Mathematical Institute

Master Thesis

Statistical Science for the Life and Behavioural Sciences

Comparing the delta method, individual level bootstrap, and cluster level bootstrap to compute standard errors of two-level scalability coefficients

a simulation study

Author:

Letty Koopman

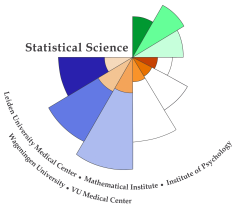
s1576380

Supervisor:

Prof. dr. M.J. de Rooij

Leiden University

December, 2016



Abstract

Several methods are available to estimate standard errors of a coefficient. Two-level Mokken scale analysis is an ordinal scaling technique that accounts for a multilevel test structure, where the subjects to be scaled are scored by various raters. Key in this analysis are the two-level scalability coefficients. Recently, standard errors of these coefficients have been estimated using the delta method. It is uncertain whether this method results in biased standard error estimates. The individual level bootstrap method is often regarded as an unbiased estimation method of standard errors. An extension of this method is the cluster level bootstrap, which maintains the dependency structure in the data. This simulation study compares these three methods on their bias, efficiency, coverage, and computation time. Results indicate that the difference in bias, efficiency, and coverage favoured the individual level bootstrap, although the difference was in most conditions very close to zero. Since computation time was much higher for the bootstrap methods, the delta method is preferred in practice.

Keywords: cluster level bootstrap, delta method, individual level bootstrap, standard errors, two-level scalability coefficients.

Comparing the delta method, individual level bootstrap, and cluster level bootstrap to compute standard errors of two-level scalability coefficients

1 Introduction

Tests with a multilevel structure occur often in educational and psychological research, characterized by subjects who are assessed by multiple raters. An example of such a measurement instrument is the ICALT observation instrument, a multiple-item questionnaire measuring the teaching skills of trainee teachers throughout the Netherlands. Typically, each trainee teacher is assessed by, say, five instructors, so a single trainee teachers skill is measured multiple times, each time by a different individual. The five instructors average test score is the measured value of the trainee teachers teaching skills. Another example is course evaluations of Dutch universities, in which courses are rated by multiple students using a standardized multi-item questionnaire. The intention in these specific situations is to represent or develop a scale in order to measure a latent trait of a *subject*, for example the trainee teacher, while information is only available from the *raters*, e.g. the instructors. This results in multilevel test data, since the raters are nested within the subjects. If this multilevel structure is ignored when investigating the quality of such measurement instruments, the outcome cannot be trusted and will look more favourable than it actually is (Crisan, Van de Pol, & Van der Ark, 2016). Specifically, ignoring the multilevel structure results in

quality estimates that do not distinguish in the degree to which the item responses are determined by the subject or by the rater.

A technique to accommodate for multilevel structure in test construction is multilevel Mokken scale analysis (Snijders, 2001; Crisan et al., 2016). Mokken scale analysis (e.g., Mokken, 1971; Sijtsma & Molenaar, 2002) is an ordinal scaling method often used for the construction or evaluation of tests in the social and behavioural sciences. This scaling method is based on non-parametric item response theory (NIRT) models, more specifically the monotone homogeneity model (MHM). In this model no particular shape of the item response functions is assumed. Specifically, the three assumptions underlying the MHM are unidimensionality of the latent trait, local independence of responses given the latent trait, and monotonicity of the item response functions. An important diagnostic tool in Mokken scale analysis to assess the quality of a test are coefficients H, based on Loevinger (1948). These coefficient essentially reflect the homogeneity of item pairs (H_{ij}), of items in regard to the rest of the scale (H_i), and of the scale itself (H). The values can be used to assess how well subjects can be ordered on the scale. For multilevel test data, all H coefficients have a within- and a between-rater version (Snijders, 2001). The within-rater scalability coefficients H^W reflect the consistency of response patterns on items within raters, and are comparable to the H coefficients in the one-level situation. In other words, H^W reflects to which degree the items form a cumulative scale. The between-rater scalability coefficients H^B reflect the consistency of response patterns on items between raters within the same subject and is a measure of the degree to which the subjects' latent trait determines the item responses. Off

special interest is the ratio H^B/H^W, where higher values represent a higher influence of the subjects' latent trait relative to influences of the rater itself. The values of, as well as the ratio between, the within- and between-rater scalability coefficients enable evaluation of the test quality.

Evaluating a questionnaire requires well-approximated standard errors for the scalability coefficients. Standard errors give a more complete description of the coefficients and can determine the precision of the estimate. In addition, ignoring them can result in incorrect inferences, such as including items that do not have a value significantly larger than a particular threshold (Kuijpers, Van der Ark, & Croon, 2013). In 1971, Mokken described computation of standard errors for small sets of dichotomous items. Since all possible item response patterns were necessary to compute the standard errors, this process was both time consuming and computer intensive when datasets became larger. In 2013, Kuijpers et al. derived standard errors for the one-level scalability coefficients in a marginal modelling framework using the delta method. Recently, a similar strategy has been applied to estimate the standard errors for the two-level scalability coefficients by Koopman (2016). However, it is unknown how well this method performs for the two-level scalability coefficients and therefore it is relevant to investigate the bias, efficiency, and coverage of this method, and compare it to methods that are known for their unbiased estimates. The individual level bootstrap is a well known method often considered robust for estimating standard errors in a wide range of situations (Efron & Tibshirani, 1993). In addition, Van Onna (2004) showed that the bootstrap resulted in good approximations of the distribution and variance of the one-level H coefficients. However, due to

the multilevel structure of the discussed data the variance might be overestimated. Therefore, it might be more appropriate to expand this method to a cluster level bootstrap, where the resampling is based on the subject-level, which is the cluster, rather than the individual level.

The intention of this simulation study is to compare these three methods on both theoretical and practical factors, resulting in the questions: What are the differences in bias, efficiency, coverage, and computation time when computing standard errors for the two-level scalability coefficients using the delta method, the individual level bootstrap, and the cluster level bootstrap? In addition, do any differences depend on other factors known to influence the sampling error of the H coefficients? The remainder of this paper elaborates on the computation of the two-level scalability coefficients and the three methods to estimate the standard errors (Section 2), expands on simulation study (Section 3), and outlines and discusses the results (Section 4 and 5).

2 Scalability coefficients and their standard errors

2.1 Two-level scalability coefficients

In a multilevel test containing J items (indexed by i or j) with $z + 1$ answer categories (x or y), each of S subjects (s or t) is rated by R_s raters (r or p). The total number of raters is defined by $R = \sum_{s=1}^{R_s} R_s$. Score X_{sri} is defined as the item response for subject s by rater r on item i. The average

score for subject s on item i is defined as

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

Subjects are generally scaled by the mean total score across raters

$$\overline{X}_{s\cdot +} = \frac{1}{R_s} \sum_{r=1}^{R_s} \sum_{i=1}^{I} X_{sri}. \tag{2}$$

The item responses are driven by the latent trait of the subject, θ_s, and the deviation of the rater within the subject, δ_{sr}. Deviations δ_{sr} are independent and identically distributed. Each item i has item-step response function $p_{ix}(\theta_s + \delta_{sr}) = P(X_{sri} \geq x|\theta_s, \delta_{sr})$, which is the probability of obtaining at least item score x given θ_s and δ_{sr}. These functions can be averaged to only depend on θ_s, which is defined as $p_{ix}(\theta_s) = P(\overline{X}_{s\cdot i} \geq x|\theta_s) = E_\delta[p_{ix}(\theta_s + \delta_{sr})]$, with expectation E_δ referring to the distribution of δ_{sr}. If the MHM holds then the subjects can be ordered on latent variable θ according to their total score $\overline{X}_{s\cdot +}$.

2.1.1 Guttman errors

Guttman error computation is a main element of the two-level scalability coefficients. Let item-step g reflect a Boolean statement indicating whether a particular score is at least x. This takes the value 1 if the step is passed, that is, $X_{sri} \geq x$, or value 0 if the step is failed, that is, $X_{sri} < x$. Let the item-step popularity be defined by the cumulative probability of scoring at least score x on item i, that is, $P(X_{sri} \geq x)$.

Assuming an equal number of categories per item, there are $2z$ item-steps for each item-pair. Within the same item the item-step ordering is fixed, but the ordering of item-steps from two different items are based on their item popularity. In a perfect Guttman scale no further item-steps are passed once a step is failed. In this light, a Guttman error happens when a less popular item-step is passed after failing a more popular step. For example, if the popularity of item b is less than the popularity of item a, item-score pattern $(X_a = 0, X_b = 1)$ is considered a Guttman error. Expanding this example, let the item-step ordering of two trichotomous items be

$$X_a \geq 1, X_b \geq 1, X_a \geq 2, X_b \geq 2. \tag{3}$$

Item-steps $X_a \geq 0$ and $X_b \geq 0$ are omitted, since $P(X_{sri} \geq 0)$ equals 1 per definition, rendering them uninformative. Evaluating each item-step in Equation 3 as value u_g^{xy} results in vector $\mathbf{u}^{xy}$. Guttman errors exist for item-score patterns (0, 1), (0, 2), and (1, 2). For pattern (1, 2), $\mathbf{u}^{12}$ = (1, 1, 0, 1), where the third step is failed, while the fourth step is passed. The *weight* of this error is indicative for the degree of deviation from a perfect Guttman scale (Molenaar, 1991). The weight for pattern (1, 2) is 1, since only one item-step is failed before another item step is passed. Guttman weights w_{ij}^{xy} for score x on item i and score y on item j can be computed as

$$w_{ij}^{xy} = \sum_{h=2}^{2z} \left\{ u_h^{xy} \times \left[\sum_{g=1}^{h-1} (1 - u_g^{xy}) \right] \right\}, \tag{4}$$

(see e.g., Kuijpers et al., 2013)

2.1.2 Item-pair, item, and total scale coefficients

Scalability coefficients compare the sum of weighted observed Guttman errors F_{ij} to the sum of weighted expected Guttman errors under marginal independence of the items E_{ij}. Let $P(W)_{ij}^{xy}$ be the bivariate probability that $X_{sri} = x$ and $X_{srj} = y$, that is, the item score patterns within the raters, used in within-rater coefficients, and $P(B)_{ij}^{xy}$ the bivariate probability that $X_{sri} = x$ and $X_{spj} = y$, that is, the item score patterns between different raters within the same subject, used in the between-rater coefficients. Moreover, let P_i^x be the univariate overall probability that $X_{sri} = x$.

There are $K = J(J-1)/2$ item-pair scalability coefficients of both H_{ij}^W and H_{ij}^B. These coefficients reflect the homogeneity of item pairs within and between the raters, respectively, and are defined as

$$H_{ij}^W = 1 - \frac{F_{ij}^W}{E_{ij}} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} P(W)_{ij}^{xy}}{\sum_x \sum_y w_{ij}^{xy} P_i^x P_j^y}, \tag{5}$$

and

$$H_{ij}^B = 1 - \frac{F_{ij}^B}{E_{ij}} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} P(B)_{ij}^{xy}}{\sum_x \sum_y w_{ij}^{xy} P_i^x P_j^y}, \tag{6}$$

For each of the J items the item scalability coefficients H_i^W and H_i^B, exist. These coefficients reflect the homogeneity of items with respect to the rest of the scale, again within and between the raters, and are defined as

$$H_i^W = 1 - \frac{\sum_{j \neq i} F_{ij}^W}{\sum_{j \neq i} E_{ij}} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(W)_{ij}^{xy}}{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P_i^x P_j^y}, \tag{7}$$

and

$$H_i^B = 1 - \frac{\sum_{j \neq i} F_{ij}^B}{\sum_{j \neq i} E_{ij}} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(B)_{ij}^{xy}}{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P_i^x P_j^y}, \tag{8}$$

The homogeneity of the total scale is summarized by scalability coefficients H^W and H^B, defined as

$$H^W = 1 - \frac{\sum \sum_{j \neq i} F_{ij}^W}{\sum \sum_{j \neq i} E_{ij}} = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(W)_{ij}^{xy}}{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P_i^x P_j^y}. \tag{9}$$

and

$$H^B = 1 - \frac{\sum \sum_{j \neq i} F_{ij}^B}{\sum \sum_{j \neq i} E_{ij}} = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(B)_{ij}^{xy}}{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P_i^x P_j^y}. \tag{10}$$

The ratio of the between- to within-rater scalability coefficients H^B/H^W results in higher values when item responses are to a larger degree determined by the latent trait of the subject rather than by rater effects. Snijders (2001) suggested as values $H^B \geq .01$ and $H^W \geq .02$ to be reasonable, and $H^B/H^W \geq 0.3$ to be a reasonable and $H^B/H^W \geq 0.6$ to be excellent as a scale to order subjects on.

Let $\mathbf{1}(\mathrm{X} = \mathrm{x})$ represent an indicator function returning value 1 if X is x, and 0 otherwise. The bivariate and univariate probabilities are estimated from the data with the following formulas

$$\widehat{P(W)}_{ij}^{x,y} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r=1}^{R_s} \mathbf{1}(X_{sri} = x, X_{srj} = y), \tag{11}$$

$$\widehat{P(B)}_{ij}^{x,y} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{R_s(R_s - 1)} \sum_{p \neq r} \sum^{R_s} \mathbf{1}(X_{sri} = x, X_{spj} = y). \tag{12}$$

Since the simple estimation of the univariate probability $P_i^x = \sum \mathbf{1}(X_{sri} = x)/R$ is biased when there is a relation between θ and R_s, its estimate is based on the relative frequencies with

$$\widehat{P}_i^x = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r=1}^{R_s} \mathbf{1}(X_{sri} = x). \tag{13}$$

The estimated two-level scalability coefficients are obtained by implementing these estimates in Equation 5 to 10.

2.2 Standard error estimation methods

The three standard error estimation methods for this study are the delta method, the individual level bootstrap, and the cluster level bootstrap.

2.2.1 Delta method

The delta method approximates the variance of the transformation of a variable (e.g., Agresti, 2002; Sen & Singer, 1993). Let $\mathbf{n}$ be a vector with observed item response patterns in the test data. For two dichotomous items a and b this vector results in

$$\mathbf{n} = \begin{pmatrix} n_{ab}^{00} \\ n_{ab}^{01} \\ n_{ab}^{10} \\ n_{ab}^{11} \end{pmatrix}. \tag{14}$$

Let $\mathbf{V_n}$ be the variance-covariance matrix of vector $\mathbf{n}$, let $\mathbf{D}(\mathbf{n})$ be the diagonal matrix with $\mathbf{n}$ on the diagonal, and let $\mathbf{G} \equiv \mathbf{G}(\mathbf{n})$ be the matrix of first partial derivatives of $\mathbf{g}(\mathbf{n})$. Assuming $\mathbf{n}$ is sampled from a multinomial

distribution, its variance-covariance matrix is $\mathbf{V_n} = \mathbf{D(n)} - \mathbf{n}N^{-1}\mathbf{n}^T$. Let the scalability coefficients be a transformation of vector $\mathbf{n}$, that is, $H = \mathbf{g(n)}$. According to the delta method, its variance is approximated by

$$\begin{aligned} \mathbf{V_{g(n)}} &\approx \mathbf{G}\ \mathbf{V}_n\ \mathbf{G}^T \\ &= \mathbf{G}\ \left[\mathbf{D(n)} - \mathbf{n}N^{-1}\mathbf{n}^T\right]\ \mathbf{G}^T \\ &= \mathbf{G}\ \mathbf{D(n)}\ \mathbf{G}^T - \mathbf{G}\ \mathbf{n}\ N^{-1}\ \mathbf{n}^T\ \mathbf{G}^T. \end{aligned} \tag{15}$$

Because scalability functions are homogeneous functions of order 0, $\mathbf{g(n)}$ does not change when $\mathbf{n}$ is multiplied by the same constant t, $\mathbf{g}(t\mathbf{n}) = \mathbf{g(n)}$. Consequently, according to Euler's homogeneous function theorem (e.g., Weisstein, 2011), $\mathbf{Gn} = \mathbf{0}$, and Equation 15 simplifies to $\mathbf{G}\ \mathbf{D(n)}\ \mathbf{G}^T$. To result in the standard errors as estimated by the delta method $\widehat{se}(H) = \text{diag}\sqrt{\mathbf{V_{g(n)}}}$

Recently, the delta method has been implemented to estimate the standard errors for the two-level scalability coefficients by rewriting the scalability coefficients as a vector function of the data, which enables easy derivation of the matrix of first-order partial derivatives of the vector functions. With these derivatives, the delta method can be applied (Koopman, 2016; see also Kuijpers et al., 2013).

2.2.2 Individual level bootstrap

A commonly used, robust method to estimate standard errors and confidence intervals is the individual level, non-parametric bootstrap (e.g., Efron & Tibshirani, 1993). This method resamples the observed data with replacement to gain insight in the variability of the estimated coefficient. The algorithm for standard error estimation is

1. Select B independent bootstrap samples $X_1, X_2, ..., X_B$ of size R with

replacement from data set X.

2. Compute the relevant statistic for each bootstrap sample b ($b = 1, ..., B$), in current situation the two-level scalability coefficients $\widehat{H}_b$.

3. Estimate the standard error se_i by computing the standard deviation of the statistic over the bootstrap samples

$$\widehat{se}(\widehat{H}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{H}_b - \overline{H})^2}, \tag{16}$$

where

$$\overline{H} = \frac{1}{B}\sum_{b=1}^{B}\widehat{H}_b. \tag{17}$$

The developments in computational power has enabled the use of bootstrap in various contexts. However, when computation of the statistic is time consuming, the bootstrap will be quite computer intensive, making it of less practical value. In 2004, Van Onna found that the distribution of the one-level coefficient H could be well approximated by the bootstrap. It appeared to follow a normal distribution, as expected from the central limit theorem, which justifies the use of symmetric confidence intervals around the mean, that is, $H \pm z_{\alpha/2}se$.

2.2.3 Clustered level bootstrap

The two-level coefficient are used for multilevel test data, where there is a dependency structure present in the data. This means that the item responses from raters within a subject are expected to be more alike than from raters

between two subjects, since the responses depend on θ_s. A way to accommodate for this dependency is by adjusting the individual level bootstrap to a cluster level bootstrap (e.g., Deen & De Rooij, 2016; Sherman & Le Cessie, 1997; Cheng, Yu, & Huang, 2013). In a cluster bootstrap, the dependency between ratings within subjects remains intact by resampling at the subject level (which is the cluster) rather than the individual level. This way all observations within a subject are retained. This reduces the variability of the estimate, since the cluster bootstrap samples reflect a more similar data structure in comparison to the original data set. The estimated standard errors of the cluster level bootstrap se_c is thus similarly computed as se_i, only the bootstrap samples differ. Since the cluster bootstrap provides robust standard errors for multilevel data, it will be a good comparison to the standard error estimation as computed by the delta method. However, the additional effort of clustering the bootstrap might be unnecessary when the individual level bootstrap performs well.

2.3 Hypotheses

In the one-level situation, bias of the standard errors appeared to be negligible for multiple factors, i.e., distance between item steps, sample size, number of items, number of answer categories, and item discrimination (Kuijpers, Van der Ark, Croon, & Sijtsma, 2016). In addition, coverage of the 95% confidence intervals was close to .95, although this became poorer when the sample size was smaller and when dichotomous items were used. Since the within-rater scalability coefficients are similar to the one-level coefficients, the results are expected to be comparable. However, additional

manipulations specific for the multilevel situation need to prove their effects, like the number of raters per subject and the ratio of the variance of θ to δ. Since the standard errors of the between-scalability coefficients have not been investigated, there are no expectations about their bias and coverage. Since previous research did not include efficiency as a outcome measure, it is currently expected that their efficiency is similar. Regarding the two bootstrap procedures, it is expected that the cluster bootstrap outperforms the regular bootstrap, especially for the H^B coefficient, since this method retains the dependency structure between raters in the same subject. It is yet unknown whether the performance of either bootstrap will differ from the delta method.

3 Method

3.1 Simulation model

The data for the simulation study is generated using the parametric graded response model (Samejima, 1969). In the one-level situation, this model is a special case of the monotone homogeneity model in Mokken scale analysis (Hemker, Sijtsma, Molenaar, & Junker, 1996). For the moment it is assumed that this relation also holds in the two-level situation and this model will thus be used to simulate the data. The graded response model is used to compute the probability of scoring at least $x \geq 1, 2, ..., z$ on item i for rater r in subject s according to discrimination parameter α_i, difficulty parameter β_{ix}, and the combination of the latent trait and rater deviation $\theta_{sr} = \theta_s + \delta_{sr}$,

resulting in

$$P(Y_{sri} \geq x|\theta_{sr}) = \frac{e^{[\alpha_i(\theta_{sr}-\beta ix)]}}{1+e^{[\alpha_i(\theta_{sr}-\beta ix)]}}. \tag{18}$$

The distributions of the latent trait values θ_s and random deviation per rater δ_{sr} are normal with mean 0 and variance depending on the condition, as described below. The item responses per rater are sampled from a multinomial distribution using the computed probabilities.

3.2 Design

3.2.1 Fixed variables

To keep the simulation study manageable, the number of subjects, the number of items, and item type are fixed:

- *Subjects*: The number of subjects S is set at 30, which is a realistic sample size for level 2 in various multilevel analyses. Although it has been shown that the number of subjects influence the magnitude and the sampling error of the two-level scalability coefficients (Crisan, 2015) and it had a small effect on the coverage of the one-level coefficients, it did not influence the bias of the standard errors in one-level coefficients (Kuijpers et al., 2016).

- *Items*: The number of items J is fixed at three. Although this is a small set of items, the H-values are a weighted mean of the pairwise H_{ij} coefficients for all item-pairs, and therefore are expected to give similar results on the outcome variables regardless of J. Previous simulation studies indicated no effect of number of items J on the bias for the one-

level coefficients H, a small effect on the coverage, and did not or only marginally influenced the sampling error of the two-level scalability coefficients (Kuijpers et al., 2016; Crisan, 2015).

- *Item type*: The number of item categories is fixed at $z + 1 = 3$. This is a common value, for example with answer categories 'true', 'somewhat true', 'not true'. Although higher values of z are known to improve the coverage by reducing bias of the H-estimates, it had no effect on the bias of the standard errors (Kuijpers et al., 2016). In addition, z had no or a small effect on the magnitude and sampling error of two-level coefficients (Crisan, 2015).

- *Bootstrap samples*: The number of bootstrap samples is fixed at $B = 1000$. The bootstraps are balanced, ensuring that each observation occurs an equal number of times in the bootstrap. This can reduce the variance of the estimation, resulting in a more efficient estimator (Efron & Tibshirani, 1993; Chernick, 2008). For the individual level bootstrap a vector is created in which the total set of raters are numbered from 1 to R, which is replicated B times. This vector is randomly shuffled and formatted to an R by B matrix, after which the appropriate data rows are selected for the actual bootstrap sample. For the cluster level bootstrap not the raters but the subjects are numbered from 1 to S, again replicated B times and randomly shuffled in an S by B matrix. Finally, the raters belonging to the subjects are selected from the original data for the actual bootstrap sample.

3.2.2 Independent variables

The independent variables are the number of raters per subject R_s, ratio of within- to between-subject variance $\sigma^2_\theta/\sigma^2_\delta$, discrimination parameter α and difficulty parameter β. These variables are manipulated in as follows:

- *Number of raters*: This variable consists of three levels. The number of raters per subject is set equal at either $R_s = 5$ or $R_s = 15$, or ranges between $R_s = [5, 15]$ (sampled from a discrete uniform distribution). This will indicate whether the number of raters per subject as well as equality of that number has an effect on the outcome variables. The total number of raters R ranges between 150 and 450. Although there was no effect of sample size on the bias of standard errors in the one-level situation (Kuijpers et al., 2016), it is unclear what the effect is on the two-level coefficients, and more specifically how the rater-subject ratio affects the outcomes.

- *Ratio of within- to between-subject variance*: This variable consists of three variables and reflects the degree that item responses are determined by the subject (θ_s) and by the rater (δ_{sr}). In the first situation $\sigma^2_\theta = \sigma^2_\delta = 0.5$, indicating an equal influence on the item response. In the second situation $\sigma^2_\theta = 0.8$ and $\sigma^2_\delta = 0.2$, indicating a larger influence of the subject. In the third situation $\sigma^2_\theta = 0.2$ and $\sigma^2_\delta = 0.8$, indicating only a minor influence of the subject. In the last situation it is least appropriate to scale subjects on the average test scores, since the item responses reflect only to a small degree the θ value, while the rater effects are large. In other words, the rater disagreement is largest

in the last condition. The covariance between θ and δ is assumed zero, resulting in a variance $\sigma^2_{(\theta+\delta)}$ equal to 1 in all cases. Although the ratio of variances affect the magnitude and sampling error of the H estimates (Crisan, 2015), it is unclear whether there is any effect on the outcome variables.

- *Item discrimination*: This variable consists of three levels. The item discrimination parameter α is either kept constant across items with a high discrimination ($\alpha = 2$) or a low discrimination ($\alpha = 0.5$), or the discrimination varies across items with equidistant values in the interval $\alpha = [0.5, 2]$. Kuijpers et al. (2016) found no bias for various equal α values across items for the one-level situation, but failed to investigate a situation where α varied across items. When α varies, the item ordering depends on θ. Since computation of the Guttman errors and thus the H-values requires ordering of the item steps for the total sample, this might introduce bias. Parameter α did affect the magnitude of both two-level coefficients, and the sampling error of H^B.

- *Item difficulty*: This variable consists of two levels. The distance between item-steps reflects the difference in difficulty of an item. The larger the difference between item-steps, the larger the difference in item difficulty. The distance between item-steps was either large as reflected by equidistant difficulty values between $\beta = [-2, 2]$, or small with values ranging between $\beta = [-0.5, 0.5]$. There was no effect of item difficulty on the bias of the standard errors of the H-estimates in

the one-level situation (Kuijpers et al., 2016), although it had a small effect on the magnitude and sampling error of the two-level coefficients.

3.2.3 Dependent variables

The dependent variables in this study are bias and efficiency of the estimated standard errors (se), coverage of the 95% confidence intervals, and computation time. These outcome variables are computed for the three methods to estimate the standard errors, that is, the delta method, the individual level bootstrap, and cluster level bootstrap. Computations are performed for the total scale coefficients H^W, H^B, and the ratio H^B/H^W, and in general will be denoted by H.

- *Bias of the standard errors*: To determine the bias, a true value of the standard error need to be known. Since this value is unknown, it is estimated by the sampling variation of H across Q replications per condition, using the standard deviation (sd)

$$sd(\widehat{H}) = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}(\widehat{H}_q - \overline{H})^2}, \tag{19}$$

where

$$\overline{H} = \frac{1}{Q}\sum_{q=1}^{Q}\widehat{H}_q. \tag{20}$$

The bias for method m is then computed by

$$bias.se_m = \frac{1}{Q}\sum_{q=1}^{Q}[se_m(\widehat{H}_q) - sd(\widehat{H})] \tag{21}$$

- *Efficiency*: A measure for efficiency is the root mean square error (RMSE), which incorporates both the variance and the bias of the standard error estimator. The RMSE for method m is estimated by

$$\widehat{RMSE}_m = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}[se_m(\widehat{H}_q) - sd(\widehat{H})]^2}. \tag{22}$$

- *Coverage of confidence intervals*: To study the coverage of the 95% intervals for the three methods, population values H are computed by generating a finite population per condition. To generate such a population, the sample of subjects is set to $S = 200\,000$, consequently leading to a population size between $R = 1$ and 3 million raters, depending on R_s. Parametric confidence intervals will be computed for each replication q with $\widehat{H} \pm 1.96se$, where the coverage is calculated as the proportion of times the population value H falls in the confidence interval, which should reflect the confidence level. In addition, non-parametric confidence intervals are computed for the bootstrap samples with the 2.5 and 97.5 percentile as CI $= [p_{0.025}, p_{0.975}]$. This might reflect a more robust interval, although Van Onna (2004) showed for the one-level coefficients that there is hardly any difference between the non-parametric and the parametric interval due to normality of the H coefficients.

- *Computation time*: The computation time will be saved and averaged for each estimation method.

3.3 Statistical analyses

The data simulation is programmed in R and the statistical analyses will be performed in either R or SPSS. The two-level scalability coefficients are computed using the functions described in Koopman (2016). Code is available upon request from the author. The fully crossed study design results in 54 conditions. The number of replications per condition is $Q = 1000$. To improve computation time the data simulation and standard error computations are performed on a high performance computing cluster. Summary descriptives are computed and visualized for the outcome variables per scalability coefficient. In addition, bias and efficiency of the standard error estimates for the different H-coefficients will be subjected to statistical analyses to examine the effect of estimation method and the various conditions. Since the three methods of estimating the standard errors are nested within each replication (i.e., they are based on the same data sample) there is expected to be a strong dependence. Therefore, it is appropriate to consider repeated measures analysis of variance (RM-ANOVA). Since sample sizes are large ($Q = 1000$), significant results are selected on their effect size η_p^2 rather than the p-value of the F-test, where only medium and large effects are selected (i.e., $\eta_p^2 > 0.06$ and 0.14, respectively). The effects of interest concern the within-subject main effect of estimation method, as well as interactions between the estimation method and the independent variables. For completeness, the between-subject effects, thus the effects of the independent variables, will be discussed as well. Estimated averages for the relevant effects will gain insight in the effects. Since the coverage reflects

a proportion, an Agresti-Coull interval is constructed around the estimated coverage. Such an interval is appropriate since the replication size Q is large and the expected p of 0.95 is close to 1, making the interval more reliable than alternative intervals (Agresti & Coull, 1998).

4 Results

The simulation of $Q = 1000$ replications for 54 conditions resulted in 54000 rows in the final data frame. For 11 of these rows (0.02%) the cluster bootstrap could not estimate the standard errors for the between to within scalability ratio H^B/H^W. These missing values had in common that α was 0.5, indicating a low item discrimination, R_s was equal (either 5 or 15) and for all but one the item-step difficulty β ranged between -0.5 and 0.5, reflecting small item-step distance. As a result, the total number of replications for se(H^B/H^W) was limited to 53989. This section will discuss subsequently the descriptive and inferential results for the bias, efficiency, coverage, and practical performance of the different standard error estimation methods. For all RM-ANOVAs the sphericity assumption was violated. Since the estimated ε for each test was below 0.75, the Greenhouse-Geisser correction was used (e.g., Field, 2013). Relevant output of the RM-ANOVAs can be found in the Appendix.

4.1 Population values

The population values of the two-level scalability coefficients and their standard deviations (Equation 19) are displayed in Table 1. The results are in

line with Crisan (2015) for all coefficients. The magnitude of the within scalability coefficient H^W mainly depends on discrimination parameter α, being larger for higher discrimination values. Furthermore, H^W is slightly smaller for lower item-step distances. Similar effects are present for the between scalability coefficient H^B, as well as an effect of the ratio of between- to within-subject variance σ^2_θ, where the value is larger when there is a higher variance across subjects and thus a smaller variance among the raters within a subject. The ratio coefficient H^B/H^W only depends on the variance between and within the subjects. The standard deviations range between 0.023 and 0.063 for the within- and between-rater scalability coefficients, but can be much larger for the ratio between the coefficients, ranging between 0.088 and 18.150, where the larger values are present when R_s varies, $\sigma^2_\theta = 0.5$, $\alpha = 0.5$, and $\beta = 0.5$. This is an indication that the ratio coefficient is not stable across different samples.

4.2 Bias

Figure 2 displays the interaction of bias for the estimation methods with each independent variable. As is clear for se(H^W) and se(H^B), the bias is close to zero for all methods. Furthermore, there is a medium to strong negative correlation between bias and standard deviation in the population (r_{H^W} = -0.461, r_{H^B} = -0.626, for both $p < .001$). This means that the bias is on average lower (but more extreme) when the standard deviation is higher. For se(H^W) all methods are negatively biased ($\overline{bias}_{se(H^W)} = -0.006, s = 0.008, \min = -0.029, \max = 0.002$). For se($H^B$) the bias varies around 0, although the average bias is still negative ($\overline{bias}_{se(H^B)} =$

Table 1:
The standard deviations of population values of the scalability coefficients, with the H values between brackets, of the sampling distribution when $S = 30$, $J = 3$, $z + 1 = 3$, per level of the independent variables.

	$sd(H^W)$	$sd(H^B)$	$sd(H^B/H^W)$
$R_s = 5$	0.063(0.284)	0.052(0.140)	3.529(0.494)
$R_s = vary$	0.049(0.283)	0.040(0.140)	13.864(0.496)
$R_s = 15$	0.040(0.284)	0.034(0.140)	1.036(0.495)
$\sigma^2_\theta = 0.2$	0.047(0.284)	0.032(0.056)	2.500(0.197)
$\sigma^2_\theta = 0.5$	0.049(0.284)	0.044(0.140)	13.334(0.494)
$\sigma^2_\theta = 0.8$	0.056(0.284)	0.050(0.225)	2.594(0.795)
$\alpha = 0.5$	0.043(0.068)	0.023(0.034)	18.154(0.498)
$\alpha = vary$	0.053(0.236)	0.041(0.117)	0.186(0.496)
$\alpha = 2$	0.056(0.548)	0.062(0.270)	0.088(0.492)
$\beta = 2$	0.054(0.312)	0.045(0.156)	3.127(0.499)
$\beta = 0.5$	0.047(0.255)	0.039(0.124)	9.159(0.492)

$-0.003, s = 0.017, \min = -0.046, \max = 0.023$). The bias of the standard error of the ratio H^B/H^W appears to be highly unstable for most values ($\overline{bias}_{se(H^B/H^W)} = 4.097\text{e+}08, s = 8.596\text{e+}08$, min = -137.472, max = 1.781e+10).

4.2.1 Bias se(H^W)

The RM-ANOVA for bias of the standard errors for H^W revealed there was a large main effect of method and a large interaction effect of method with R_s (see the Appendix). This indicates that the effect of estimation method differs according to the levels of R_s. More specifically, Table 2 (visualized in Figure 2) contains the estimated means, and shows that the delta method

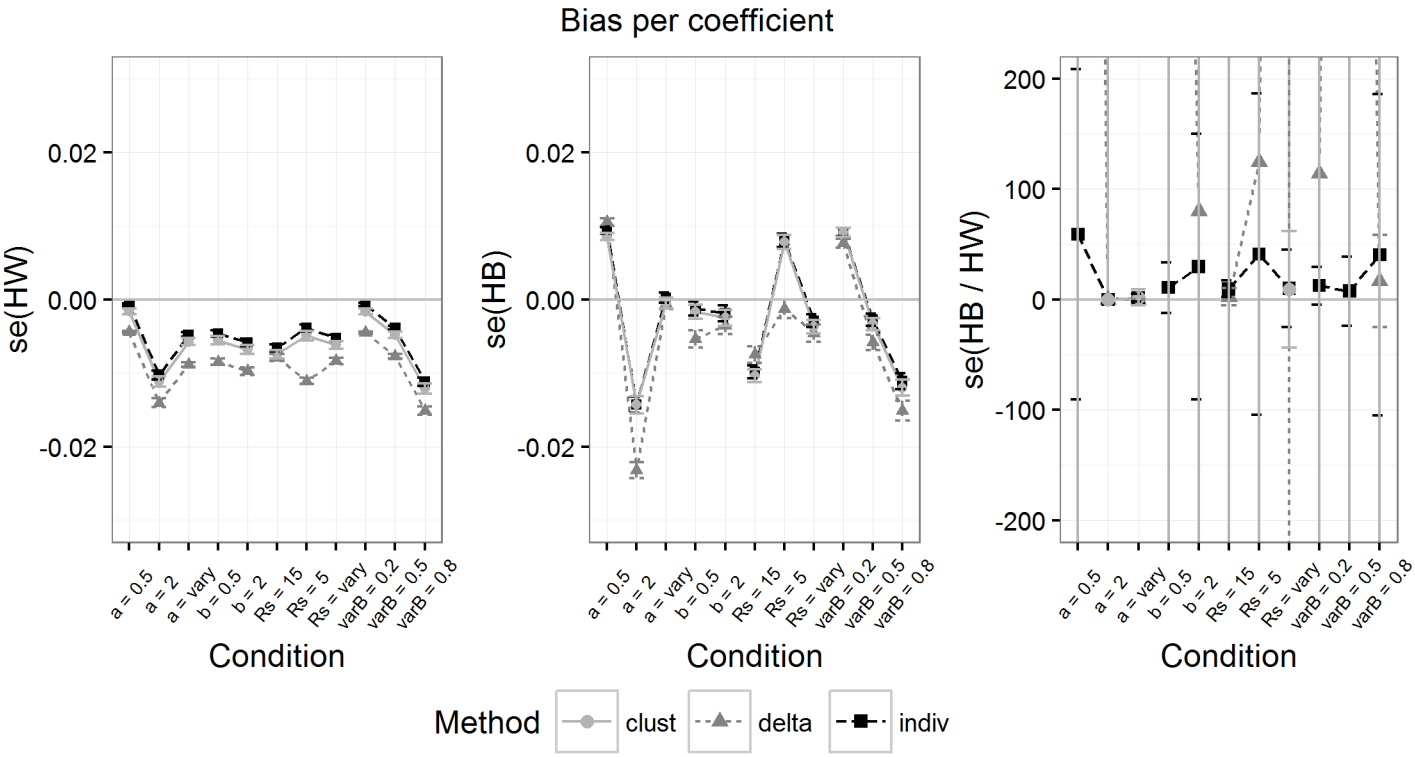


Figure 2: Average bias of the standard error estimation methods for discrimination parameter α (a), difficulty parameter β (b), raters per subject R_s, and between-subject variance σ_θ^2 (varB), per H-coefficient, with 95% confidence bars. Note that the scale of the y-axis for se(H^B/H^W) differs from the other two plots.

resulted especially in a more extreme bias when the number of raters was small, thus 5, whilst the bias of the bootstraps was slightly more extreme for more raters. When the number of raters was higher, the difference between the methods decreased. In general the bootstrap methods outperform the delta method. In addition, the individual level bootstrap performs slightly better than the cluster level bootstrap. The difference between the average bias ranged between 0.001 and 0.004, which is clear throughout the different

conditions, as was displayed in Figure 2.

Table 2:
The average bias for the interaction between the method and number of raters R_s for coefficient se(H^W).

Method	$R_s = 5$	$R_s =$ varies	$R_s = 15$	*Average*
Delta	-0.011	-0.008	-0.008	-0.009
Individual	-0.004	-0.005	-0.007	-0.005
Clustered	-0.005	-0.006	-0.007	-0.006

In general, regardless of estimation method for standard errors, the variance of within to between subject ratio σ_θ^2 and the item discrimination α had a large effect on the bias. In addition, there is a large interaction effect between these two variables. Specifically, the bias increases when the value for σ_θ^2 or α became larger. This effect is more pronounced in when they both became larger, see Table 3.

Table 3:
The average bias for the interaction between σ_θ^2 and α for coefficient se(H^W).

Condition	$\alpha = 0.5$	$\alpha =$ varies	$\alpha = 2$	*Average*
$\sigma_\theta^2 = 0.2$	-0.001	-0.003	-0.003	-0.002
$\sigma_\theta^2 = 0.5$	-0.002	-0.005	-0.009	-0.005
$\sigma_\theta^2 = 0.8$	-0.003	-0.012	-0.023	-0.013
Average	-0.002	-0.006	-0.012	

4.2.2 Bias se(H^B)

The RM-ANOVA for the standard error of coefficient H^B revealed a medium main effect of method and two large interaction effects, one of method with

R_s and one of method with α, which can also be seen in Figure 2. This indicates that the effect of estimation method depends both on R_s and α. Similarly to the bias of the standard errors for H^W, the difference between the estimation methods is largest for $R_s = 5$ and for $\alpha = 2$ (see Table 4). In contrast to se(H^W) the absolute bias is smaller for the delta method when the number of raters is equal. In general, the difference between the delta method and the bootstrap methods is smaller in comparison to the bias of se(H^W), yet still consistent throughout the conditions. Furthermore, no difference is present for the two bootstrap methods.

Table 4:
The average bias for the interaction between the estimation method and number of raters R_s or α for coefficient se(H^B).

Method	$R_s = 5$	$R_s =$ varies	$R_s = 15$	$\alpha = 0.5$	$\alpha =$ varies	$\alpha = 2$	*Average*
Delta	−0.001	−0.005	−0.007	0.011	−0.001	−0.023	−0.004
Individual	0.008	−0.003	−0.010	0.009	0.000	−0.014	−0.002
Clustered	0.008	−0.004	−0.010	0.009	0.000	−0.014	−0.002
Average	0.005	−0.004	−0.009	0.009	0.000	−0.017	

In general there were large effects on the bias of standard error of H^B for R_s, σ^2_θ, and α. In addition there was, similar to the average bias of se(H^W), a large interaction effect between σ^2_θ and α. According to the averages in Table 4, it appears that the negative bias becomes more extreme when the number of raters is larger. Regarding the interaction effect, Table 5 shows that the bias is positive when α and σ^2_θ are small, and negative when they are larger.

Table 5:
The average bias for the interaction between σ^2_θ and α for coefficient se(H^B).

Condition	$\alpha = 0.5$	$\alpha =$ varies	$\alpha = 2$	*Average*
$\sigma^2_\theta = 0.2$	0.013	0.010	0.003	0.009
$\sigma^2_\theta = 0.5$	0.009	−0.001	−0.020	−0.004
$\sigma^2_\theta = 0.8$	0.006	−0.010	−0.035	−0.013
Average	0.009	0.000	−0.017	

4.2.3 Bias se(H^B/H^W)

As is clear from Figure 2, the bias for the standard errors of the ratio H^B/H^W is for particular conditions very high and unstable. Specifically, for the delta method the bias ranges between -0.01 to 2215.00 ($\overline{bias}_{se_d(H^B/H^W)} =$ 738.00), for the individual level bootstrap between -0.01 and 59.14 ($\overline{bias}_{se_i(H^B/H^W)} =$ 20.29), and for the cluster bootstrap between 0.00 and 3.69e+09 ($\overline{bias}_{se_c(H^B/H^W)} =$ 1.23e+09). It appears that the individual level bootstrap shows the least variability. The RM-ANOVA showed no effect of method, neither of the other variables tested. Since all effect sizes were zero, the results are not displayed in the Appendix.

4.3 Efficiency

The RMSE, which represents the efficiency of an estimate, incorporates the variance and the bias of an estimate. It is therefore expected that a similar trend as in the bias of the methods is present. Figure 3 shows that for se(H^W) the difference between the two bootstrap methods increased, where the individual level bootstrap is more efficient than the cluster bootstrap,

while the difference between the cluster bootstrap and the delta method is less obvious in most conditions. For se(H^B) it seems as though the bootstrap estimates are slightly more efficient in comparison to the delta method, although the difference is not consistent. As was clear from the bias, the se(H^B/H^W) estimates are often not stable, resulting in large error bars of the RMSE estimate. In general $\overline{RMSE}_{se(H^W)} = 0.011$ ($s = 0.003$, min = 0.001, max = 0.029), $\overline{RMSE}_{se(H^B)} = 0.018$ ($s = 0.004$, min = 0.001, max = 0.046), and $\overline{RMSE}_{se(H^B/H^W)} = 6.046\text{e+}10$ ($s = 6.044\text{e+}10$, min = 0, max = 4.169e+11). Additionally, there is a medium to strong correlation between efficiency and the standard deviation in the population ($r_{H^W} = 0.473$, $r_{H^B} = 0.553$, for both $p < .001$). This means that the efficiency value is higher (thus less efficient) when the standard deviation is higher.

4.3.1 RMSE se(H^W)

The RM-ANOVA indicated similar results for the RMSE as the bias for se(H^W), which are a large effect of method and a medium interaction effect between method and R_s (see the Appendix). It appears that the delta method is least and the individual bootstrap most efficient when the number of raters is small (see Table 6). In addition, all three methods differ in their efficiency, with the individual level bootstrap being most efficient.

In general there were large effects of σ_θ^2 and α on bias, as well as a large interaction effect between the two, see Table 7. The efficiency of the estimates became poorer when σ_θ^2 was larger, especially when α was larger as well.

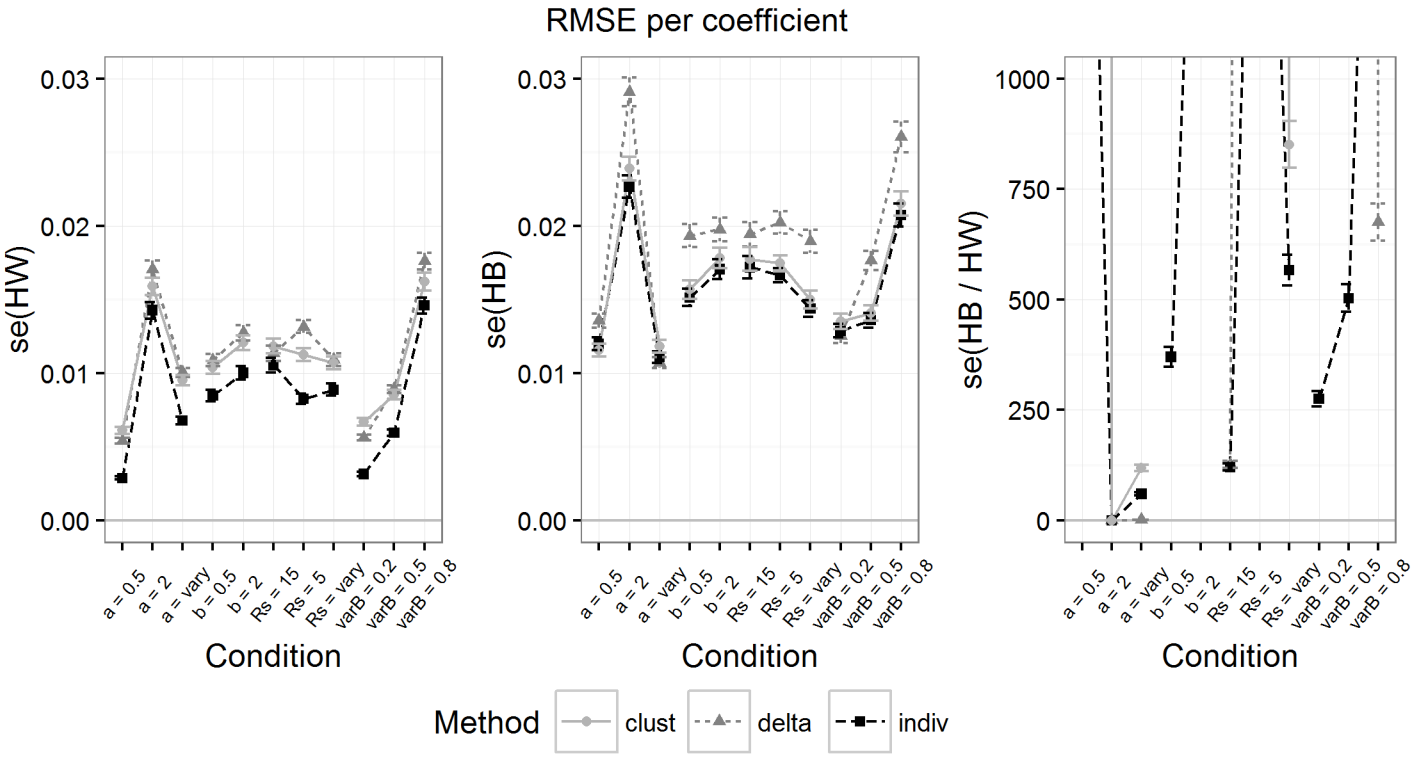


Figure 3: Average RMSE of the standard error estimation methods for discrimination parameter α (a), difficulty parameter β (b), raters per subject R_s, and between-subject variance σ^2_θ (varB), per H-coefficient, with 95% confidence bars. Note that the scale of the y-axis for se(H^B/H^W) differs again from the other two plots.

4.3.2 RMSE se(H^B)

The RM-ANOVA indicated a large main effect of method, a medium interaction effects of method with σ^2_θ and a large interaction effect of method with α. In addition, two three-way interaction effects were present, between method, R_s, and σ^2_θ and between method, σ^2_θ, and α (see the Appendix). To enhance interpretation, Figure 4 displays the three-way interaction plots. According to this Figure, the efficiency for the three methods differs espe-

Table 6:
The average RMSE for the interaction between the estimation method and number of raters R_s for coefficient se(H^W).

Method	$R_s = 5$	$R_s =$ varies	$R_s = 15$	*Average*
Delta	0.011	0.008	0.008	0.009
Individual	0.006	0.006	0.007	0.006
Clustered	0.009	0.008	0.009	0.008

Table 7:
The average bias for the interaction between σ_θ^2 and α for coefficient se(H^W).

Condition	$\alpha = 0.5$	$\alpha =$ varies	$\alpha = 2$	*Average*
$\sigma_\theta^2 = 0.2$	0.003	0.004	0.005	0.004
$\sigma_\theta^2 = 0.5$	0.004	0.006	0.010	0.006
$\sigma_\theta^2 = 0.8$	0.004	0.012	0.023	0.013
Average	0.004	0.007	0.013	

cially when R_s and σ_θ^2 were small, whereas they are more similar when R_s is larger. Furthermore, the delta method is least efficient method when R_s and α are at their largest values, but most efficient when R_s and α are small. The two bootstrap methods give similar results.

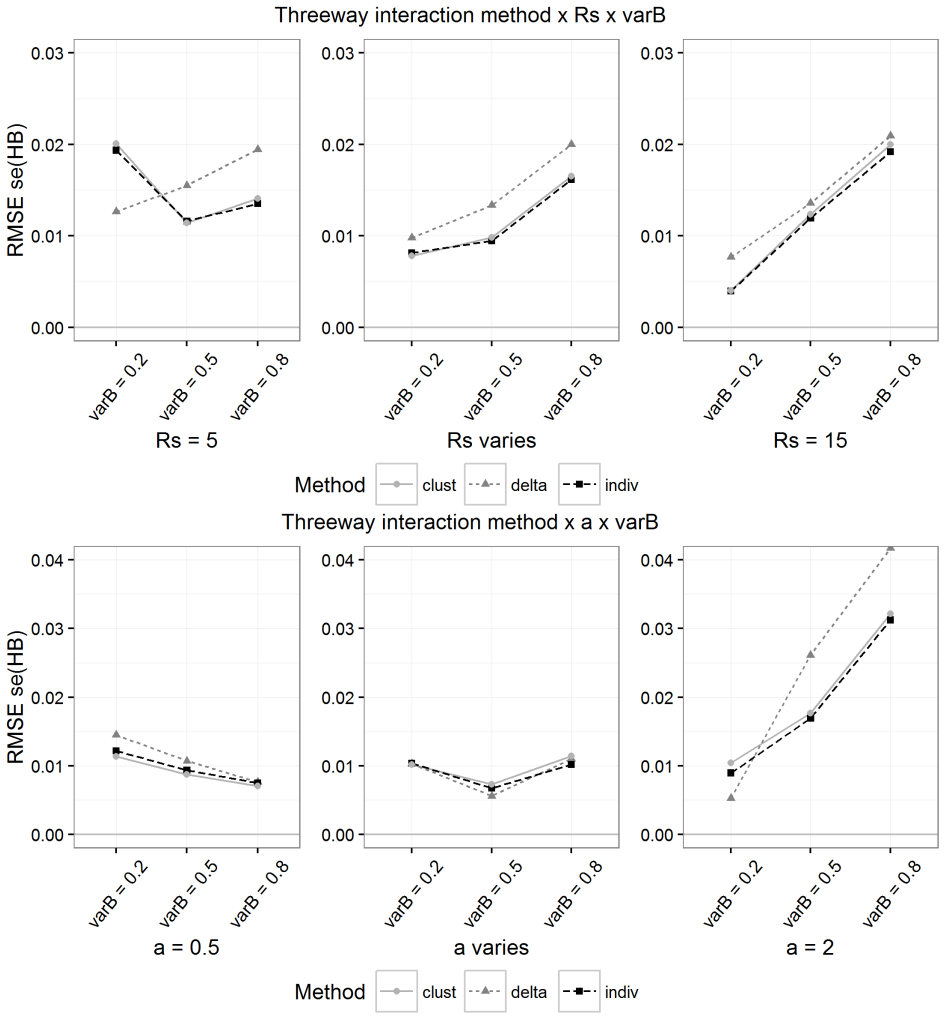


Figure 4: The three-way interactions effects between method, R_s, and σ^2_θ and method, σ^2_θ, and α.

Furthermore, there was a large effect of R_s and α on RMSE in general, as well as a large interaction effects between R_s and α, between R_s and σ^2_θ, and between σ^2_θ and α. Additionally there is a three-way interaction between R_s, σ^2_θ, and α. Table 8 displays the average RMSE for the three-way interaction, indicating that the efficiency is worst when σ^2_θ and α are large, regardless of

R_s, but when σ_θ^2 and α are smaller, than efficiency improves if R_s is larger.

Table 8:
The average RMSE for the three-way interaction between R_s, σ_θ^2, and α for coefficient se(H^B).

$\mathbf{R_s = 5}$	$\alpha = 0.5$	$\alpha =$ varies	$\alpha = 2$	*Average*
$\sigma_\theta^2 = 0.2$	0.019	0.018	0.015	0.017
$\sigma_\theta^2 = 0.5$	0.010	0.007	0.020	0.013
$\sigma_\theta^2 = 0.8$	0.007	0.011	0.035	0.016
Average	0.017	0.010	0.019	0.015
R_s varies	$\alpha = 0.5$	$\alpha =$ varies	$\alpha = 2$	*Average*
$\sigma_\theta^2 = 0.2$	0.012	0.009	0.004	0.009
$\sigma_\theta^2 = 0.5$	0.008	0.004	0.020	0.011
$\sigma_\theta^2 = 0.8$	0.006	0.011	0.036	0.017
Average	0.009	0.008	0.020	0.012
$\mathbf{R_s = 15}$	$\alpha = 0.5$	$\alpha =$ varies	$\alpha = 2$	*Average*
$\sigma_\theta^2 = 0.2$	0.007	0.004	0.005	0.005
$\sigma_\theta^2 = 0.5$	0.004	0.007	0.027	0.013
$\sigma_\theta^2 = 0.8$	0.003	0.016	0.042	0.020
Average	0.004	0.009	0.025	0.013

4.3.3 RMSE se(H^B/H^W)

Equal to the analysis of the bias statistics, there were no effects of the independent variables on the RMSE of the standard errors for coefficient H^B/H^W.

4.4 Coverage

Figure 5 visualizes the coverage per coefficient. In general the coverage for H^W is below the desired value of 0.95 ($\overline{cov}_{se(H^W)} = 0.900$), with the individual level estimates being closest to the desired coverage. The coverage of H^B varies more, but its average is equal to the coverage of se(H^W) ($\overline{cov}_{se(H^B)} = 0.900$), with the two bootstrap estimates being closer to the desired interval than the delta method. For the ratio se(H^B/H^W) the average coverage match the desired value ($\overline{cov}_{se(H^B/H^W)} = 0.948$), but it can still be too high or too low depending on the condition. For all coefficients, the coverage is worst when $\alpha = 2$, $R_s = 15$, or $\sigma^2_\theta = 0.8$. Contrastingly, the coverage is highest when $\alpha = 0.5$, $R_s = 5$, or $\sigma^2_\theta = 0.2$. As a comparison, during the simulation the non-parametric confidence intervals was computed for the bootstrap replications. For the H^W coefficient the coverage was 0.912 and 0.897 for the individual and cluster bootstrap, respectively, for the H^B coefficient these values were 0.896 and 0.890, and for the ratio H^B/H^W 0.933 and 0.924, lower values than the parametric interval estimates.

4.5 Practical performance

The computational time between the delta method and either bootstrap method differed for this simulation study on average by 2.42 minutes (about 200 times as long), where the bootstrap method lasts longer ($t(53999) =$ -545.91, $p < 0.001$). However, this simulation was quite small in setup; only three items in dataset are highly uncommon. As an example, on a regular laptop, for $S = 50$ subjects, $R_s = 20$ raters per subject, $J = 10$ items, and

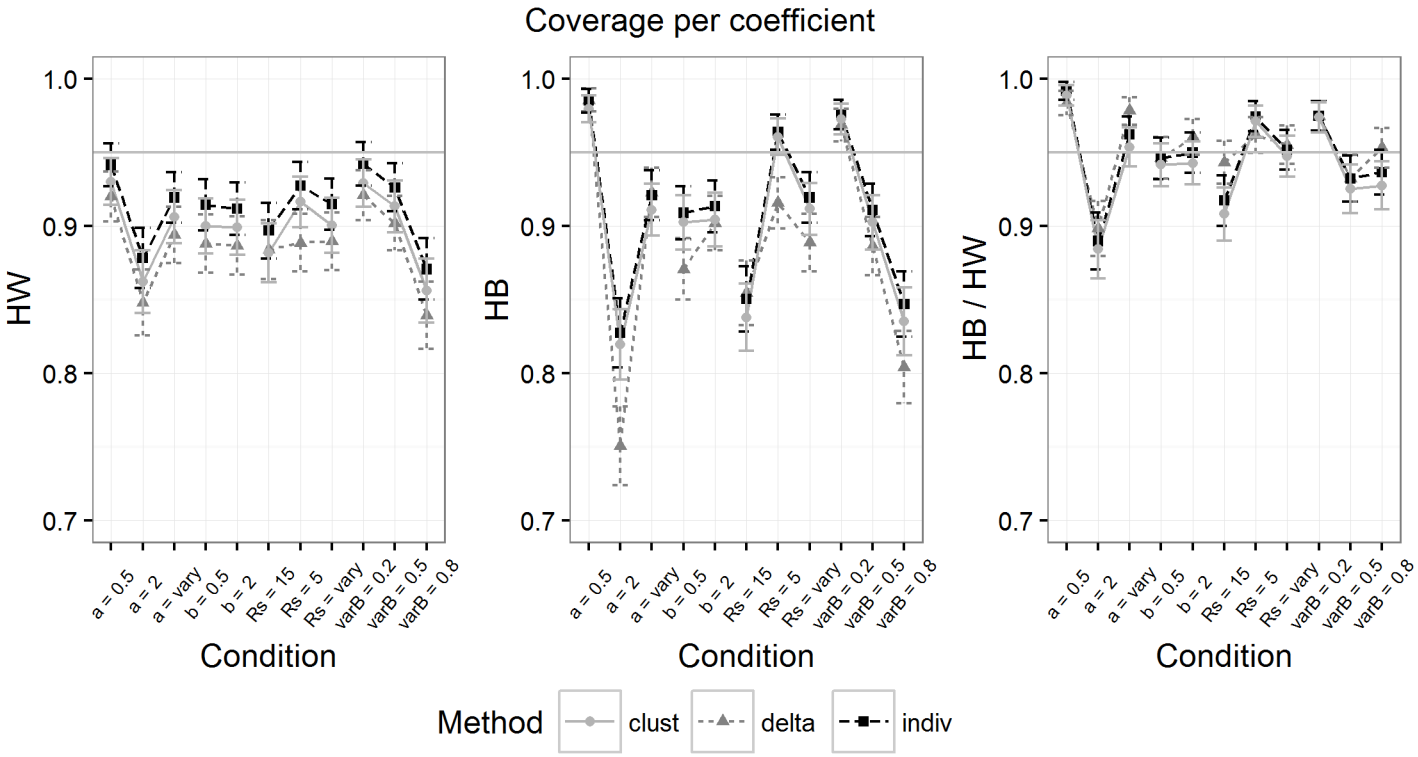


Figure 5: Coverage of the 95% confidence intervals for discrimination parameter α (a), difficulty parameter β (b), raters per subject R_s, and between-subject variance σ^2_θ (varB), per H coefficient, with 95% Agresti Coull confidence bars.

$z + 1 = 5$ answer categories, the duration of the delta method is about 1.3 minute and a bootstrap with $B = 1000$ about 4 hours, also approximately 200 times as long.

5 Discussion

The intention of this simulation study was to compare three methods to estimate the standard errors for two-level scalability coefficients on their bias,

efficiency, coverage, and computation time. Computation of the standard errors of the ratio coefficient H^B/H^W was problematic on rare occasions for the cluster bootstrap. Although the cause is uncertain, it is assumed that the estimated H^W coefficient was zero in at least one of the bootstrap samples, making it impossible to estimate the ratio H^B/H^W. It is expected that this does not cause any practical problems since it occurs only rarely and can be avoided by ignoring the missing values in the bootstrap samples when computing the standard error.

The results indicate that bias appears to be negligible for all conditions of the standard errors of both the H^W and the H^B scalability coefficient. Although the delta method is more (negatively) biased than the two bootstrap methods, the average difference is only 0.004. In addition, the difference between the delta method and bootstrap methods is especially present when the number of raters per subject is low ($R_s = 5$) and the item discrimination is equal and high ($\alpha = 2$). While for the bias there was a very small difference between the two bootstrap methods, it appears that the individual level bootstrap is more efficient in estimating the standard errors for the H^W coefficient in comparison to the cluster level bootstrap. However, the RMSE on average differs with 0.003, which can be considered small. For H^B the delta method performs slightly poorer compared to the bootstrap methods, especially when α is equal and high, compared to the bootstrap methods. The bootstrap methods themselves are highly similar.

Contradicting expectations, the cluster bootstrap did not outperform the individual level bootstrap, regardless of the nested structure of the data. Apparently, the dependency among raters within a subject remain sufficiently

in tact for the individual level bootstrap as compared to the cluster level bootstrap. This might be explained by the fact that rater deviations δ_{sr} are random deviations from θ_s, making it unnecessary to maintain the full rater sample per subject. In contrast, in longitudinal data a certain development is expected over time, observations that are closer together in time might have a higher correlation. When only a selection of time points are retained in the bootstrap sample, this development is less obvious, making the variability of the estimates too large. In current situation such a particular dependency is not present since the raters are independent conditional on θ_s, making it less relevant to preserve all raters per subject. As a result, the variance of the individual level bootstrap estimates is lower, especially for se(H^W), resulting in a more efficient estimate.

The standard errors of the ratio coefficient H^B/H^W appear to be highly unstable, the variability being smallest for the individual level bootstrap. This indicates that estimates of the ratio coefficient itself are highly unstable too, since small changes in H^W and H^B can result in large changes of their ratio, especially when H^W is small. Therefore, it is advised to first evaluate the the H^W and H^B coefficients, and if their values are sufficiently large the ratio coefficient can be interpreted to gain information on whether the item responses are mainly influenced by the subjects or the raters. The standard errors can be used to verify whether the values of the coefficients are significantly larger than the threshold set.

The coverage for both coefficients H^W and H^B was lower than the desired value of 95%. It appears that the difference between the methods where the cluster bootstrap was worse than the individual level bootstrap,

and the delta method was worse than the cluster bootstrap. This trend is most obvious for the H^W coefficient. For the H^B coefficient the delta method performs especially worse if the item discrimination α is large and equal, when the variance of the subjects (σ_θ^2) is large, and when the number of raters per subject R_s is small. These results are not all explained by the effects found in the bias of the standard errors, but might arise from biased estimates of the coefficients themselves, which has not been investigated as such. It is expected that the coverage improves when more polytomous items are used and more subjects and raters are included in the sample (Kuijpers et al., 2016). The coverage for the ratio se(H^B/H^W) reflects the desired value of 95% for most conditions. Since the estimates are for most of these conditions are unstable, the standard errors are very large and as a consequence so are the confidence intervals. This way, the population value is likely to fall in the interval.

The computation time is much larger for the bootstrap methods in comparison to the delta method. The number of bootstrap samples used in this study is $B = 1000$, which is not uncommon for a bootstrap method. The improvement in bias and efficiency of the scalability coefficients might be regarded insufficient to warrant the use of the bootstrap, especially since the non-parametric confidence intervals do not improve the coverage.

Limitations of this study include the fact that data is generated with the parametric graded response model. The results might be different when data is simulated according to a non-parametric model. In addition, various other variables might influence the results but are not taken into account in this study. For example, it is currently assumed that the latent trait of the

subject and the rater deviations are independent, that is, there is no relation between θ and δ. However, various scenarios exist where this assumption does not hold. For example, when subjects have extreme latent trait values, the raters might show more agreement on the item scores. Then the rater deviations show less variation when θ_s is either very high or very low. Also, other assumptions might be violated, such as having items with item-step response functions that decrease at some point or when item responses depend not only on θ and δ. Further research might focus on the effects of such manipulations on the standard errors of the two-level scalability coefficients.

Summarized, the three methods to estimate the standard errors for the two-level scalability coefficients differed only marginally in their bias, efficiency, and coverage, with a small preference for the individual level bootstrap. However the bootstrap method quickly becomes quite computer intensive, up to hours of computation time for medium to large datasets. Unless the computation time of the two-level H-coefficients will be improved, this method is of little practical value. Therefore, the delta method is currently advised to use as method for standard error estimation of the two-level scalability coefficients.

Appendix: Output RM-ANOVA

Table 9:
Results of the full-factorial RM-ANOVA for bias of se(H^W), with the degrees of freedom (df), F-value, right-tail probability p, and effect size value η_p^2.

Within effect	*df*	*F*	*p*	η_p^2
method	1.06	19693.42	<0.01	**0.267**
method x R_s	2.11	4759.54	<0.01	**0.150**
method x σ_θ^2	2.11	0.98	0.38	<0.001
method x α	2.11	13.84	<0.01	0.001
method x β	1.06	2.73	0.07	<0.001
method x R_s x σ_θ^2	4.22	0.92	0.46	<0.001
method x R_s x α	4.22	4.22	<0.01	<0.001
method x R_s x β	2.11	2.22	0.11	<0.001
method x σ_θ^2 x α	4.22	0.37	0.84	<0.001
method x σ_θ^2 x β	2.11	0.92	0.40	<0.001
method x α x β	2.11	34.04	<0.01	0.001
method x R_s x σ_θ^2 x α	8.45	1.09	0.37	<0.001
method x R_s x σ_θ^2 x β	4.22	0.69	0.61	<0.001
method x R_s x α x β	4.22	4.40	<0.01	<0.001
method x σ_θ^2 x α x β	4.22	0.41	0.81	<0.001
method x R_s x σ_θ^2 x α x β	8.45	0.66	0.74	<0.001
Error	56945.37			
Between effect				
Intercept	1.00	212242.75	<0.01	**0.797**
R_s	2.00	230.59	<0.01	0.008
σ_θ^2	2.00	44066.52	<0.01	**0.620**
α	2.00	34902.78	<0.01	**0.564**
β	1.00	1712.71	<0.01	0.031
R_s x σ_θ^2	4.00	478.22	<0.01	0.034
R_s x α	4.00	377.06	<0.01	0.027
R_s x β	2.00	384.91	<0.01	0.014
σ_θ^2 x α	4.00	12178.95	<0.01	**0.475**
σ_θ^2 x β	2.00	619.69	<0.01	0.022
α x β	2.00	58.63	<0.01	0.002
R_s x σ_θ^2 x α	8.00	190.90	<0.01	0.028
R_s x σ_θ^2 x β	4.00	147.73	<0.01	0.011
R_s x α x β	4.00	50.86	<0.01	0.004
σ_θ^2 x α x β	4.00	206.79	<0.01	0.015
R_s x σ_θ^2 x α x β	8.00	56.89	<0.01	0.008
Error	53946.00			

Note: Letter x denotes an interaction effect. Medium and large η_p^2 values are printed in boldface.

Table 10:
Results of the full-factorial RM-ANOVA for bias of se(H^B), with the degrees of freedom (df), F-value, right-tail probability p, and effect size value η_p^2.

Within effect	*df*	*F*	*p*	η_p^2
method	1.50	8034.01	<0.01	**0.130**
method x R_s	2.99	13664.54	<0.01	**0.336**
method x σ_θ^2	2.99	447.38	<0.01	0.016
method x α	2.99	11692.36	<0.01	**0.302**
method x β	1.50	1860.46	<0.01	0.033
method x R_s x σ_θ^2	5.98	153.20	<0.01	0.011
method x R_s x α	5.98	444.00	<0.01	0.032
method x R_s x β	2.99	6.308	<0.01	<0.001
method x σ_θ^2 x α	5.98	34.90	<0.01	0.003
method x σ_θ^2 x β	2.99	46.20	<0.01	0.002
method x α x β	2.99	350.32	<0.01	0.013
method x R_s x σ_θ^2 x α	11.97	38.24	<0.01	0.006
method x R_s x σ_θ^2 x β	5.98	2.46	0.02	<0.001
method x R_s x α x β	5.98	17.81	<0.01	0.001
method x σ_θ^2 x α x β	5.98	14.18	<0.01	0.001
method x R_s x σ_θ^2 x α x β	11.97	0.90	0.54	<0.001
Error	80681.77			
Between effect				
Intercept	1.00	29898.88	<0.01	**0.357**
R_s	2.00	69828.74	<0.01	**0.721**
σ_θ^2	2.00	159744.04	<0.01	**0.856**
α	2.00	253452.71	<0.01	**0.904**
β	1.00	19.03	<0.01	<0.001
R_s x σ_θ^2	4.00	80.08	<0.01	0.006
R_s x α	4.00	163.60	<0.01	0.012
R_s x β	2.00	81.16	<0.01	0.003
σ_θ^2 x α	4.00	30379.78	<0.01	**0.693**
σ_θ^2 x β	2.00	1527.04	<0.01	0.054
α x β	2.00	575.77	<0.01	0.021
R_s x σ_θ^2 x α	8.00	127.39	<0.01	0.019
R_s x σ_θ^2 x β	4.00	233.37	<0.01	0.017
R_s x α x β	4.00	21.23	<0.01	0.002
σ_θ^2 x α x β	4.00	409.93	<0.01	0.029
R_s x σ_θ^2 x α x β	8.00	111.04	<0.01	0.016
Error	53946.00			

Note: Letter x denotes an interaction effect. Medium and large η_p^2 values are printed in boldface.

Table 11:
Results of the full-factorial RM-ANOVA for MSE of se(H^W), with the degrees of freedom (df), F-value, right-tail probability p, and effect size value η_p^2.

Within effect	*df*	*F*	*p*	η_p^2
method	1.16	7.45e-5	<0.01	**0.146**
method x R_s	2.32	1.64e-5	<0.01	**0.070**
method x σ_θ^2	2.32	1.30e-5	<0.01	0.056
method x α	2.32	8.93e-5	<0.01	0.039
method x β	1.16	1.39e-6	<0.01	0.003
method x R_s x σ_θ^2	4.63	2.53e-6	<0.01	0.023
method x R_s x α	4.63	1.88e-6	<0.01	0.017
method x R_s x β	2.32	7.46e-7	<0.01	0.003
method x σ_θ^2 x α	4.63	3.42e-6	<0.01	0.030
method x σ_θ^2 x β	2.32	6.80e-7	<0.01	0.003
method x α x β	2.32	2.17e-8	<0.01	<0.001
method x R_s x σ_θ^2 x α	9.26	7.85e-7	<0.01	0.014
method x R_s x σ_θ^2 x β	4.63	2.28e-7	<0.01	0.002
method x R_s x α x β	4.63	2.11e-8	0.03	<0.001
method x σ_θ^2 x α x β	4.63	2.04e-7	<0.01	0.002
method x R_s x σ_θ^2 x α x β	9.26	4.88e-8	<0.01	0.001
Error	62456.78			
Between effect				
Intercept	1.00	101037.97	<0.01	**0.652**
R_s	2.00	352.87	<0.01	0.013
σ_θ^2	2.00	38561.77	<0.01	**0.588**
α	2.00	33188.16	<0.01	**0.552**
β	1.00	2463.79	<0.01	0.044
R_s x σ_θ^2	4.00	872.69	<0.01	**0.061**
R_s x α	4.00	654.81	<0.01	0.046
R_s x β	2.00	258.91	<0.01	0.010
σ_θ^2 x α	4.00	19596.07	<0.01	**0.592**
σ_θ^2 x β	2.00	1155.22	<0.01	0.041
α x β	2.00	616.80	<0.01	0.022
R_s x σ_θ^2 x α	8.00	585.58	<0.01	**0.080**
R_s x σ_θ^2 x β	4.00	137.32	<0.01	0.010
R_s x α x β	4.00	54.45	<0.01	0.004
σ_θ^2 x α x β	4.00	419.83	<0.01	0.030
R_s x σ_θ^2 x α x β	8.00	40.30	<0.01	0.006
Error	53946.00			

Note: Letter x denotes an interaction effect. Medium and large η_p^2 values are printed in boldface.

Table 12:
Results of the full-factorial RM-ANOVA for RMSE of se(H^B), with the degrees of freedom (df), F-value, right-tail probability p, and effect size value η_p^2.

Within effect	*df*	*F*	*p*	η_p^2
method	1.48	10659.52	<0.01	**0.165**
method x R_s	2.97	362.56	<0.01	0.013
method x σ_θ^2	2.97	4362.55	<0.01	**0.139**
method x α	2.97	8383.72	<0.01	**0.237**
method x β	1.48	532.43	<0.01	0.010
method x R_s x σ_θ^2	5.93	1597.40	<0.01	**0.106**
method x R_s x α	5.93	809.83	<0.01	0.057
method x R_s x β	2.97	14.25	<0.01	0.001
method x σ_θ^2 x α	5.93	5480.45	<0.01	**0.289**
method x σ_θ^2 x β	2.97	35.24	<0.01	0.001
method x α x β	2.97	84.67	<0.01	0.003
method x R_s x σ_θ^2 x α	11.86	530.06	<0.01	**0.073**
method x R_s x σ_θ^2 x β	5.93	41.41	<0.01	0.003
method x R_s x α x β	5.93	76.78	<0.01	0.006
method x σ_θ^2 x α x β	5.93	61.37	<0.01	0.005
method x R_s x σ_θ^2 x α x β	11.86	13.19	<0.01	0.002
Error	79962.74			
Between effect				
Intercept	1.00	260302.56	<0.01	**0.828**
R_s	2.00	1342.40	<0.01	0.047
σ_θ^2	2.00	32798.25	<0.01	**0.549**
α	2.00	77672.70	<0.01	**0.742**
β	1.00	1723.58	<0.01	0.031
R_s x σ_θ^2	4.00	7107.45	<0.01	**0.345**
R_s x α	4.00	7742.43	<0.01	**0.365**
R_s x β	2.00	48.66	<0.01	0.002
σ_θ^2 x α	4.00	42297.31	<0.01	**0.758**
σ_θ^2 x β	2.00	173.41	<0.01	0.006
α x β	2.00	1782.96	<0.01	**0.062**
R_s x σ_θ^2 x α	8.00	1649.67	<0.01	**0.197**
R_s x σ_θ^2 x β	4.00	48.97	<0.01	0.004
R_s x α x β	4.00	138.78	<0.01	0.010
σ_θ^2 x α x β	4.00	650.55	<0.01	0.046
R_s x σ_θ^2 x α x β	8.00	71.02	<0.01	0.010
Error	53946.00			

Note: Letter x denotes an interaction effect. Medium and large η_p^2 values are printed in boldface.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: 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

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 & Sons.

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

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, & M. Wiberg (Eds.), *Quantitative psychology research.* New York, NY: Springer.

Deen, M., & De Rooij, M. J. (2016). *A comparison of mixed models and linear regression with cluster bootstrap for longitudinal data: two monte carlo experiments.* (Manuscript submitted for publication)

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

Field, A. (2013). *Discovering statistics using ibm spss statistics.* London: Sage.

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, V. E. C. (2016). *Standard errors of scalability coefficients in two-level mokken scale analysis* (Unpublished master's thesis). University of Amsterdam, Amsterdam, The Netherlands.

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 mokkens scalability coefficients. *Applied Psychological Measurement*, *40*, 331-345. doi: 10.1177/0146621616638500

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507-529. doi: 10.1037/h0055827

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

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

Samejima, F. (1969). *Estimation of latent ability using a response pat-*

tern of graded scores. (Psychometrika monograph supplement No. 17). Richmond, VA: Psychometric Society. Retrieved from `http://www.psychometrika.org/journal/online/MN17.pdf`

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

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* (p. 319-338). New York, NY: Springer. doi: 10.1007/978-1-4613-0169-1_17

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

Weisstein, E. W. (2011). *Euler's homogeneous function theorem.* Retrieved June 12, 2016, from `http://mathworld.wolfram.com/EulersHomogeneousFunctionTheorem.html`