Nonparametric item response theory for multilevel test data

Koopman, L.

Publication date
2022

Chapter 5

Range-Preserving Confidence Intervals and Significance Tests for Scalability Coefficients in Mokken Scale Analysis

Abstract

Mokken's scalability coefficients take values on the interval $(-\infty, 1]$. The sampling distribution of scalability coefficients is skewed near the boundary, so Wald-based confidence intervals and significance tests may be biased. We introduce a transformation of the scalability coefficients and their standard errors, which can be used to construct range-preserving confidence intervals and significance tests. We demonstrated that for scalability coefficients away from the boundary, the properties of this range-preserving method are similar to the properties of the Wald-based method, but the range-preserving method outperforms the Wald-based method if the coefficient is close to unity. The range-preserving method can be applied to all types of scalability coefficients, in nonclustered and clustered data. Its implementation in software is discussed.

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

5.1 Introduction

Mokken scale analysis is a popular scaling method used in questionnaires and is based on nonparametric item response theory models (see, e.g., Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017, for an elaborate introduction). The most popular aspect of Mokken scale analysis is scalability coefficients, which can be used to construct questionnaires from a larger set of items or to evaluate questionnaires that have a fixed set of items (Sijtsma & Van der Ark, 2017). Let I denote the total number of items, indexed by i or j $(i, j = 1, 2, \ldots, I)$. There are three types of scalability coefficients: item-pair scalability coefficient H_{ij} is a normed correlation between items i and j, item scalability coefficient H_i is a normed item–rest correlation that can be considered a discrimination index, and total-scale coefficient H is the weighted sum of the H_is across all items, for which higher values indicate a more accurate ordering of respondents (e.g., Sijtsma & Molenaar, 2016, p. 309). The standard errors of the three types of scalability coefficients were derived using the delta method as $SE_{H_{ij}}$, SE_{H_i} and SE_H, respectively (Kuijpers et al., 2013). Snijders (2001a) generalized the coefficients to two-level scalability coefficients for multi-rater data, in which multiple raters score the subjects of interest. Two-level scalability coefficients consist of within-rater and between-rater coefficients, which provide information on the scalability on the respondent- and the group-level, respectively (see also, Koopman, Zijlstra, & Van der Ark, 2020). Within-rater coefficients have a similar interpretation to Mokken's coefficients.

A Mokken scale is defined as a set of items for which

$$\begin{array}{lll} H_{ij} & > 0 & \text{for all item-pairs } (i, j), \\ H_i & \geq c > 0 & \text{for all items } i, \end{array} \tag{5.1}$$

where c is some positive lower bound for which $c = .3$ is often used (Mokken, 1971, p. 184). All scalability coefficients can take values from $-\infty$ to 1. If the items are statistically independent, the scalability coefficients equal 0; if the items are perfectly correlated, the scalability coefficients equal 1. The strength of a scale can be classified as follows:

$$\begin{array}{l} .3 \leq H < .4 \text{ weak scale}, \\ .4 \leq H < .5 \text{ medium scale}, \\ .5 \leq H \text{ strong scale}. \end{array} \tag{5.2}$$

For more information on suggested thresholds for two-level scalability coefficients, see Snijders (2001a). The actual minimum of scalability coefficients depends on the marginal frequencies (Sijtsma & Molenaar, 2002, p. 59). Away from the boundary, the sampling distribution of scalability coefficients is approximately normal (Koopman, Zijlstra, & Van der Ark, 2020; Mokken, 1971, pp. 166–167), but if a coefficient is close to the boundary or the SE is large, the sampling distribution is skewed to the left.

The point estimates of a scalability coefficient and its SE in sample data can be

combined by a normal approximation Wald-based confidence interval or significance test (Koopman, Zijlstra, & Van der Ark, 2020; Kuijpers et al., 2013). Two-sided confidence intervals are useful to determine the strength of total-scale coefficient H with confidence (Eq. 5.2), whereas one-sided significance tests are useful to test the two criteria of a Mokken scale (Eq. 5.1; Koopman et al., 2022). If the sampling distribution of the scalability coefficients is skewed, Wald-based confidence intervals and significance tests may be biased. This can result in deteriorated coverage of the confidence interval, inclusion of values larger than 1 in the confidence interval, and inflated Type I error rates of the significance tests. In this chapter, we propose a range-preserving confidence interval and significance test using a logarithmic transformation that can be applied to all scalability coefficients, both in nonclustered data (i.e., obtained by a simple random sampling design) and clustered data (i.e., obtained by a cluster sampling design). We compare the performance of the Wald-based and range-preserving methods in terms of coverage and Type I error rate using simulated data. Applications of the range-preserving methods in software are demonstrated.

5.2 Sampling Distribution of Scalability Coefficients

The sampling distribution of both Mokken's and Snijders' scalability coefficients are asymptotically normal (Mokken, 1971, pp. 166-167; Koopman, Zijlstra, & Van der Ark, 2020, respectively). Therefore, it is common practice to use normal-theory approaches to confidence interval estimation and significance testing. Let $\widehat{H}$ denote the point estimate of H with standard error $SE_{\widehat{H}}$. Figure 5.1 shows six histograms of the empirical sampling distribution of $\widehat{H}$ for a range of population values for H, created with 10,000 simulated datasets using 100 respondents and 10 dichotomous items. For H away from the boundary of 1, the distribution is approximately normal (as is expected according to asymptotic theory), but as H comes closer to the boundary, the distribution becomes increasingly skewed. For skewed sampling distributions, normal-theory approaches may be biased, in which case, range-preserving approaches are desirable because they only take values on the possible range of the coefficient and tend to be more accurate and reliable (Efron & Tibshirani, 1993, Section 13.7).

Confidence interval and significance tests can be applied to the scalability coefficients by using point estimates of the item-pair coefficients $\widehat{H}_{ij}$, item coefficients $\widehat{H}_i$, and total-scale coefficient $\widehat{H}$, along with $SE_{\widehat{H}_{ij}}$, $SE_{\widehat{H}_i}$, and $SE_{\widehat{H}}$, respectively. Two-sided confidence intervals of H are appropriate to estimate whether a scale is weak, medium, or strong (Eq. 5.2). One-sided significance tests (or one-sided confidence intervals) are appropriate to evaluate the two criteria of a Mokken scale (Eq. 5.1; Koopman et al., 2022). For the first criterion, the null hypothesis is $H_{ij} = 0$ and the alternative hypothesis is $H_{ij} > 0$ for each item pair (i, j). For the second criterion, the null hypothesis is $H_i = c$ and the alternative hypothesis is $H_i > c$ for each item i.

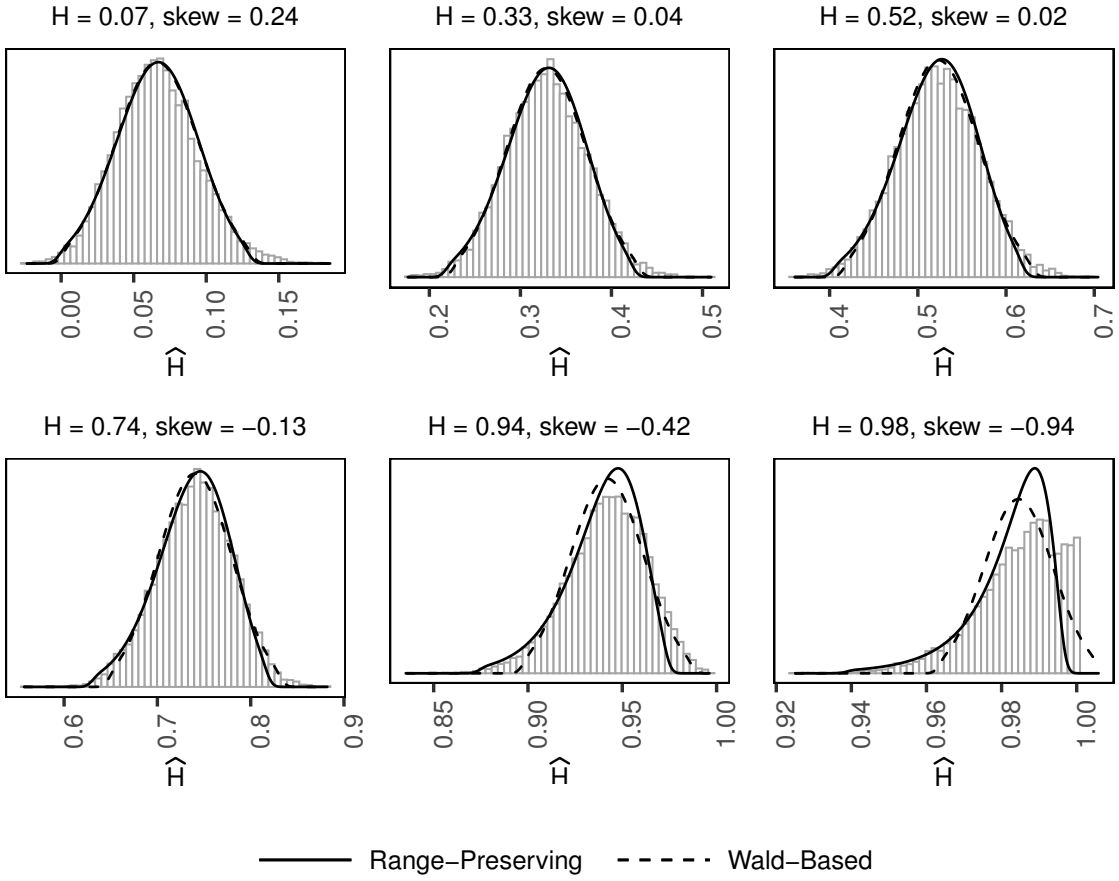

Figure 5.1: Six empirical distributions of total-scale coefficient H, with Wald-based (dashed line) and range-preserving (solid line) approximations of the sampling distribution based on the average $\widehat{H}$ and $SE^2_{\widehat{H}}$ across the datasets. The distribution is based on 10,000 simulated datasets with 10 dichotomous items and 100 respondents.

5.2.1 Wald-Based Methods

Wald-based methods assume a normal sampling distribution. A two-sided confidence interval contains two confidence limits. Let $z_{\alpha/2}$ denote the z score pertaining to significance level $\alpha/2$. Then, the two-sided $(1-\alpha)\times 100\%$ Wald-based confidence interval (denoted CI) is computed as

$$\text{CI} = \widehat{H} \pm z_{\alpha/2} \times SE_{\widehat{H}}. \tag{5.3}$$

Consider a two-sided 95% CI, $z_{\alpha/2} \approx 1.96$. Note that the upper confidence limit may exceed the boundary of 1, which is the maximum value of H. One-sided CIs also exist and can be constructed by replacing $z_{\alpha/2}$ in Eq. 5.3 with z_α and by selecting the confidence limit of interest, which is the lower limit for H_{ij} and H_i. For a one-sided 95% CI $z_\alpha \approx 1.645$.

The Wald-based significance test is a z test to standardize the difference between $\widehat{H}$ and the value of H under the null-hypothesis to a z score. For example, using the null

hypothesis $H = c$, z is computed as

$$z = \frac{\widehat{H} - c}{SE_{\widehat{H}}}. \tag{5.4}$$

The corresponding one-sided p value can be found in the standard normal z table.

A problem with the Wald-based method is that the sampling distribution is skewed for very high values of H or SE, in which case the results cannot be trusted.

5.2.2 Range-Preserving Methods

A confidence interval is range-preserving if its values are in the possible range of the parameter of interest. We propose a strategy to compute a range-preserving interval and to apply a similar strategy to compute a z score, which we collectively refer to as range-preserving methods. Range-preserving methods also apply asymptotic normal theory, but rather than using the original estimate $\widehat{H}$, which is bounded by 1, confidence interval and z scores are computed using a transformation of $\widehat{H}$ and its SE.

Let $g(\widehat{H})$ denote the transformation of $\widehat{H}$, and let $\log(x)$ denote the natural logarithm of x,

$$g(\widehat{H}) = -\log(1 - \widehat{H}). \tag{5.5}$$

The range for the transformed scalability coefficient is the real space $(-\infty, \infty)$. Let $g^{-1}(\widehat{H})$ denote inverse of $g(\widehat{H})$, and let $\exp(x)$ denote the exponential of x

$$g^{-1}(g(\widehat{H})) = 1 - \exp(-g(\widehat{H})) = \widehat{H}. \tag{5.6}$$

Let $g'(\widehat{H})$ denote the first derivative of $g(\widehat{H})$ with respect to $\widehat{H}$. By the chain rule (Stewart, 2008, p. 197),

$$g'(\widehat{H}) = \frac{d}{d\widehat{H}}\, g(\widehat{H}) = \frac{1}{1 - \widehat{H}}. \tag{5.7}$$

Using the delta method (Agresti, 2012, pp. 577–594), the SE of $g(\widehat{H})$, $SE_{g(\widehat{H})}$, is then approximated as

$$\begin{aligned} SE_{g(\widehat{H})} &\approx \sqrt{[g'(\widehat{H})]^2 SE_{\widehat{H}}^2} \\ &= SE_{\widehat{H}}/(1 - \widehat{H}). \end{aligned} \tag{5.8}$$

To obtain the range-preserving confidence interval (denoted CI*), we first construct a Wald-based confidence interval using the result of Eqs. 5.5 and 5.8,

$$\begin{aligned} \mathrm{CI}_{g(\widehat{H})} &= g(\widehat{H}) \pm z_{\alpha/2} \times SE_{g(\widehat{H})} \\ &= -\log(1 - \widehat{H}) - z_{\alpha/2} \times SE_{\widehat{H}}/(1 - \widehat{H}). \end{aligned} \tag{5.9}$$

Then, this interval is transformed back to the original scale of H, which reflects the

range-preserving confidence interval:

$$\begin{aligned} \mathrm{CI}^* &= 1 - \exp(-CI_{g(\widehat{H})}) \\ &= 1 - \exp(\log(1-\widehat{H}) \pm z_{\alpha/2} \times SE_{\widehat{H}}/(1-\widehat{H})). \end{aligned} \tag{5.10}$$

The range-preserving z score (denoted z^*) is computed by transforming both $\widehat{H}$ and c,

$$z^* = \frac{g(\widehat{H}) - g(c)}{SE_{g(\widehat{H})}}. \tag{5.11}$$

If $\widehat{H} = 1$, then the SE is estimated as $SE_{\widehat{H}} = 0$, resulting in $g(\widehat{H}) = \infty$ and an undefined $SE_{g(\widehat{H})}$, z, and z^*. In that case, we define CI^* as $[1, 1]$ and evaluate z and z^* as significant.

Similarly, confidence intervals and z scores can be computed for item pairs as CI_{ij} and z_{ij} and for items as CI_i and z_i, with superscript * for range-preserving results, by replacing $\widehat{H}$ and $SE_{\widehat{H}}$ in Eqs. 5.3, 5.4, 5.10, and 5.11 with $\widehat{H}_{ij}$ and $SE_{\widehat{H}_{ij}}$ or with $\widehat{H}_i$ and $SE_{\widehat{H}_i}$ respectively.

Multivariate Case. The range-preserving transformation is easily generalized to the multivariate case, which is useful to, for example, construct a variance–covariance matrix for a set of transformed item-pair or item coefficients. Let $\mathbf{H} = [H_{(1)}, H_{(2)}, \ldots, H_{(k)}, \ldots, H_{(K)}]^T$ denote a transposed vector containing K scalability coefficients $H_{(k)}$, $(k = 1, 2, \ldots, K)$. The transformation of $\mathbf{H}$ is

$$g(\mathbf{H}) = [g(H_{(1)}), g(H_{(2)}), \ldots, g(H_{(k)}), \ldots, g(H_{(K)})]^T. \tag{5.12}$$

Let $\mathbf{G} = \frac{\partial g(\mathbf{H})}{\partial \mathbf{H}^T}$ be the Jacobian of $g(\mathbf{H})$, that is, the matrix of first-order partial derivatives with respect to $\mathbf{H}$. Let $\bigoplus$ indicate the direct sum. For $g(\mathbf{H})$,

$$\mathbf{G} = \bigoplus_{1}^{K} g(\mathbf{H}). \tag{5.13}$$

$\mathbf{G}$ is a diagonal matrix with the first derivative of $g(H_{(k)})$ (Eq. 5.7) on the kth diagonal element and zero on the off-diagonal elements. Let $\mathbf{V_H}$ denote the variance–covariance of $\mathbf{H}$, $V_{(k)}$ the variance of $H_{(k)}$, and $V_{(k,l)}$ the covariance between $H_{(k)}$ and $H_{(l)}$. Applying the multivariate delta method, the variance–covariance matrix of $g(\mathbf{H})$, $\mathbf{V}_{g(\mathbf{H})}$, is approximated by

$$\mathbf{V}_{g(\mathbf{H})} \approx \mathbf{G}\mathbf{V}_\mathbf{H}\mathbf{G} \tag{5.14}$$

$\mathbf{V}_{g(\mathbf{H})}$ is a diagonal matrix for which the kth diagonal element equals $V_k/(1 - H_k)^2$ and the off-diagonal element (k, l) equals $V(k, l)/[(1 - H_{(k)})(1 - H_{(l)})]$. In data samples, $\mathbf{H}$ and $\mathbf{V_H}$ in Eqs. 5.12 to 5.14 are replaced by their estimates $\widehat{\mathbf{H}}$ and $\mathbf{V}_{\widehat{\mathbf{H}}}$, respectively, to get estimates $g(\widehat{\mathbf{H}})$ and $\mathbf{V}_{g(\widehat{\mathbf{H}})}$.

5.2.3 Approximating the Sampling Distribution

In Figure 5.1, the Wald-based and range-preserving approximations of the distribution are plotted over the distributions. This visualization shows that for H not close to the boundary of 1, the approximated distributions are similar (upper panels), but close to the boundary the range-preserving approximation (solid line) approaches the distribution more accurately than the Wald-based approximation (dashed line), especially in the left tail (lower panels). Note that the left tail is of interest because the one-sided significant tests evaluate whether H_{ij} or H_i is significantly larger than some hypothesized value. Hence, the left tail is compared to the hypothesized value.

5.3 Simulation Study

We performed a small-scale simulation study to investigate the coverage of the two-sided confidence interval and Type I error rate of the one-sided significance test for the Wald-based and range-preserving methods.

5.3.1 Method

We simulated data for 10 dichotomous items using a two-parameter logistic model (Birnbaum, 1968, p. 458). The difficulty parameter was fixed to equidistant values between -1 and 1 across the items.

Independent Variables

Item discrimination: The magnitude of the total-scale scalability coefficient was manipulated by increasing a_i in the two-parameter logistic model. The higher the discrimination, the better the item can distinguish between respondents, which results in higher scalability of the item, and thus the total scale. There were six levels in which a_i varied across items at equidistant values between 0 and 1: between 0.8 and 2, between 1 and 4, between 2 and 8, between 10 and 25, or between 25 and 75, resulting in $H = .07$ (unscalable), .33 (weak), .52 (strong), .74 (very strong), .94 (extremely strong), and .98 (near unity), respectively.

Sample size: The sample size N was 100, 500, or 1,000. Although 100 respondents is not considered sufficient for a Mokken scale analysis (Straat et al., 2014), the difference between the methods is expected to be more distinct.

Method: The Wald-based and range-preserving methods were used to compute the dependent variables.

Dependent Variables

We evaluated the following dependent variables, for which the population value H was determined by using the mean of $\widehat{H}$ across all replications within a condition, assuming

it was unbiased.

Coverage: The coverage of the two-sided 95% confidence interval was determined to be the proportion of times H was included in the 95% CI or CI*. Its value should be close to .95.

Type I error rate: The Type I error rate of the one-sided significance test was determined as the proportion of times the p value of z or z^* was below significance level .05. Its value should be close to the significance level. Statistics z and z^* were computed by replacing c by H in Eqs. 5.4 and 5.11.

Analysis

Method was a within-subject variable, and all other independent variables were between-subject variables. There were $4 \times 3 = 12$ conditions and for each condition 10,000 datasets were simulated. Data were simulated in software package R (R Core Team, 2020) using the function `simdata()` from package `mirt` (Chalmers, 2012)[1].

5.3.2 Results

Figure 5.2 shows the coverage rates of CI and CI* and the Type I error rates of z and z^* for all conditions. CI* outperformed CI in all conditions—more substantially for conditions where H approached its upper boundary of 1. Overall, the coverage rates were poorer in the conditions with 100 respondents compared to the conditions with more respondents, especially for the highest two H conditions. In general, for the two highest H conditions, the average undercoverage of CI was divided in 9.2% on the left side and 1.2% on the right side, indicating that the CI had mainly undercoverage in the left tail of the distribution, whereas the right tail was overcovered. The undercoverage of CI* was divided more symmetrically, with 5.3% on the left tail and 3.6% on the right tail. When looking only at the 500 and 1,000 respondents conditions, the undercoverage of CI was 5.5% in the left tail and 1.2% in the right tail, compared to 2.8% in both tails for CI*. This indicates that the tails of the sampling distribution are better approximated using the range-preserving method compared to the Wald-based method.

The Type I error rate of z^* was close to that of z in most conditions, but for conditions in which $H \geq .74$, z^* outperformed z. For 100 respondents, the Type I error rate of both z and z^* was below the nominal value for the lowest H condition, but improved with increased sample size.

Note that for the condition with 100 respondents and $H = .98$, in approximately 10% of the replications $\widehat{H}$ was (very close to) 1 (i.e., $\widehat{H} > .999$) and $SE_{\widehat{H}}$ was estimated (very close to) 0 (i.e., the mean of $SE_{\widehat{H}} = .0002$, compared to a mean of .0086 for $\widehat{H} < .999$). Regardless of the method, this estimation issue made it problematic to construct accurate intervals or to perform accurate tests, resulting in deteriorated coverage and Type I error rates for both methods.

[1]Syntax files are available to download from the Open Science Framework: https://osf.io/5m827/

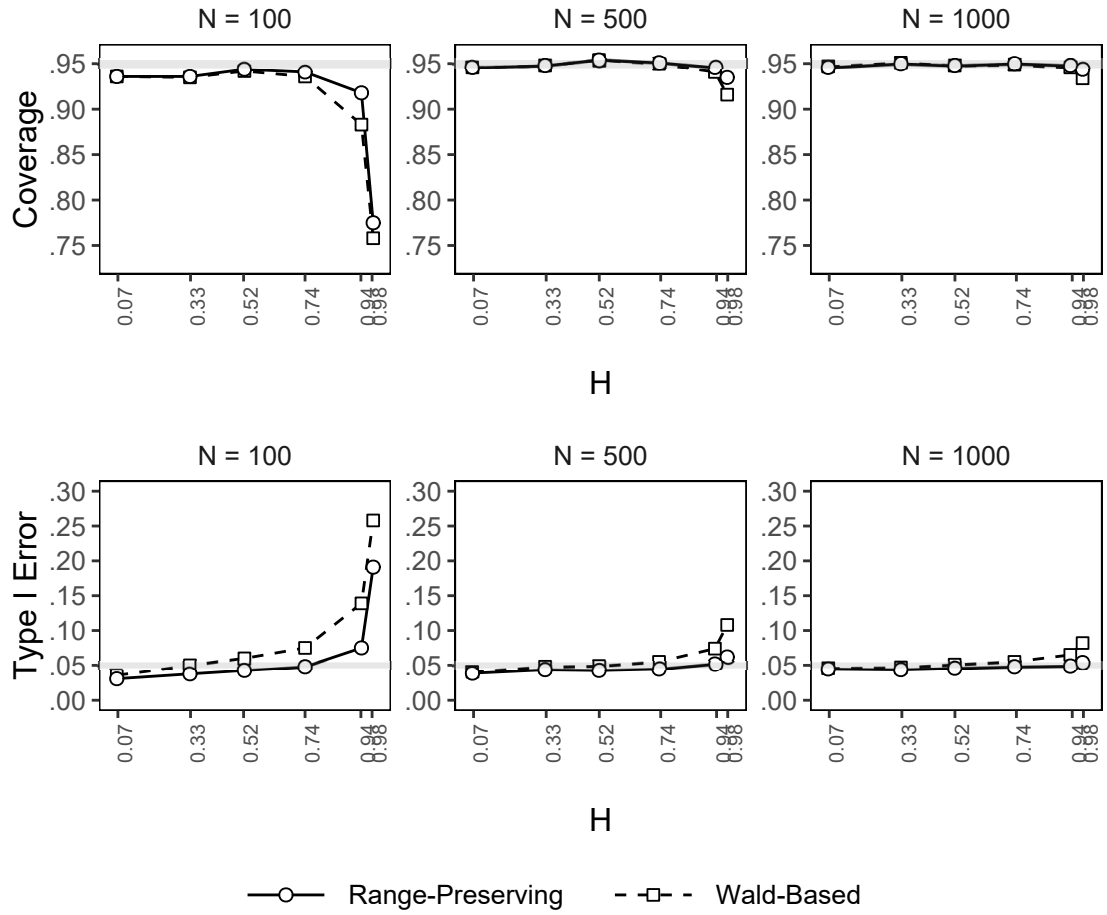

Figure 5.2: Coverage rates of the two-sided confidence interval (top panels) and Type I error rates of the one-sided significance test (bottom panels) for the Wald-based (dashed line) and range-preserving (solid line) method. The row panels represent the sample size N. Each panel displays the population values H on the horizontal axis.

5.4 Implementation in Software

The range-preserving methods are implemented in R (R Core Team, 2020) in the package `mokken` (Van der Ark, 2007, 2012). Here we give an overview of how to get CI^* and z^* for Mokken's scalability coefficients in nonclustered and clustered data and for Snijders' two-level scalability coefficients in multi-rater data, using scores of 639 students nested in 30 schools on 7 items measuring their well-being with teachers. The first column in the dataset contains a grouping variable, which we will ignore for nonclustered computations but which we use for clustered data. Throughout we will use the significance level $\alpha = .05$ and a null hypothesis for coefficients $c = .3$. Wald-based results can be obtained by replacing `"RP"` by `"WB"` in the R code. Let `R>` denote the R prompt. The R script and output are available to download from the Open Science Framework: https://osf.io/5m827/.

```
R> # Preliminary code:
R> # Load package, get data
R> # Set significance level and value c
R> library(mokken)
R> data(SWMD)
R> X <- SWMD[, -1] # item scores SWMD
R> groups <- SWMD[, 1] # grouping variable
R> alpha <- .05 # Significance level
R> c <- .3 # Null hypothesis value
R> ## Mokken's scalability coefficients in nonclustered data:
R> # Point estimates, standard errors,
R> # and two-sided range-preserving confidence intervals
R> coefH(X, ci = 1 - alpha, type.ci = "RP")
R> # Range-preserving z-scores using null hypothesis c
R> coefZ(X, lowerbound = c, type.z = "RP")
R> ## Mokken's scalability coefficients in clustered data:
R> # Point estimates, standard errors,
R> # and two-sided range-preserving confidence intervals
R> coefH(X, ci = 1 - alpha, type.ci = "RP", level.two.var = groups)
R> # Range-preserving z-scores using null hypothesis c
R> coefZ(X, lowerbound = c, type.z = "RP", level.two.var = groups)
R> ## Snijders' two-level scalability coefficients:
R> # Point estimates, standard errors,
R> # and two-sided range-preserving confidence intervals
R> MLcoefH(SWMD, ci = 1 - alpha, type.ci = "RP")
R> # Range-preserving z-scores using null hypothesis c
R> MLcoefZ(SWMD, lowerbound = c, type.z = "RP")
```

5.5 Discussion

We proposed a method to compute range-preserving confidence intervals and significance tests, which we implemented in the R package `mokken`. Simulation results showed that for H not close to 1, Wald-based and range-preserving methods are very similar and both are useful. However, for very strong scales ($H > .7$), the range-preserving methods return more accurate results and are preferred over the Wald-based method, especially for the left tail of the sampling distribution (which is used in the one-sided significance tests). The results were poorer for only 100 respondents, confirming that larger samples are desirable (Straat et al., 2014). Note that we only investigated range-preserving methods for scalability coefficients in nonclustered data. Whether the results are similar in clustered data and for two-level scalability coefficients is a topic for further research.

In our method, we used $(-\infty, 1]$ as the range for H. However, the actual minimum of

scalability coefficients depends on the marginal frequencies (Sijtsma & Molenaar, 2002, p. 59). This minimum has the undesirable property that it must be estimated and thus varies across finite samples. We explored an alternative and more complex logistic transformation that takes the estimated minimum into account. The results were very similar to the results obtained using the logarithmic transformation presented in this chapter, so we did not investigate this method any further.

A limitation of the logarithmic transformation is that the value 1 can not be included in the interval (although values very close to 1 can), as this value corresponds to ∞ on the transformed scale. However, 1 is a possible value for scalability coefficients, both in the population and in data samples. Alternative transformations that can include 1 may approximate the sampling distribution more closely. However, this will not solve the deteriorated coverage and Type I error rates for very high H entirely because there remain samples where the SE can not be estimated because $\widehat{H} = 1$.