} equals zero.

Let $F_{ij}^W = \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{srj} = y)$ be the sum of all weighted within-rater Guttman errors in item pair (i, j) and let $E_{ij} = \sum_x \sum_y w_{ij}^{xy} P(X_i = x) P(X_j = y)$ be the sum of all expected weighted Guttman errors in item pair (i, j) under marginal independence. The within-rater scalability coefficient H_{ij}^W for item-pair (i, j) is then defined as

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

Let $F_{ij}^B = \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{spj} = y), (p \neq r)$ be the sum of all weighted between-rater Guttman errors in item pair (i, j). Replacing F_{ij}^W with F_{ij}^B in Equation 2 results in the between-rater scalability coefficient

$$H_{ij}^B = 1 - \frac{F_{ij}^B}{E_{ij}}. \tag{3}$$

Dividing the two coefficients results in ratio coefficient $H_{ij}^{BW} = H_{ij}^B / H_{ij}^W$. Note that if $F_{ij}^W = F_{ij}^B$, then $H_{ij}^B = H_{ij}^W$ and $H_{ij}^{BW} = 1$. As for single-level scalability coefficients, the two-level scalability coefficients for items (H_i^W, H_i^B) are defined as $H_i = 1 - \sum_{j \neq i} F_{ij} / \sum_{j \neq i} E_{ij}$ and the two-level scalability coefficients for the total scale (H^W, H^B) are defined as $H = 1 - \sum_i \sum_{j>i} F_{ij} / \sum_i \sum_{j>i} E_{ij}$ (e.g., Crisan et al., 2016; Snijders, 2001). In samples, the scalability coefficients are estimated by using the sample proportions; for computational details, see Snijders (2001; also see Crisan et al., 2016; Koopman et al., 2017).

Within-rater coefficient H^W reflects the consistency of item-score patterns within raters, and its interpretation is similar to the single-level scalability coefficients of Mokken (1971). Between-rater coefficient H^B reflects the consistency of item-score patterns between raters of the same subject. The maximum value of within- and between-rater scalability coefficients equals 1, reflecting a perfect relation between the items, within and between raters of the same subject. Under the discussed IRT models, if the distribution of $\theta_s + \delta_{sr}$ is equally or more dispersed than the distribution of θ_s, $0 \leq H^B \leq H^W$ (Snijders, 2001). As the population of subject-rater combinations becomes more homogeneous (i.e., the variance of $\theta_s + \delta_{sr}$ becomes smaller), coefficient H^W decreases. Likewise, as the population of subjects becomes more homogeneous (i.e., the variance of θ_s becomes smaller), coefficient H^B decreases. Ratio coefficient H^{BW} provides useful information on the between- to within-rater variability: the larger the

variance of δ_{sr} (i.e., the rater effect) is compared to the variance of θ_s (i.e., the subject effect), the smaller the consistency of item-score patterns between raters of the same subject is relative to the consistency of item-score patterns within raters, and the smaller H^B is compared to H^W. As a result, H^{BW} decreases as the rater effect increases. For example, if H^{BW} is close to 1, the test score is hardly affected by the individual raters and only few raters per subject are necessary to scale the subjects, whereas if H^{BW} is close to 0, the raters almost entirely determine the item responses and scaling subjects is not sensible.

For a satisfactory scale, Snijders (2001) suggested heuristic criteria $H_{ij}^W \geq .1$, H_i^W and $H^W \geq .2$, $H_{ij}^B \geq 0$, and H_i^B and $H^B \geq .1$. In addition, he proposed that ratio value $H^{BW} \geq .3$ is reasonable and $H^{BW} \geq .6$ is excellent, with similar interpretations for H_{ij}^{BW} and H_i^{BW}. In single-level data, an often-used lower bound is .3 (Mokken, 1971, p. 185). Due to the availability of multiple parallel measurements per subject (i.e., multiple raters), the heuristics for two-level scalability coefficients are lower. The value of total-scale coefficients can be increased by removing items with low item scalability from the item set. In Mokken scale analysis for single-level data, there exists an item selection procedure based on single-level scalability coefficients, but this is not yet available for multi-rater data. In addition to Snijders's criteria, the authors suggest that the confidence intervals (CIs) of the H coefficients should be used in evaluating the quality of a scale. Kuijpers et al. (2013) advised comparing the CI with the heuristic criteria: For example, a scale can only be accepted as strong when the lower bound of the 95% CI is at least .5. A less conservative approach is to require the lower bound for all H coefficients to exceed zero. Items that fail to meet these criteria may be adjusted or removed from the item set.

Standard Error of Two-Level Scalability Coefficients

Analytical Standard Errors

The delta method approximates the variance of the transformation of a variable by using a first-order Taylor approximation (e.g., Agresti, 2012, pp. 577-581; Sen & Singer, 1993, pp. 131-152). Recently, Koopman et al. (in press) applied the delta method to derive standard errors for two-level scalability coefficients. Let $\mathbf{n}$ be a vector of order $(m+1)^I$ containing the frequencies of all possible item-score patterns, each pattern taking the form $n^{x_1 x_2 \ldots x_I}_{1\,2\,\ldots\,I}$. The patterns are ordered lexicographically with the last digit changing fastest, such that $\mathbf{n} = [n^{00\ldots0}_{12\ldots I} n^{00\ldots1}_{12\ldots I} \ldots n^{mm\ldots m}_{12\ldots I}]^T$. Vector $\mathbf{n}$ is assumed to be sampled from a multinomial distribution with varying multinomial parameters per subject (Vágó, Kemény, & Láng, 2011). Vector $\mathbf{p}_s$ contains the probabilities of obtaining the item-score patterns in vector $\mathbf{n}$ for subject s, with expectation $E(\mathbf{p})$ for a randomly selected subject. Suppose that for each subject $R_1 = R_2 = \ldots = R_S = R$. In addition, let $E(\mathbf{x})$ denote the expectation of vector $\mathbf{x}$, and $\mathrm{Diag}(\mathbf{x})$ a diagonal matrix with $\mathbf{x}$ on the diagonal. Then the variance-covariance matrix of $\mathbf{n}$ equals

$$\mathbf{V_n} = SR[\mathrm{Diag}(E(\mathbf{p})) - E(\mathbf{p})E(\mathbf{p})^T] + SR(R-1)[E(\mathbf{pp}^T) - E(\mathbf{p})E(\mathbf{p})^T] \quad (4)$$

(Koopman et al., in press; Vágó et al., 2011).

Let $\mathbf{g(n)}$ be the transformation of vector $\mathbf{n}$ to a vector containing the scalability coefficients $\mathbf{g(n)} = [H^B H^W H^{BW}]^T$. Let $\mathbf{G} \equiv \mathbf{G(n)}$ be the matrix of first partial derivatives of $\mathbf{g(n)}$. According to the delta method, the variance of $\mathbf{g(n)}$, $\mathbf{V}_{(\mathbf{g(n)})}$, is approximated by

$$\mathbf{V_{g(n)}} \approx \mathbf{G V_n G}^T \quad (5)$$

The covariance matrix of the scalability coefficients can be estimated as $\hat{\mathbf{V}}_{\mathbf{g(n)}}$ by using the sample estimates for $\mathbf{G}$ and $\mathbf{V_n}$. For two-level scalability coefficients, Koopman et al. (in press) derived matrix $\mathbf{G}$ in Equation 5. Because the derivations are rather cumbersome and lengthy, they are omitted here. The interested reader is referred to Koopman et al. (in press). The estimated delta-method standard errors $SE_d(H)$ are obtained by taking the diagonal of $(\hat{\mathbf{V}}_{\mathbf{g(n)}})^{1/2}$.

Bootstrap Standard Errors

The nonparametric bootstrap is a commonly used and easy to implement method to estimate standard errors (see, for example, Efron & Tibshirani, 1993; Van Onna, 2004). This method resamples the observed data with replacement to gain insight in the variability of the estimated coefficient. The bootstrap requires that all resampled observations are independent and identically distributed. Because in the two-level data structure the observations within subjects are expected to correlate, a standard bootstrap will not work. The cluster bootstrap accommodates for this dependency by resampling the subjects, thereby retaining all raters of that subject (see, for example, Deen & De Rooij, in press; Field & Welsh, 2007; Harden, 2011; Ng, Grieve, & Carpenter, 2013; Sherman & Le Cessie, 1997).

A bootstrap procedure is balanced if each observation occurs an equal number of times across the B bootstrap samples. Balancing the bootstrap can reduce the variance of the estimation, resulting in a more efficient estimator (Chernick, 2008, p. 131; Efron & Tibshirani, 1993, pp. 348-349). The following algorithm is used to estimate a standard error with a balanced cluster bootstrap.

1. For a bootstrap of size B, replicate the S subjects from data $\mathbf{X}$ B times and randomly distribute these replications in a $B \times S$ matrix $\mathbf{S}$.
2. Create B cluster-bootstrap data sets $\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*$. To obtain $\mathbf{X}_b^*$, take the bth row of the $\mathbf{S}$ matrix; $\mathbf{X}_b^*$ consists of the observed ratings of all raters from the bootstrap subjects.
3. Compute the scalability coefficients H_b^W, H_b^B, and H_b^{BW} for each bootstrap data set $\mathbf{X}_b^*$.
4. Estimate the bootstrap standard errors $SE_b(H)$ by computing the standard deviation of the H_b coefficient across the bootstrap samples.

Resampling at subject-level ensures that the bootstrap samples reflect a similar data structure as the original data set. The cluster bootstrap allows observations within subjects to correlate, but observations between subjects should be independent. The correlation structure may differ per subject, and need not be known.

Method

Simulated data were used to investigate the bias of the two-level scalability coefficient estimates, bias of the standard error estimates, and coverage of the Wald-based CIs. To keep the simulation study manageable (and readable), a completely crossed design was avoided. Instead, bias and coverage were first investigated in a small study that included the most important independent variable, the rater effect σ_δ, and the two standard error estimation methods (the main design). Because the rater effect determines the scalability of subjects for a given test, it is considered the most important independent variable. Second, in a series of small studies with specialized designs the effects of other independent variables were investigated using the most promising standard error estimation method. Finally, remarkable results were further investigated in post hoc simulations.

Data Simulation Strategy

Computation of the scalability coefficients and their standard errors by means of the delta method only assumes that the item scores follow a multinomial distribution with varying multinomial parameters across subjects (Koopman et al., in press). The cluster bootstrap assumes that data between subjects are independent. Both assumptions hold under the discussed two-level IRT models, given that each subject has a unique set of raters. The authors used a parametric hierarchical rater model to generate data, parameterized as follows:

$$\begin{array}{lcl} \theta_s & \sim & i.i.d.N(0, \sigma_\theta^2), s = 1, .., S \\ \xi_{si} & \sim & \text{Graded response model, } i = 1, ..., J, \text{ for each } s \\ \delta_{sr} & \sim & i.i.d.N(0, \sigma_\delta^2), r = 1, ..., R_s, \text{ for each } s \\ X_{sri} & \sim & \text{Signal detection model, for each } s,\ r,\ i \end{array} \tag{6}$$

Latent trait values θ_s were sampled from a normal distribution with mean 0 and variance σ_θ^2. Ideal ratings ξ_{si} were obtained using a graded response model (Samejima, 1969). This model was used because it is the parametric version of the monotone homogeneity model that underlies Mokken scale analysis (Hemker, Sijtsma, Molenaar, & Junker, 1996). For latent trait value θ_s, item discrimination parameter α_i, and item-step location parameter β_{ix}, the probability of ideal rating $\xi_{si} \geq x$ ($x = 1, 2, \ldots, m$) according to the graded response model is

$$P(\xi_{si} \geq x|\theta_s) = \frac{\exp[\alpha_i(\theta_s - \beta_{ix})]}{1 + \exp[\alpha_i(\theta_s - \beta_{ix})]}. \tag{7}$$

Note that $P(\xi_{si} \geq 0|\theta_s) = 1$ and $P(\xi_{si} \geq m + 1|\theta_s) = 0$ by definition. Ideal ratings ξ_{si} were sampled from a multinomial distribution using the probabilities $P(\xi_{si} = x|\theta_s) = P(\xi_{si} \geq x|\theta_s) - P(\xi_{si} \geq x + 1|\theta_s)$ for each subject s and item i.

Rater deviations δ_{sr} were sampled from a normal distribution with mean 0 and variance σ_δ^2. For deviation δ_{sr} and ideal rating ξ_{si}, the probability of observed score $X_{sri} = x$, $P(X_{sri} = x|\xi_{si}, \delta_{sr})$, was obtained from a discrete signal detection model. In this model, the probabilities are proportional to a normal distribution in x with mean $\xi_{si} + \delta_{sr}$ and rating variance τ_r^2; that is,

$$P(X_{sri} = x|\xi_{si}, \delta_{sr}) \propto \exp\left\{-\frac{[x - (\xi_{si} + \delta_{sr})]^2}{2\tau_r^2}\right\} \tag{8}$$

(also, see Patz et al., 2002). The computed probabilities $P(X_{sri} = x|\xi_{si}, \delta_{sr})$ for the $m + 1$ answer categories were normalized to sum to 1. Finally, observations X_{sri} were sampled from a multinomial distribution with parameter $P(X_{sri} = x|\xi_{si}, \delta_{sr})$.

Main Design

Independent variables. *Rater effect* σ_δ had four levels, each reflecting a different degree of rater effect: $\sigma_\delta = 0.25$ (very small), $\sigma_\delta = 0.50$ (small), $\sigma_\delta = 0.75$ (medium), and $\sigma_\delta = 1$ (large). Because the rater effect determines the scalability of subjects for a given test, it is considered the most important independent variable. As noted earlier, both the subject effect σ_θ and the rater effect σ_δ affect the magnitude of the scalability coefficients. By setting $\sigma_\theta + \sigma_\delta = 2$, the magnitude of H^W was similar across the four levels of rater effect, which facilitated comparison. H^B and H^{BW} decreased as σ_δ increased.

Table 1. Population Values of the Two-Level Scalability Coefficients H^W, H^B, and H^{BW} and the *SD* of the Sampling Distribution for the Four Conditions of σ_δ in the Main Design.

σ_δ	0.25		0.50		0.75		1.00	
	H	*SD*	*H*	*SD*	*H*	*SD*	*H*	*SD*
H^W	.437	.037	.418	.034	.435	.029	.479	.025
H^B	.415	.038	.316	.038	.214	.036	.126	.032
H^{BW}	.948	.010	.756	.036	.483	.057	.262	.058

Standard-error estimation method had two levels: the delta method and the bootstrap method. These methods were applied to each level of rater effect.

Other variables in the main design were fixed: The *number of subjects* was $S = 100$, and each subject was rated by the independent group of raters of size $R_s = 5$. The *number of items* was $I = 10$, and each item had $m + 1 = 5$ *answer categories. Item discrimination* was equal for each item at $\alpha_i = 1$ (Equation 7), the *item-step location* parameter β_{ix} (Equation 7) had equidistant values between values –3 and 3, and *rating variance* $\tau_r^2 = 0.5^2$ (Equation 8).

Dependent variables. The scalability coefficients H and standard errors of the estimates *SE* were computed for the three classes of the two-level total-scale scalability coefficients (H^W, H^B, and H^{BW}). Item-pair and item scalability coefficients were not computed because the total-scale coefficient can be written as a normalized weighted sum of the H_{ij} or H_i coefficients (Mokken, 1971, pp. 150-152). Therefore, it is expected that potential bias of H_{ij} or H_i is reflected in H. In the *specialized design*, the authors investigated conditions with two items; in that case, $H_{ij} = H_i = H$.

Bias of the estimated H coefficient. Bias reflects the average difference between the sample estimate and population value of H. Let H_q be the estimated scalability coefficient of the qth replication. The bias was determined across Q replications as $\text{Bias}(H) = Q^{-1} \sum_{q=1}^{Q} (H_q - H)$. The population values (Table 1) were determined based on a finite sample of 1,000,000 subjects and five raters per subject. Table 1 shows that H^B and H^{BW} decrease as rater effect σ_δ increases. As the rater effect in Table 1 increases the difference between H^B and H^W becomes larger. Therefore, the correlation between the sample estimates of H^B and H^W will be larger for small rater effects than for large rater effects. On average, a relative Bias(H) of 10% reflects a value of 0.044. Therefore, absolute bias values below 0.044 is considered satisfactory.

Bias of the estimated standard errors. Let SE_q be the standard error of the qth replication, and *SD* the population standard error, then $\text{Bias}(SE) = Q^{-1} \sum_{q=1}^{Q} [SE_q - SD]$. The population *SD* values (Table 1) were determined by the standard deviation of H_q across the Q replications and is assumed to be representative of the true standard deviation of the sampling distribution of H, under the conditions of the main design. On average, a relative Bias(*SE*) of 10% reflects a value of 0.004. Therefore, absolute bias values below 0.004 is considered satisfactory.

Coverage. Coverage of the 95% CIs was computed as the proportion of times, in Q replications, the population value H was included in the Wald-based confidence interval $\text{CI}_q = H_q \pm 1.96 SE_q$. This interval is selected because the distribution of the two-level scalability coefficients is asymptotically normal (Koopman et al., in press). There were $Q = 1,000$ replications per condition, and $B = 1,000$ balanced bootstrap samples per replication.

Analyses. The simulation study was programmed in R (R Core Team, 2018). and partly performed on a high performance computing cluster. The scalability coefficients and delta method standard errors were computed using the R-package mokken (Van der Ark, 2007, 2012; also, see Koopman et al., in press). The main design had eight conditions (two standard error

Table 2. Rater Effect (σ_δ) and Rating Variance (τ_r^2) Values for the Number of Answer Categories ($m+1$) Specialized Design.

		Rater effect σ_δ			
$m+1$	τ_r	0.25	0.50	0.75	1.00
2	.3	0.18	0.27	0.35	0.45
3	.4	0.20	0.33	0.48	0.65
5	.5	0.25	0.50	0.75	1.00
6	.5	0.30	0.70	1.00	1.20

Note. $m+1=5$ is the level from the main design.

estimation methods × four rater effect levels). Summary descriptives were computed and visualized for relevant outcome variables for all scalability coefficients. An Agresti–Coull CI (Agresti & Coull, 1998) was constructed around the estimated coverage using R-package binom (Dorai-Raj, 2014) to test whether it deviated from the desired value .95.

Specialized Designs

Each specialized design varied one of the independent variables that had been fixed in the main design. The levels of rater effect σ_δ remained unchanged (σ_δ = 0.25, 0.50, 0.75, and 1.00), to allow for the detection of potential interaction effects.

Independent variables. The following variables defined the specialized designs:

Number of subjects S was 50, 100 (as in main design) 250, or 500.
Number of raters per subject R_s had six conditions. Let $U\{a,b\}$ denote a discrete uniform distribution with minimum a and maximum b. In the six conditions R_s ($s=1,\ldots,S$) were sampled from $U\{2,2\}$, $U\{5,5\}$ (as in main design), $U\{30,30\}$, $U\{4,6\}$, $U\{3,7\}$, and $U\{5,30\}$, respectively. Hence, in the first three conditions, each subject had the same number of raters, and in the last three conditions the number of raters differed across subjects.
Rating variance τ_r^2 had four conditions. In three conditions, τ_r was fixed at 0.25, 0.50 (as in main design), and 0.75, respectively. In the fourth condition τ_r was sampled for each rater from an exponential distribution with mean $\lambda^{-1}=0.5$.
Number of items I was 2, 3, 4, 6, 10 (as in main design), or 20.
Number of answer categories $m+1$ had four levels: 2 (dichotomous items), 3, 5 (as in main design), and 7. The parameters of the signal detection model were adjusted according to the number of answer categories, to ensure that the magnitude of the scalability coefficients remained similar to those in the main design (Table 2).
Item discrimination parameter α_i had four levels. In three conditions α_i was kept constant for each item at 0.5, 1.0 (as in main design), or 1.5. In the last condition, the item discrimination varied across items at equidistant values between 0.5 and 1.5.
Distance between item-step location parameters β_{ix} had four levels. In the first three conditions, value β_{ix} ranged between –4.5 and 4.5, between –3 and 3 (as in main design), or between –1.5 and 1.5. In the last condition, the item-step locations were equal for the same item-steps across items, and ranged between –3 and 3 within items (i.e., $\beta_{i1}=-3, \beta_{i2}=-1.5, \beta_{i3}=1.5, \beta_{i4}=3$ for all i).

Table 3. Population Values for H^W, H^B, and H^{BW} for the Specialized Designs Item Discrimination α_i, Item-Step Location β_{ix}, and Rating Variance τ_r^2, for Rater Effect $\sigma_\delta = .5$.

	α_i				β_{ix}				τ_r			
	0.5	1	1.5	Varied	1.5	3	4.5	Equal	0.25	0.50	0.75	Varied
H^W	.185	.418	.569	.381	.377	.418	.439	.400	.464	.418	.357	.384
H^B	.125	.316	.439	.284	.327	.316	.270	.252	.343	.316	.269	.270
H^{BW}	.675	.756	.772	.747	.866	.756	.616	.630	.738	.756	.752	.704

Table 4. Bias of Estimated Coefficients (H) and of the Estimated Standard Errors (SE).

	Bias(H)			Bias(SE) delta			Bias(SE) bootstrap		
σ_δ	H^W	H^B	H^{BW}	H^W	H^B	H^{BW}	H^W	H^B	H^{BW}
0.25	–.000	–.001	–.002	.002	.002	**.006**	**–.007**	**–.007**	–.002
0.50	–.001	–.002	–.007	.002	.001	.004	**–.008**	**–.009**	**–.010**
0.75	.001	–.002	–.009	.003	.002	.004	**–.007**	**–.009**	**–.016**
1.00	.001	–.003	–.008	.003	.003	**.006**	**–.007**	**–.009**	**–.016**

Note. Bias that exceeds the boundary of .044 and .004 for SE and H^W, respectively, is printed in boldface.

Dependent variables and analyses. The dependent variables and statistical analyses were the same for the specialized designs and the main design. The specialized designs item discrimination, item-step location, and rating variance had an effect on the magnitude of (some of) the population H values, see Table 3. Population SDs were similar to those in the main design, but increased for fewer items and smaller sets of subjects or raters.

Post Hoc Simulations

Some exploratory simulations were performed to investigate aberrant results from the main and specialized designs.

Results

Main Design

Bias of all two-level scalability coefficients was close to zero across the different levels of rater effect σ_δ (Table 4, left panel).

Bias of the delta method standard error estimates was generally close to zero, but the bootstrap standard error estimates were negatively biased (Table 4, last two panels). As a result, coverage of the 95% CIs was too low for the cluster bootstrap, with values ranging between .82 and .88 across the different conditions and coefficients (Figure 1). The delta method coverage is excellent for the between-rater coefficient, but is conservative for the within-rater coefficient H^W if rater effect σ_δ is large (Figure 1). In addition, coverage of the ratio coefficient H^{BW} tends to be too high, especially if the rater effect is nearly absent. The high coverage may be explained by the small σ_δ value. For $\sigma_\delta = .25, H^B \approx H^W$, hence there is hardly any variation of H^{BW} across different samples, indicated by a true standard error of .01 (Table 1). The bias of the estimated standard error was .006 (Table 4, first row, sixth column), which is identical to the bias in

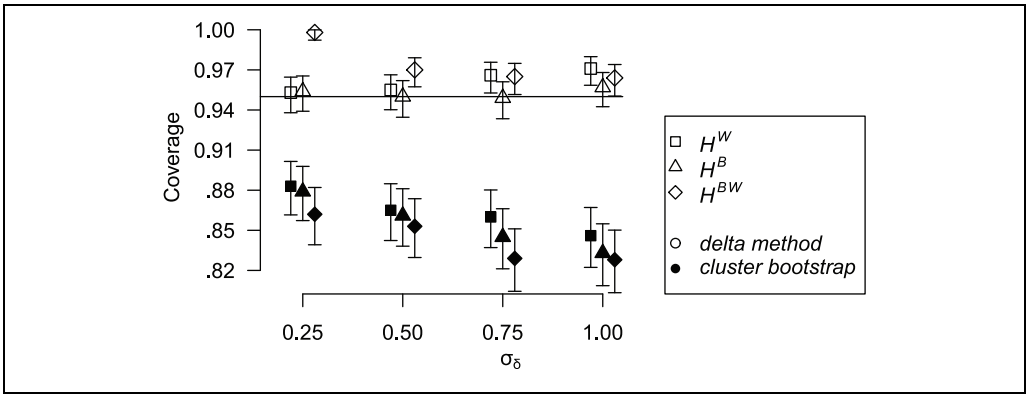


Figure 1. Plot of the coverage of the 95% confidence interval of the two-level scalability coefficients, for different levels of rater effect σ_δ and the two standard error estimation methods.
Note. Error bars represent the 95% Agresti–Coull confidence interval.

the $\sigma_\delta = 1$ condition (Table 4, last row, sixth column), for which the true standard error is .058 (Table 1). Relative to their true standard error, the bias of .006 was 60% for $\sigma_\delta = .25$, and only 10% for $\sigma_\delta = 1$. Therefore, coverage was much larger in the $\sigma_\delta = .25$ condition compared with the $\sigma_\delta = 1$ condition, even though the bias was equal.

Specialized Designs

For all conditions in the specialized designs, the bias of the point estimates of the two-level scalability coefficients was satisfactory with values between –.004 and .014. Because of the poor performance in the main design, the bias and coverage of the cluster-bootstrap standard errors were not computed in the specialized designs, so all results for the standard errors pertain to the delta method. Number of subjects, S, number of answer categories, $m + 1$, item discrimination, α_i, item-step location, β_{ix}, and rating variance, τ_r^2, had little or no effect on the bias of the estimated standard errors and the coverage of the Wald-based CI. As in the main design, for H^W and H^B, bias was satisfactory and coverages were accurate; whereas for H^{BW}, the bias was occasionally unsatisfactory–bias(SE) $\leq$ .008–and coverages conservative. Number of raters, R_s, and number of items, I had an effect (Table 5). No interaction effect was found between rater effect (σ_δ) and the specialized design variables. Therefore, results are discussed only for $\sigma_\delta = 0.5$.

For unequal numbers of raters, the standard errors of the two-level scalability coefficients were too conservative (Table 5, left panel) and the coverage of the CIs too high (Figure 2, left plot, right-hand side of the plot). The overestimation was stronger if the variation of R_s was larger. As in the main design with five raters, the standard errors were also too conservative for H^{BW} in the condition with two raters (Figure 2, left plot).

For two and three items, the standard errors were underestimated for the between-rater coefficient H^B (Table 5, right panel). As a result, coverage was too low (Figure 2, right plot).

Post Hoc Simulations

It was unexpected that the cluster bootstrap in the main design performed poorly in estimating the standard errors of the two-level scalability coefficients, resulting in poor coverage values. Apparently, the cluster bootstrap does not correctly approximate the sampling distribution of H

Table 5. Bias of the Delta Method Standard Errors (*SE*) for the Two-Level Scalability Coefficients H^W, H^B, and H^{BW} for Specialized Designs of Number of Raters (R_s) and Number of Items (I).

R_s	H^W	H^B	H^{BW}	I	H^W	H^B	H^{BW}
2	.002	.002	**.009**	2	.002	**−.009**	−.003
5	.002	.001	.004	3	.001	−.004	.000
30	.000	.000	.001	4	.002	−.001	.003
4-6	.004	**.005**	**.008**	6	.001	.001	**.006**
3-7	**.013**	**.015**	**.017**	10	.002	.001	.004
5-30	**.032**	**.037**	**.035**	20	.002	.002	.003

Note. Bias that exceeds the boundary of .004 is printed in boldface.

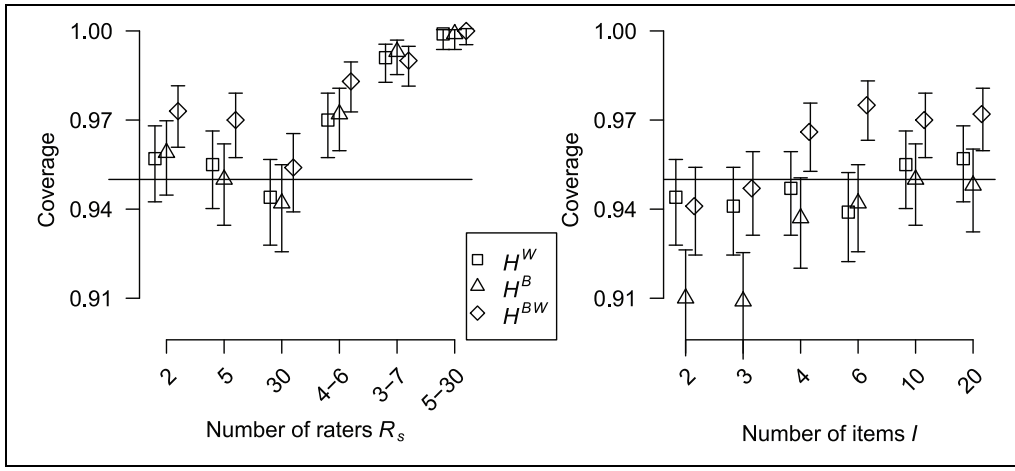


Figure 2. Coverage plots for the two-level scalability for different number of raters and items, respectively.

Note. Error bars represent the 95% Agresti–Coull confidence interval.

in the population. An explanation may be that the cluster bootstrap ignores the assumption that the raters should be a random sample of the population of raters. Therefore, an alternative, two-stage bootstrap is proposed (for a similar bootstrap procedure, see Ng et al., 2013). At Stage 1, the clusters are resampled as in the cluster bootstrap and at Stage 2, the raters of the selected subjects are resampled. Compared with the cluster bootstrap, the two-stage bootstrap resulted in substantial improvements in the standard error estimates and the coverages (Table 6, rows 1 and 2). In an effort to further improve the coverage rates of the two-stage bootstrap, the percentile and bias-corrected accelerated interval were also computed (see, for example, Efron & Tibshirani, 1993, pp. 170-187, for a detailed description). These two methods use the empirical distribution of H to construct an interval, rather than assuming a normal distribution. The coverages of the percentile and bias-corrected accelerated intervals were equal to or lower than the coverages of the Wald-based intervals. Because the bias and coverages of the two-stage bootstrap are still inferior to those of the delta method (Table 4, third row), the delta method remains the preferred method.

There were two odd results in the specialized designs: the relatively poor results of the standard error estimates for unequal group sizes and for a set of two items. The standard error

Table 6. Post Hoc Results of the Bias(*SE*) and Coverage for the Two-Stage and Cluster Bootstrap and the Delta Method, the Arithmetic and Harmonic Mean of R_s, and item-pairs H_{ij} with Two, Four, and 10 Items, for H^W, H^B, and H^{BW}, and Main Design Condition With $\sigma_\delta = .5$.

		Bias (*SE*)			Coverage		
		H^W	H^B	H^{BW}	H^W	H^B	H^{BW}
Method							
Two-stage bootstrap		–.003	–.004	**–.007**	**.930**	**.930**	**.880**
Cluster bootstrap		**–.008**	**–.009**	**–.010**	**.865**	**.861**	**.853**
Delta method		.002	.001	.004	.955	.950	**.970**
R_s	Mean						
4-6	A	.004	**.005**	**.008**	**.970**	**.972**	**.983**
	H	.003	.003	**.007**	**.965**	**.965**	**.979**
3-7	A	**.013**	**.015**	**.017**	**.991**	**.993**	**.990**
	H	**.009**	**.011**	**.013**	**.984**	**.984**	**.989**
5-30	A	**.032**	**.037**	**.021**	**.999**	**.999**	**1.00**
	H	**.018**	**.021**	**.021**	**.992**	**.994**	**.999**
Number of Items							
2		.002	**–.009**	–.003	.944	**.910**	.941
4		.002	–.001	**.011**	.945	.938	**.983**
10		.002	.003	**.019**	.950	.953	**.989**

Note. Bias that exceeds the boundary of .004 and coverages where .95 is outside the Agresti–Coull interval are printed in boldface. The two-stage bootstrap results are based on 100 replications. The H_{ij} results are averaged across all item-pairs. A = arithmetic mean and H = harmonic mean of R_s.

estimates of the two-level scalability coefficients rapidly increased if the variation in number of raters across subject became larger. For unequal number of raters across subjects, R in Equation 4 was estimated by the (arithmetic) sample mean $\widehat{R} = S^{-1} \sum_{s=1}^{S} R_s$. As a solution, the authors estimated R by the harmonic mean, which is lower than the arithmetic mean if group sizes differ, and is computed as $\widehat{R} = S / \sum_{s=1}^{S} R_s^{-1}$. Using the harmonic mean improved the bias of the standard error and the coverage compared to the use of the arithmetic mean (Table 6, rows 4-9). However, the estimates were still too conservative, and equal group sizes are preferred.

The standard error of between-rater coefficient H^B was underestimated for sets of two items. Although, in general, testing with a small set of items is discouraged (see, for example, Emons, Sijtsma, & Meijer, 2007), this condition was of interest because for only two items, the total-scale coefficient H^B is equal to item-pair coefficient H_{ij}^B. To investigate whether bias in the standard error of item-pair coefficient H_{ij}^B persisted for larger sets of items, the coefficients and their standard errors were computed in a new condition with four items and in the main design with 10 items (both for $\sigma_\delta = .5$). As is shown in Table 6, bottom three rows, bias of H_{ij}^B standard errors vanished as the number of items increased. However, Table 6 also shows that the standard error estimates estimates and coverages of item-pair ratio coefficient H_{ij}^{BW} were increasingly conservative, more than the total-scale coefficient H^{BW}.

Discussion

Point estimates of the two-level scalability coefficients were unbiased in all conditions, with bias values approximately zero. Standard errors were mostly unbiased if the delta method was used but not for the traditional cluster bootstrap. A two-stage cluster bootstrap was proposed that partially mitigated the bias, yet the delta method remains the preferred method.

The delta method resulted in unbiased standard error estimates for both the within- and between-rater scalability coefficients H^W and H^B, respectively. For large rater effects, the coverage of the within-rater coefficient H^W was slightly conservative. However, if the rater effect is large, standard errors are of less interest, because the test will be determined of poor quality based on the (unbiased) coefficients alone. Standard error estimates and coverages for ratio coefficient H^{BW} were conservative, especially if H^{BW} was close to its upper bound 1. In this latter situation, standard errors are also of less interest, because if the coefficient estimate is so high, so is its interval estimate.

For all coefficients, the delta method overestimated the standard error if the number of raters was unequal across subjects, especially if the variation was larger. Post hoc simulations showed some improvements if the harmonic mean of the group size was used rather than the arithmetic mean, but equal group sizes are recommended. In addition, for small sets of items the standard errors between-rater coefficient H^B were too liberal. Post hoc simulations showed that the standard errors of the total scale and the item-pair between-rater coefficients are unbiased, provided that a scale consists of at least four items.

The results of this study demonstrate that, in general, the estimated scalability coefficients and delta method standard errors are accurate and can therefore be confidently used in practice. When the scalability of a multi-rater test is deemed satisfactory, a related (but different) topic concerns the reliability. For a given test, Snijders (2001) presented coefficient alpha to determine how many raters are necessary for reliable scaling of the subjects. Note that the magnitude of the scalability coefficients is not affected by the number of raters. Alternatively, generalizability theory provides a more extensive selection of methods to investigate reliability (generalizability) of multi-rater tests (see, for example, Shavelson & Webb, 1991).

The application of two-level scalability coefficients and their standard errors is not limited to multi-rater data. They may also be applied in research with multiple (random) circumstances or time points in which the same questionnaire is completed. Also, the items may be replaced by a fixed set of situations in which a particular skill is scored using a single item. The standard errors examined in this article are also useful for single-level Mokken scale analysis for data from clustered samples (e.g., children nested in classes) because the single-level standard error will typically underestimate the true standard error (see, for example, Koopman et al., in press). Future research may focus on how the point and interval estimates can be useful to select a subset of items from a larger set of items.

Appendix

Illustrative Example

Table A1 shows two small constructed data examples, each with two subjects and five raters per subject on two three-category items. The same item scores are present in both data sets, but Rater 4 of Subject 1 and Rater 5 of Subject 2 are exchanged in the second data set.

For both data sets in Table A1, the item-step ordering is $Z_{i1}, Z_{j1}, Z_{i2}, Z_{j2}$. Therefore, consistent item-score patterns are $(0, 0)$, $(1, 0)$, $(1, 1)$, $(2, 1)$, and $(2, 2)$, whereas patterns $(0, 1)$, $(0, 2)$, $(1, 2)$, and $(2, 0)$ are Guttman errors. Within raters, Guttman error $(0, 1)$ occurs once in each data set (Rater 2 of Subject 2). In the first data set, there are five between-rater Guttman errors $(0, 1)$ (for Subject 2, Rater 1 scored 0 on X_i, whereas Raters 2, 4, and 5 scored 1 on X_j, and Rater 2 scored 0 on X_i, whereas Raters 4 and 5 scored 1 on X_j), four between-rater Guttman errors $(1, 2)$, and five between-rater Guttman errors $(2, 0)$, summing up to 14 between-rater Guttman errors. In the second data set there are only three $(0, 1)$ and two $(1, 2)$ between-rater Guttman errors, summing up to five.

Table A1. Two Small Constructed Multi-Rater Data Examples, One With a Large Rater Effect and One With a Small Rater Effect.

	Data set 1:Large rater effect										Data set 2:Small rater effect									
	$s = 1$					$s = 2$					$s = 1$					$s = 2$				
$r =$	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Item X_i	2	2	2	1	1	0	0	1	1	2	2	2	2	2	1	0	0	1	1	1
Item X_j	1	2	2	0	1	0	1	0	1	1	1	2	2	1	1	0	1	0	1	0
	$\overline{X}_1 = 1.4$					$\overline{X}_2 = 0.7$					$\overline{X}_{2\cdot} = 1.6$					$\overline{X}_2 = 0.5$				
	$H^W = .762$					$H^B = .167$			$H^{BW} = .219$		$H^W = .762$					$H^B = .702$			$H^{BW} = .922$	
95% CI	[0.343, 1.181]					[−0.231, 0.565]			[−0.288, 0.726]		[0.349, 1.175]					[0.435, 0.970]			[0.441, 1.402]	

Note. 95% CI is the 95% Wald-based confidence interval. Both data sets have two subjects (s), and each rated by a unique set of five raters (r) on two three-category items (X_i and X_j).

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

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

Acknowledgment

The authors thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster to conduct our Monte Carlo simulations.

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: This research was funded by the Netherlands Organization for Scientific Research (NWO; Grant 406.16.554).

ORCID iD

Letty Koopman https://orcid.org/0000-0003-3832-2542

References

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). New York, NY: John Wiley.

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

Bull, S., Darlington, G., Greenwood, C., & Shin, J. (2001). Design considerations for association studies of candidate genes in families. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *20*, 149-174. doi:10.1002/1098-2272(200102)20:2<149::AID-GEPI1>3.0.CO;2-A

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

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

Crisan, D. R., Van de Pol, J. E., & Van der Ark, L. A. (2016). Scalability coefficients for two-level polytomous item scores: An introduction and an application. In L. A. Van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research: The 80th annual meeting*

of the Psychometric Society, Beijing, 2015 (pp. 139-154). New York, NY: Springer. doi:10.1007/978-3-319-38759-8_11

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

Deen, M., & De Rooij, M. (in press). ClusterBootstrap: An R package for the analysis of clustered data using generalized linear models with the cluster bootstrap *Behavior Research Methods*. doi: 10.3758/s13428-019-01252-y

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

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

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

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

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

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

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

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

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (in press). Standard errors of two-level scalability coefficients *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12174

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. doi:10.1177/0081175013481958

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

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

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

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

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

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

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

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

R Core Team (2018). R: *A language and environment for statistical computing* [computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/)

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

Reise, S. P., Meijer, R. R., Ainsworth, A. T., Morales, L. S., & Hays, R. D. (2006). Application of group-level item response models in the evaluation of consumer reports about health plan quality. *Multivariate Behavioral Research*, *41*, 85-102. doi:10.1207/s15327906mbr41016

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* [Psychometrika Monograph Supplement No. 17]. Richmond, VA: Psychometric Society.

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.

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

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

Snijders, T. A. B. (2001). Two-level non-parametric scaling for dichotomous data. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 319-338). New York, NY: Springer. doi:10.1007/978-1-4613-0169-1_17

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

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

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

Van der Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research*, *49*, 127-152. doi:10.1080/001 31880701369651

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

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