Nonparametric item response theory for multilevel test data

Koopman, L.

Publication date
2022

Link to publication

Citation for published version (APA):
Koopman, L. (2022). *Nonparametric item response theory for multilevel test data*. [Thesis, fully internal, Universiteit van Amsterdam].

Chapter 7

A Two-Step, Test-Guided Mokken Scale Analysis, for Nonclustered and Clustered data

Abstract

Mokken scale analysis (MSA) is an attractive scaling procedure for ordinal data. Two of MSA's prime features are the scalability coefficients and the automated item selection procedure (AISP). The AISP partitions a (large) set of items into scales based on the observed item scores; the resulting scales can be used as measurement instruments. There exist two issues in MSA: First, point estimates, standard errors, and test statistics for scalability coefficients are inappropriate for clustered item scores, which are omnipresent in quality-of-life research data. Second, the AISP insufficiently takes sampling fluctuation of Mokken's scalability coefficients into account. We solved both issues by providing point estimates and standard errors for the scalability coefficients for clustered data and by implementing a Wald-based significance test in the AISP algorithm, resulting in a test-guided AISP (T-AISP), that is available for both nonclustered and clustered test scores. We integrated the T-AISP into a two-step, test-guided MSA for scale construction. The procedure was demonstrated on clustered item scores obtained from administering a questionnaire on quality of life in schools.

Chapter 7 is published as: Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2022). A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. *Quality of Life Research*, 31(1), 25–36. Springer. doi: 10.1007/s11136-021-02840-2

7.1 Introduction

Nonparametric item response theory (NIRT) models (Mokken, 1971; Sijtsma & Molenaar, 2002) are flexible measurement models that put relatively few restrictions on the data compared to parametric item response theory (PIRT) models, such as the Rasch model (Rasch, 1960), the graded response model (Samejima, 1969) and the partial credit model (Masters, 1982). Therefore, NIRT models will fit the data relatively well. NIRT models have the attractive property that if the model fits the data, the sum score can be used to order respondents on the latent trait measured by the test (see Van der Ark & Bergsma, 2010, for details). Using the sum score for measurement is very common in quality of life research, but providing justification for using the sum score for measurement, for example by showing that a NIRT model fits the data, is less common.

NIRT models have two major applications. First, NIRT models can be used as standalone measurement models. As NIRT models have relatively good fit compared to PIRT models, NIRT models are preferred for constructing tests and questionnaires that require ordering respondents, such as using the test scores to order respondents from high to low on the ability to cope independently, to select the 30% most capable respondents for a special training program, or to construct ordinal test norms such as percentile scores. Second, NIRT models can be used preliminary to PIRT models. Below, we discuss methods to identify items that do not fit the NIRT model, and that should be removed from the test or questionnaire. As all popular PIRT models are special cases of NIRT models (Van der Ark, 2001), the items selected for removal under the NIRT model will not fit under the PIRT model either. Removing these badly fitting items prior to PIRT analysis will simplify the PIRT analysis.

Mokken scale analysis (MSA; Mokken, 1971; see also, e.g., Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017; Van Schuur, 2003; and Wind, 2017 for elaborate introductions to MSA and its methods) is a scaling method that consists of various tools to investigate the fit of NIRT models. The most popular tools of MSA are the *scalability coefficients* (or H-coefficients; Mokken, 1971, pp. 152-185), a diagnostic tool to evaluate whether the items form a scale, and the *automated item selection procedure* (AISP; Mokken, 1971, pp. 190-194; Van Abswoude, van der Ark, & Sijtsma, 2004, see also). The AISP selects items from a larger set of items into scales. Items that are not selected into a scale either violate the assumptions of the NIRT model, or have poor discriminatory power. In addition to the scalability coefficients and the AISP, MSA contains a set of methods to investigate whether the specific assumptions of the NIRT model holds. With regard to data collection, MSA assumes that the item scores are obtained using simple random sampling; that is, the respondents in the sample must not be clustered in groups, such as classes, hospitals, or geographical regions.

MSA is especially suitable for constructing quality of life (related) measures, as the NIRT properties are often sufficient for the intended use of the scales (Sijtsma et al., 2008). Recent examples of quality of life questionnaires that have been analyzed using

MSA include the Heart disease-specific health-related Quality of Life (HeartQoL; Huber et al., 2020), the Participation and Activity Inventory for Children and Youth (PAI-CY; Elsman et al., 2020), and the Rotterdam Diabetic Foot Study Test Battery (RDF Rinkel et al., 2019). In these studies, the data were collected using a simple random sample, yielding *nonclustered* test scores. A simple random sample is not always preferable in quality of life research, due to constraints in funding, time, or the sampling frame or due to a substantial preference for including multiple levels. In such cases, quality of life questionnaires are administered to respondents who are nested in groups, yielding *clustered* test scores (i.e., obtained by a cluster or multi-stage sampling design). Some authors who used such a sampling design include Elley et al. (2005), who investigated health-related quality of life related variables in 233 older patients from 42 general practitioners; S.-K. Chen et al. (2013), who investigated the quality of life of 1392 high-school and middle-school students nested in school classes; and Fisher & Li (2004), who investigated the effects of a neighborhood walking program on quality of life among 182 older adults from 56 different neighborhoods in Portland, Oregon. In these examples, the interest is in measuring the trait of the patients, students, and older adults (level 1), whereas the grouping variables at level 2 (general practitioners, school classes, and neighborhoods) were considered to be a nuisance.

In this chapter, we discuss two issues in MSA: First, point estimates, standard errors, and test statistics for scalability coefficients are unavailable for clustered data. Hence, currently scalability coefficients should not be used to evaluate clustered test scores. As the AISP also uses the point estimates and test statistics for scalability coefficients, the AISP should not be applied to clustered test scores either. Second, the AISP insufficiently takes into account sampling fluctuations, explained in detail hereunder. As a result, the AISP may be too liberal. We solved both issues by providing point estimates and standard errors for the scalability coefficients for clustered data and by incorporating a z-tests in a test-guided AISP (T-AISP) that tests all relevant hypotheses, and that is available for both nonclustered and clustered test scores. We integrated the T-AISP into a comprehensive two-step, test-guided MSA, to guide the analysis for nonclustered and clustered data.

The remainder of this chapter is organized as follows: First, we discuss the scalability coefficients and the AISP, and elaborate on issues when using the scalability coefficients for clustered data and issues when using the AISP for both nonclustered and clustered data. Second, we propose solutions and introduce a T-AISP that tackles the issue pertaining to significance testing and the estimation methods for clustered data. Third, we incorporate the proposed methods in a two-step, test-guided MSA that can be applied to nonclustered and clustered data. Finally, using the two-step procedure, we analyzed data from the two-dimensional Dutch quality of life at school questionnaire Schaal Welbevinden met Docenten en Klasgenoten [Scale Well-Being with Teachers and Classmates] (SWMDK; Peetsma et al., 2001).

7.2 Scalability Coefficients

There are three types of scalability coefficients that can be used as a diagnostic tool for NIRT model fit, and discriminatory power. Item-pair scalability coefficient H_{ij} (for the pair of items i and j) is a normed correlation between the two items (e.g., Van Abswoude, van der Ark, & Sijtsma, 2004). Item-scalability coefficient H_i is a normed item-rest correlation (i.e., the correlation between item i and the total score on the remaining items) and can be regarded as a discrimination index (e.g., Van Abswoude, van der Ark, & Sijtsma, 2004; Zijlmans et al., 2018). 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 (Mokken, 1971, p. 152; Sijtsma & Molenaar, 2002, pp. 57-68). For a set of items $\min(H_{ij}) \leq \min(H_i) \leq H \leq \max(H_i) \leq \max(H_{ij})$ (Sijtsma & Molenaar, 2002, p. 58).

Mokken (1971, pp. 184-185) called a set of items a scale (further referred to as a Mokken scale) if two criteria are met:

$$H_{ij} > 0 \text{ for all item pairs;} \tag{7.1}$$

$$H_i \geq c \text{ for all items,} \tag{7.2}$$

with c being some positive lowerbound, for which $c = .3$ is often used. The first criterion follows from the assumptions of Mokken's NIRT models, which result in positive inter-item correlations. The second criterion is a practical requirement that ensures only sufficiently discriminating items are selected into the scale (see also Sijtsma & Molenaar, 2002, p. 68). Mokken (1971, p. 185) provided benchmarks to determine the strength of a scale: $H \geq 0.5$ reflects a strong scale, $H \geq 0.4$ a medium scale, and $H \geq 0.3$ a weak scale. The stronger the scale, the more accurately persons can be ordered on the latent trait by means of their total score (Sijtsma & Molenaar, 2002, p. 68).

7.2.1 Estimating Scalability Coefficients

So far, properties of scalability coefficients in the population were discussed. In research, we have finite samples, and scalability coefficients must be estimated from the sample data, where $\widehat{H}_{ij}$, $\widehat{H}_i$, and $\widehat{H}$ denote the estimate of population value H_{ij}, H_i, and H, respectively, and $SE_{\widehat{H}_{ij}}$, $SE_{\widehat{H}_i}$, and $SE_{\widehat{H}}$ denote the standard error of the estimate. For nonclustered data, Mokken (1971 p. 166; Sijtsma & Molenaar, 2002, p. 49) derived estimates of scalability coefficients, and Kuijpers et al. (2013) derived estimates for the standard errors. These estimation methods are referred to as *one-level methods*. For computational details, see Chapter 6, Section 1 and 2. One-level methods assume the data are obtained from a simple random sample.

In clustered data, respondents are nested within groups, violating the assumption of an independent simple random sample that underlies the one-level methods. A typical aspect of clustered data is positive within-group dependency, in which test scores of respondents within the same group (or cluster) are more similar than test scores of respondents in different groups. A commonly used statistic to express within-group dependency is the intraclass correlation (ICC), which is the expected correlation between two test scores in the same group (Stapleton et al., 2016). ICCs between 0 to 0.5 are common for measures of quality of life and related concepts (e.g., Atkinson et al., 1997; Elley et al., 2005; Fan et al., 2011; Gulliford et al., 1999; Mok, 2002). In general, accounting for dependency in the data is advised if the ICC > 0, as it can severely affect outcomes of statistical analyses (Geldhof et al., 2014; Maas & Hox, 2005; Stapleton et al., 2016). For example, ignoring positive within-group dependency is a well-known cause of underestimated standard errors (e.g., Maas & Hox, 2005; Hox, 2010, p. 294; although in rare cases overestimation may also happen, e.g., Moerbeek et al., 2003). As a result, confidence intervals will be too narrow, making the estimates appear more precise than they actually are, and the type I error rates of significance tests will be inflated, and should not be used.

For clustered data, estimates of scalability coefficients and standard errors are not yet available. However, Snijders (2001a) derived estimates for scalability coefficients for so-called *multi-rater data*, and Koopman, Zijlstra, & Van der Ark (2020) derived estimates for their standard errors. Multi-rater data are also multilevel data, where level 2 (the group level) is of primary interest, and level 1 (the respondent level) is a nuisance. Snijders proposed scalability coefficients for both levels: Within-rater scalability coefficients H^W_{ij}, H^W_i, and H^W for level 1, and between-rater scalability coefficients H^B_{ij}, H^B_i, and H^B for level 2. The estimates of the within-rater scalability coefficients and their standard errors may provide a viable alternative for Mokken's scalability coefficients in clustered data.

7.2.2 Hypothesis Tests and Confidence Intervals for Scalability Coefficients

Mokken (1971, pp. 160-162) proposed a test of marginal independence to evaluate null hypotheses $H_{ij} = 0$, $H_i = 0$ or $H = 0$. Let S_{ij} denote the estimated covariance of item pair (i, j) and S_i the estimated standard deviation of item i, in a sample of size N. Then, Mokken's test statistic for item-pair coefficient H_{ij} is defined as

$$\Delta_{ij} = \frac{S_{ij}}{S_i S_j}\sqrt{N-1}. \tag{7.3}$$

Mokken's test statistic is also available for item coefficients,

$$\Delta_i = \frac{\sum_{(j \neq i)} S_{ij}}{S_i \sum_{(j \neq i)} S_j}\sqrt{N-1}, \tag{7.4}$$

and the total scale coefficient,

$$\Delta = \frac{\sum_i \sum_{(j>i)} S_{ij}}{\sum_i \sum_{(j>i)} S_i S_j}\sqrt{N-1} \tag{7.5}$$

(see also Van der Ark et al., 2008, Section 3.5). Asymptotically, the test statistics follow a standard normal distribution Mokken (1971, pp. 162). A one-sided significance test can evaluate the null hypothesis $H_{ij} \leq 0$ with alternative hypothesis $H_{ij} > 0$. Let z_{crit} denote the critical value (i.e., the point on the normal distribution for a given significance level α; e.g. $z_{\text{crit}} \approx 1.645$ for $\alpha = .05$), that is compared to the test statistic to determine whether to reject the null hypothesis. The null hypothesis is rejected if the test statistic exceeds z_{crit}; for example, if $\Delta_{ij} > z_{\text{crit}}$. This test assumes a simple random sample, and is therefore only suited for nonclustered data.

Recently, Koopman et al. (2021) defined a Wald-based significance test and a range-preserving significance test that both use the point estimates and standard errors of scalability coefficients to evaluate null hypotheses $H_{ij} = c$, $H_i = c$, or $H = c$, with c being some constant. The Wald-based test statistic for H_{ij} is defined as

$$z_{ij} = \frac{\widehat{H}_{ij} - c}{SE_{\widehat{H}_{ij}}}. \tag{7.6}$$

Let $g(\widehat{H})$ denote a (logarithmic) transformation of $\widehat{H}$, with standard error $SE_{g(\widehat{H})}$, and $g(c)$ the transformation of hypothesized value c (for computational details see Koopman et al., 2021). The range-preserving test statistic for H_{ij} is defined as

$$z^*_{ij} = \frac{g(\widehat{H}_{ij}) - g(c)}{SE_{g(\widehat{H}_{ij})}}. \tag{7.7}$$

The Wald-based and range-preserving test statistics are also available for item coefficients (denoted z_i and z^*_i, by replacing $\widehat{H}_{ij}$ with $\widehat{H}_i$ in Equations 7.6 and 7.7, respectively) and the total scale coefficient (denoted z and z^*, by replacing $\widehat{H}_{ij}$ with $\widehat{H}$ in Equations 7.6 and 7.7, respectively; see also Koopman et al., 2021). For nonclustered data, the range-preserving test has better type I error rates compared to the Wald-based test for very strong scales (e.g., $H > .7$; Koopman et al., 2021).

As a result of the availability of both Wald-based and range-preserving test statistics, confidence intervals around the scalability coefficients can also be either Wald-based or range-preserving. Wald-based confidence intervals have the form

$$\text{CI} = \widehat{H} \pm 1.96 \times SE_{\widehat{H}} \tag{7.8}$$

for a two-sided 95% confidence interval for the total-scale coefficient (Mokken, 1971, p. 168; Koopman, Zijlstra, & Van der Ark, 2020). However, the maximum value of scalability coefficients is 1, and if H is close to 1 or its standard error is large, Wald-based confidence intervals can include values larger than 1 and are biased due to a skewed sampling distribution. Hence, Koopman et al. (2021) proposed range-preserving confidence interval.

Let $g^{-1}\big[g(\widehat{H})\big] = \widehat{H}$ denote the inverse of $g(\widehat{H})$, then

$$\text{CI}^* = g^{-1}\big[g(\widehat{H}) \pm 1.96 \times SE_{g(\widehat{H})}\big] \tag{7.9}$$

is the two-sided 95% confidence interval of the total-scale coefficient. This interval ensures all values to be within the possible range of the coefficient, and has better coverage rates than the Wald-based interval in nonclustered data for high values of H.

7.3 Automated Item Selection Procedure

The objective of the AISP is to select as many items as possible into a scale, as long as these items meet the Mokken-scale criteria (Equations 7.1 and 7.2). Table 7.1, upper panel, provides an overview of how the Mokken scale criteria currently are evaluated in the AISP. Criterion 1 is accepted if $\Delta_{ij} > z_{\text{crit}}$ (Equation 7.3), using null hypothesis $H_{ij} \leq 0$ and alternative hypothesis $H_{ij} > 0$. Criterion 2 is accepted if $\Delta_i > z_{\text{crit}}$ (Equation 7.4), using null hypothesis $H_i \leq 0$ and alternative hypothesis $H_i > 0$, and $\widehat{H}_i \geq c$. Hence, for Criterion 2 the hypothesis $H_i \geq c$ is not tested, but evaluated on the point estimate, which may render the procedure too liberal.

Table 7.1:

Mokken Scale Criteria Evaluation by the AISP (upper panel) and the T-AISP (lower panel).

Criterion		AISP	
	Null hypothesis	Hypothesis matches criterion	Accepts criterion if
1: $H_{ij} > 0$	$H_{ij} \leq 0$	✓	$\Delta_{ij} > z_{\text{crit}}$
2: $H_i \geq c$	$H_i \leq 0$	–	$\Delta_i > z_{\text{crit}}$ and $\widehat{H}_i \geq c$

Criterion		T-AISP	
	Null hypothesis	Hypothesis matches criterion	Accepts criterion if
1: $H_{ij} > 0$	$H_{ij} \leq 0$	✓	$z_{ij} > z_{\text{crit}}$
2: $H_i > c$	$H_i \leq c$	✓	$z_i > z_{\text{crit}}$

The AISP starts with a (typically large) set of items that have been administered to a sample of respondents, so for each item in the set, item scores are available. The AISP uses the following algorithm.

1. *Select the first two items in the scale.* These two items have the highest value of $\widehat{H}_{ij}$ and both Mokken-scale criteria must have been accepted. If for no item-pair both Mokken-scale criteria are accepted, no items are selected and the AISP stops.

2. *Select the next items into the scale.* The next item selected in the the scale is the item for which both Mokken-scale criteria are accepted and that produces the highest $\widehat{H}$-value when computed on all selected items. Step 2 is repeated until there are either no more items left, or until no more items can be added for which both Mokken-scale criteria are accepted.
3. *Start the next scale.* The AISP returns to Step 1 to form a next scale using only the unselected items. If there are no more items left or if there are no more pairs of items for which the Mokken-scale criteria are accepted, the AISP stops.

Note that the value z_{crit} is adjusted in each subsequent step using a Bonferroni correction for the number of tests performed in the previous steps and the current step of the algorithm (Sijtsma & Molenaar, 2002, p. 72). Alternative algorithms for automated item selection in MSA have been suggested (e.g., Van Abswoude, Vermunt, et al., 2004; Brusco et al., 2015; Straat et al., 2013), but are not discussed.

The AISP can be applied using different values of lowerbound c. By increasing c the criterion for item scalability becomes more stringent. Stringent criteria lead to shorter scales with higher discrimination power in the sample under investigation. When applying the AISP to investigate whether a set of items form one or more Mokken scales, Hemker et al. (1995; see also Sijtsma & Van der Ark, 2017) advised to use increasing values for c (e.g., $c = 0, .05, .10, \ldots, .55$). Typically, for a unidimensional scale satisfying the NIRT model, for small values of c, all or most items are in a single large scale, as c increases most items are in a single smaller scale and the remaining items are unscalable, and as c increases further, there are only one or a few small scales and several unscalable items remain. For a multidimensional scale satisfying the NIRT model, typically for small c all or most items are in one large scale, as c increases most or all items are divided over two or more scales, and as c increases further there are two or more smaller scales and several unscalable items. Note that the number of Mokken scales does not necessarily reflect the dimensionality of the item set, this depends on the correlation between the dimensions and the level of item discrimination (I. A. M. Smits et al., 2012).

Because $H_i > c$ is not tested and only point estimates are used for the evaluation of Criterion 2, the sampling fluctuation is not taken into account and the item selection is too liberal. Ignoring sample fluctuation can result in the inclusion of items that do not contribute to (or possibly negatively affect) accurate measurement of the scale in the population. Hence, we require an alternative significance test in the AISP that meets the following requirements: It can test null hypotheses for values of c other than zero and is available for both nonclustered and clustered data.

7.4 Solving the two MSA Issues

7.4.1 Point Estimates, Standard Errors, and Statistical Tests of Scalability Coefficients for Clustered Data.

As noted earlier, no point estimates or standard errors are available for scalability coefficients for clustered data, but for multi-rater data, point estimates (Snijders, 2001a) and standard errors (Koopman, Zijlstra, & Van der Ark, 2020) have been derived. For clustered data, we derived point estimates and standard errors in Chapter 6 by slightly modifying the point estimates and standard errors for multi-rater data, which we coin the *two-level methods*. It turned out that the point estimates of the two-level method are equivalent to the point estimates of the one-level method, whereas the standard errors account for the additional variation that is typical for clustered data. For computational details and explanation of the modification, we refer to Chapter 6, Sections 1 and 2. The estimates based on the two-level method can be plugged into Equations 7.6 to 7.9 to get the test statistics and confidence intervals of scalability coefficients for clustered data.

In a small-scale simulation study (Chapter 6, Section 3) we compared one-level methods to two-level methods and Wald-based methods to range-preserving methods for Mokken's scalability coefficients in clustered data. Point estimates of the scalability coefficients were accurately estimated in all conditions. In general, estimating standard errors and confidence intervals using the two-level method produced less bias and better coverage rates than using the one-level method. This was especially true for larger ICC levels and for larger groups. For small ICC values (i.e., ICC $\leq$ .12), especially for small group sizes (i.e., below 10), the standard errors estimated using the two-level method were slightly conservative, but the absolute difference was small compared to the one-level method. Hence, we recommend to use the two-level method for estimating standard errors in clustered data. The coverages of the Wald-based and range-preserving confidence intervals were approximately similar, but the Wald-based method was slightly more symmetric. As in practice, scalability coefficients are seldom very close to 1, and all relevant hypotheses regarding scalability coefficients are in the range 0 to .55, we believe the additional value of range preserving confidence intervals in practice is limited. Hence, we recommend Wald-based confidence intervals and significance tests.

7.4.2 Test-Guided Automated Item Selection Procedure

As noted earlier, both Criterion 1 and Criterion 2 (Equations 7.1 and 7.2) should be tested in the AISP, to gain confidence that the formed scales satisfy both Mokken-scale criteria. In addition, the testing procedure should be available for both nonclustered and clustered data. We propose a test-guided AISP (T-AISP) that uses the same algorithm of the AISP, but that evaluates the criteria of a Mokken scale using different significance tests. Specifically, in the T-AISP, Mokken's Δ_{ij} and Δ_i are replaced by the Wald-based test statistics z_{ij} (Equation 7.6), with null hypothesis $H_{ij} \leq 0$ and alternative hypothesis $H_{ij} > 0$, and z_i (cf. Equation 7.6), with null hypothesis $H_i \leq c$ and alternative hypothesis $H_i > c$, respectively. Table 7.1, lower panel, provides an overview of how the Mokken scale criteria are evaluated in the T-AISP. Criterion 1 is accepted if $z_{ij} > z_{\text{crit}}$. Criterion 2 is accepted if $z_i > z_{\text{crit}}$. Consequently, the evaluation of Criterion 2 in the AISP that uses the point estimate, $\widehat{H}_i > c$ (Table 7.1, third column in upper panel), is redundant and can be removed from the algorithm. Although the original AISP is also partially test-guided (Criterion 1 is tested), we refer to the adjusted evaluation as the test-guided procedure, because only in the T-AISP algorithm both criteria of a Mokken scale are tested. We distinguish between a T-AISP using one-level methods and a T-AISP using two-level methods. The T-AISP using one-level methods computes the standard errors in z_{ij} and z_i using the one-level methods, which gives an appropriate test for nonclustered data. The T-AISP using two-level methods computes these standard errors using the two-level methods, which gives an appropriate test for clustered data. Using the Wald-based significance tests in the T-AISP results in a slightly different second Mokken-scale criterion compared to the original definition: A scale for which $H_i > c$ for all items, rather than $H_i \geq c$. Note that replacing Δ_{ij} and Δ_i in Table 7.1, upper panel, by z_{ij} and z_i, respectively, while retaining the same hypotheses, makes the AISP available for clustered data when two-level methods are used, but retains the issue that Criterion 2 is not tested.

The major difference of the T-AISP compared to the AISP is that Criterion 2 of a Mokken scale (Equation 7.2) is statistically tested, rather than evaluated using a point estimate. As a result, T-AISP is more conservative compared to the AISP, because uncertainty of $\widehat{H}_i$ is taken into account. As the point estimates become more accurate (e.g., for larger samples), the uncertainty becomes smaller, and the formed scales by the T-AISP approaches those of the AISP more closely. If there is a substantial amount uncertainty, the T-AISP will generally result in more and smaller scales and is likely to show more unscalable items compared to the AISP for the same lowerbound c. Note that the resulting scale patterns in the T-AISP can be different from patterns that emerge by using more stringent criteria for lowerbound c in the AISP, because uncertainty can differ across items, samples, and sample sizes, whereas a given c is fixed regardless of the items and sample.

7.5 A Two-Step, Test-Guided MSA, for Nonclustered and Clustered Data

The two-step, test-guided MSA is a procedure to create scales from a set of items, evaluate the strength of these scales, and perform follow-up analyses such as fit diagnostics of the NIRT models and possibly PIRT models. Figure 7.1 shows a flow chart of the procedure, which is elaborated on below.

The procedure commences with determining whether the data are clustered, based

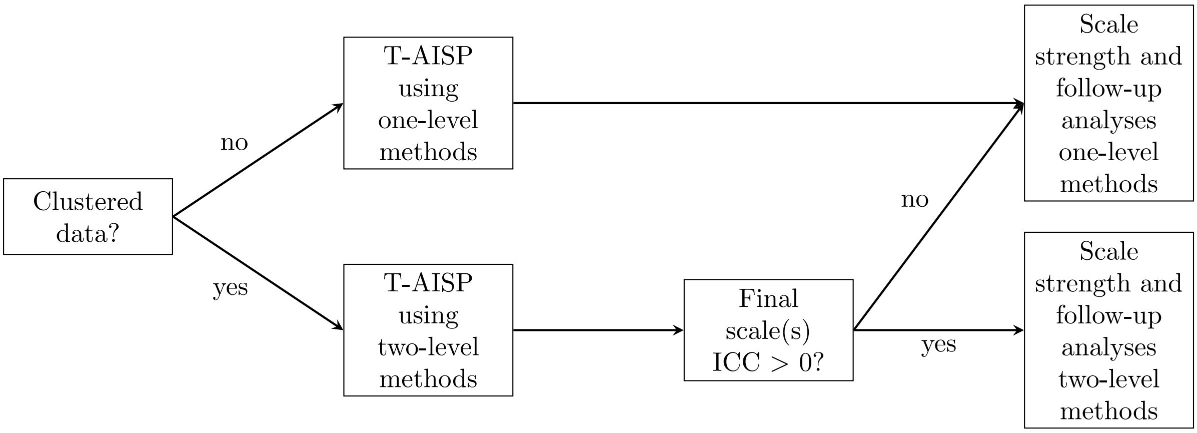
Clustered data?
no
yes
T-AISP using one-level methods
T-AISP using two-level methods
Final scale(s) ICC > 0?
no
yes
Scale strength and follow-up analyses one-level methods
Scale strength and follow-up analyses two-level methods

Figure 7.1: Flow chart of the two-step test-guided procedure for scale construction.

on the sampling design. Hence, if data were collected using a simple random sampling design, the data are nonclustered, whereas if a (two-stage) cluster sampling design was used, the data are clustered. For clustered data, we discourage investigating the ICC of a test score at this stage. The rationale is that test scores in this stage are based on a set of items for which the quality is unsure, as the goal of MSA is forming good quality scales, and poor items should not affect decisions based on the ICC. For example, unscalable items may mask within-group dependency that is possibly present when using the final scale. If one was to perform the T-AISP using one-level methods, the one-level standard errors of large-ICC items are likely to be substantially too small, and there is an increased risk of incorrectly admitting these items to the final scale.

In Step 1 the T-AISP is applied to form one or more Mokken scales. The significance tests in the algorithm are performed using one-level methods for nonclustered data and two-level methods for clustered data. The T-AISP is performed using increasing lowerbounds from 0 to .55 in steps of .05 to investigate how scales are formed and how stable they are, although the number and range of lowerbounds may be adjusted based on practical or theoretical considerations. Final scales are selected based on stability, discrimination, and possibly theoretical considerations.

For clustered data, the within-group dependency of the formed scales is evaluated. The within-group dependency of each scale is investigated by estimating the ICC of the total score per scale and performing an F-test to test the null hypothesis that the ICC is zero (for computational details, see Snijders & Bosker, 2012, pp. 19-23). This is the F-test also used in analysis of variance and assumes that the test scores within a group are normally distributed. If the F-test is not significant, it is plausible that the ICC is zero, and accounting for the nesting is not necessary in the subsequent statistical analyses. However, if the ICC is significantly larger than zero, the subsequent analyses should be estimated using multilevel data analysis methods.

Step 2 is determining the strength of the final scales and performing follow-up analyses, using one-level methods for nonclustered data and for clustered data without within-group dependency, and two-level methods for clustered data with significant within-group dependency. The strength of the scale is evaluated using a 95% Wald-based confidence interval around total scale coefficient estimate $\widehat{H}$ (Equation 7.8) to get plausible values of the population coefficient H. The fit of Mokken models is investigated using several available methods, such as conditional association and manifest monotonicity (e.g., Sijtsma & Van der Ark, 2017). Items that show (severe) misfit may be adjusted or removed (Crişan et al., 2020). Subsequently, reliability analysis may be performed or more strict measurement models (e.g., PIRT models) may be fitted.

7.6 Real-Data Example

The SWMDK is a two-dimensional scale designed to measure a student's perception of their well-being at school with teachers and classmates. The items are scored on a five-point Likert scale, ranging from 1 (*not true at all*) to 5 (*completely true*), and contain elements of relationships, interactions, and feelings towards the teachers and classmates. The SWMDK consists of seven items pertaining teachers (SWMD, items 1 to 7) and six items pertaining classmates (SWMK, items 8 to 13; Zijsling et al., 2017), although shorter versions have been used (e.g. Van der Veen & Peetsma, 2009; Thoonen et al., 2011). To our knowledge, no scalability analysis has been performed on this scale, and therefore it is unsure whether the items may be used in one scale, whether subscales should be formed, and which items should be included in the scale(s). For the SWMD, Cronbach's αs varied between .63 and .84 and for the SWMK between .68 and .83 (Thoonen et al., 2011; Van der Veen & Peetsma, 2009; Zijsling et al., 2017). Table 7.2 shows the 13 items of the SWMDK.

7.6.1 Method

The data were collected in 814 classes at 94 secondary schools in the Netherlands, as part of a large-scale cohort study COOL$^{5-18}$ (see Zijsling et al., 2017, for sampling and consent procedures). The data used in this analysis consisted of a subset of 30 classes (each from a different school) to better reflect everyday practice in quality of life research in which smaller samples are more common. The average class size was 21.30 (SD = 6.49), in a total sample of 639 students. Table 7.2 shows the means and standard deviations of the items and the total scale of the SWMDK.

The data were collected in classes by means of a cluster sampling design, resulting in clustered data. Hence, we performed the two-step, test-guided MSA instructions for clustered data to investigate the scalability of the SWMDK. We desired a stable set or sets of items that was sufficiently discriminative, preferring Mokken scales with $H_i \geq .3$ for all items. All analyses were conducted in R (R Core Team, 2020) using the R package mokken (Van der Ark, 2007, 2012), which also contains the data. The R syntax to obtain the results in this section are available to download from the Open Science Framework: http://osf.io/y7xud. The R syntax in Figure 7.2 is a shortened version.

Table 7.2:

Item Content, Mean, and Standard Deviation for Each Item and for the Total Scale of the SWMDK.

Item		*M*	*SD*
1	The teachers usually know how I feel	2.84	0.89
2	I can talk about problems with the teachers	3.18	0.92
3	If I feel unhappy, I can talk to the teachers about it	3.03	0.98
4	I feel at ease with the teachers	3.52	0.77
5	The teachers understand me	3.23	0.81
6	I have good contact with the teachers	3.34	0.83
7	I would prefer to have other teachers*	3.22	0.85
8	I have a lot of contact with my classmates	4.06	0.76
9	I would prefer to be in another class*	3.89	1.08
10	We have a nice class	3.89	0.96
11	I get along well with my classmates	4.01	0.73
12	I sometimes feel alone in the class*	4.12	0.92
13	I enjoy hanging out with my classmates	4.00	0.74
Total scale		3.57	0.53

Note. SWMDK = Schaal Welbevinden Met Docenten en Klasgenoten. The items were translated from Dutch. For the original items, see pp. 79–83 in Zijsling et al. (2017). M = mean, SD = standard deviation. Items 1 to 7 pertain teachers, items 8 to 13 pertain classmates.
* Reversely scored item that has been recoded.

7.6.2 Results

In Step 1 of the two-step, test-guided MSA, for lowerbounds $c = .00$ and .05, all items were included in one scale (Table 7.3, columns 1 and 2), implying that the first Mokken-scale criterion ($H_{ij} > 0$) was accepted. For $c = .15$ to .45, two scales were formed, one predominantly containing SWMD items and one predominantly containing SWMK items. Items 7 and 12 were not included for $c \geq .25$ and item 8 was not included for $c = .45$ (Table 7.3, columns 3 to 10). For $c \geq .50$, the items fell apart into three small subscales and several unscalable items (Table 7.3, columns 11 and 12). This reflects the typical pattern expected for a two-dimensional scale. The results supported the division into two separate scales: The first well-being with teachers and the second well-being with classmates. Item 7 and 12 were removed from the final scales as they contributed too little to accurate measurement of the final scales ($c < .25$), hence they did not meet the second Mokken-scale criterion (which we defined as $H_i > .3$ for all items). Apparently item 7 and 12 contain elements that either are reflective of another construct, or misunderstood by the students. We continued with the procedure using the six-item SWMD and the

```
R> ### Two-step, test-guided MSA
R> # Load mokken package
R> library(mokken)
R> # Obtain data: Column 1: Grouping variable. Remaining columns: Item scores.
R> data(SWMDK)
R> ### Step 1
R> # Set lowerbound at increasing values from 0 to .55 in steps of .05
R> lbs <- seq(0, .55, .05)
R> # Apply T-AISP using two-level methods
R> aisp(SWMDK[, -1], lowerbound = lbs, test.Hi = T, type.z = "WB",
R>      level.two.var = SWMDK[, 1])
R> # Investigate ICC
R> ICC(SWMDK[, 1:7]) # SWMD
R> ICC(SWMDK[, c(1, 9:12, 14)]) # SWMK
R> ### Step 2
R> # Investigate scale strength
R> coefH(SWMDK[, 2:7], level.two.var = SWMDK[, 1], ci = .95) # SWMD
R> coefH(SWMDK[, c(9:12, 14)], level.two.var = SWMDK[, 1], ci = .95) # SWMK
```

Figure 7.2: R syntax to obtain the main results of the two-step, test-guided MSA in the real-data example. `R>` denotes the R prompt and # precedes a comment.

five-item SWMK.

The average test score for the final, six-item SWMD was 3.20 ($SD = 0.68$) and of the final, five-item SWMK 3.97 ($SD = 0.68$). Table 7.4 shows the ICC per item and for the final scales. The ICC fluctuated across the items, indicating that some items have a larger group effect than others. The estimated ICC of the SWMD was .169, and was significantly larger than zero ($F(29, 609) = 5.31$, $p < .001$). The estimated ICC of the SWMK was .183, and was also significantly larger than zero ($F(29, 609) = 5.75$, $p < .001$). Hence, we continue the analyses using two-level methods.

In Step 2, the SWMD and the SWMK were evaluated as strong scales, as both (two-level) 95% confidence intervals exceeded the threshold of .5 for a strong scale (see Table 7.4, bottom row). For completeness sake, the point estimates, standard errors, and confidence intervals of the item coefficients for both scales are shown in Table 7.4. Given that Mokken's NIRT model holds, the scales could be used to order respondents using the sum scores on the scales, with all related (ordinal) measurement properties.

Comparison to the AISP

We now illustrate what would have happened if the traditional AISP would have been used that does not take into account sampling fluctuation and the nested data structure. Results showed that for $c \leq .2$ all items were included in one scale. For $c = .25$ item 12 was dropped from the one scale. For $c \geq .3$ two scales were formed, one containing

Table 7.3:

Scales Formed by the T-AISP for Increasing Values of Lowerbound c.

	c											
Item	.00	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55
1	1	1	1	1	1	1	1	1	1	1	1	0
2	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	2
5	1	1	1	1	1	1	1	1	1	1	1	2
6	1	1	1	1	1	1	1	1	1	1	0	0
7	1	1	1	1	1	0	0	0	0	0	0	0
8	1	1	1	2	2	2	2	2	2	0	3	0
9	1	1	1	2	2	2	2	2	2	2	2	3
10	1	1	1	2	2	2	2	2	2	2	2	3
11	1	1	1	2	2	2	2	2	2	2	3	0
12	1	1	0	2	2	0	0	0	0	0	0	0
13	1	1	1	2	2	2	2	2	2	2	2	0

Note. The number in each cell represents the scale to which the item was assigned. Unscalable items are denoted 0.

SWMD items (not including item 7 for $c \geq .4$) and the other containing SWMK items (not including item 12 for $c \geq .4$ and also item 8 for $c = .55$).

Comparing the results of the AISP to the results of the T-AISP demonstrates that, for this data example, similar scale patterns emerged across different values of the lowerbound. However, the AISP retained more items in less scales for larger lowerbounds. In addition, for $c = .3$ (the lowerbound that is often used as default) the AISP divided all items into two subscales, whereas the T-AISP considered item 7 and 12 unscalable. These differences are a consequence of the AISP not taking uncertainty of H_i into account, and this uncertainty is quite substantial in this small data example.

7.7 Discussion

This chapter introduced two major advancements in MSA: First, point estimates and standard errors for scalability coefficients were derived for clustered data (where respondents are nested in groups). Until now, estimating scalability coefficients and their standard errors were unavailable for clustered data. However, point estimates and standard errors for within-rater scalability coefficients, which are similar in interpretation as the original

Table 7.4:

Scalability coefficients, Standard Errors and Wald-Based Confidence Intervals Estimated Using the Two-Level Method, and ICCs for Each Item and the Total Scale of the SWMD and the SWMK.

SWMD					SWMK				
Item	$\widehat{H}$	SE	95% CI	ICC	Item	$\widehat{H}$	SE	95% CI	ICC
1	.609	.033	[.545; .674]	.120	8	.547	.036	[.477; .617]	.077
2	.641	.026	[.589; .693]	.111	9	.551	.036	[.480; .621]	.103
3	.619	.029	[.562; .676]	.103	10	.644	.025	[.595; .693]	.196
4	.634	.033	[.568; .699]	.142	11	.594	.031	[.532; .655]	.111
5	.650	.028	[.594; .705]	.082	12	-	-	-	-
6	.566	.031	[.506; .626]	.129	13	.627	.028	[.572; .682]	.097
7	-	-	-	-					
Total	.620	.026	[.570; .670]	.169	Total	.592	.025	[.543; .642]	.183

Note. $\widehat{H}$ = estimated scalability coefficient, SE = standard error, CI = confidence interval, ICC = intraclass correlation.

scalability coefficients, were available for multi-rater data (Koopman, Zijlstra, & Van der Ark, 2020), data that also have a two-level structure. We proposed a slight adaptation of the estimates that resulted in two-level methods, for which the point estimates are identical to the traditional (one-level) point estimates for scalability coefficients, but for which the standard errors are accurate and have little bias in clustered data. To keep the chapter readable, details of the estimation, accuracy, and bias of the point estimates and standard errors have been diverted to Chapter 6.

Second, a test-guided automated item selection procedure (T-AISP) was introduced. The traditional AISP tested only Criterion 1 of a Mokken scale ($H_{ij} > 0$), and evaluated Criterion 2 ($H_i > c$) by testing whether $H_i > 0$ and checking whether the point estimate $\widehat{H}_i \geq c$. By implementing a Wald-based test statistic in the T-AISP, we enabled the direct testing of both criteria of a Mokken scale. In addition, by using the newly developed standard errors based on the two-level method, the T-AISP could also be adapted to clustered data. As illustrated by a real-data example, the T-AISP is more conservative than the AISP. So when using the T-AISP, a researcher may expect that less items will end up in the scale. In addition, especially for large sets of items the computation time of the T-AISP may be considerably longer than the AISP. In future research, simulation studies using population covariance structures may show whether the T-AISP is indeed a better item-selection procedure than the AISP. A possible alternative for the Wald-based test statistic in the T-AISP would be the marginal modeling approach of Van der Ark et al.

(2008) for flexible and distribution-free testing of scalability coefficients. However, their test could only be applied to a limited number of dichotomous items. The comparison of their test to our test was beyond the scope of this chapter, but could be a topic of future research.

We integrated the two advancements into a two-step, test-guided MSA for scale construction, which is available for nonclustered and clustered data. The first step is performing a T-AISP and select final scale(s) with items that meet specified scalability criteria, using one-level methods for nonclustered and two-level methods for clustered data. For clustered data, the ICC is estimated on the final scale(s) and an F-test is performed to test whether the ICC is significantly larger than zero. In the second step, the strength of the scale(s) is determined using 95% Wald-based confidence intervals and further analyses are performed, using one-level methods for nonclustered data and for clustered data without within-group dependency, and two-level methods for clustered data with within-group dependency.

We applied the two-step, test-guided MSA to the 13-item scale SWMDK, intended to measure students' well-being at school with teachers and classmates. As data were collected using a cluster sampling design, two-level MSA methods were necessary. The T-AISP resulted in two reduced subscales (the 6-item SWMD and 5-item SWMK), both with significant within-group dependency. In the second step both scales were evaluated as strong scales. Note that the conclusions on the scalability were based on a subset of respondents from a larger dataset; when the procedure was applied to the original dataset two subscales were formed as well, but for each scale all items remained, resulting in a seven-item SWMD (evaluated as a medium to strong scale) and a six-item SWMK (evaluated as a strong scale), both with significant within-group dependency.

The next step in making MSA available for clustered data is generalizing the methods to investigate NIRT model fit, as these are critical for the implied measurement properties of Mokken's NIRT models. In addition, our proposed procedure can only handle two-level data. The procedure would benefit from a generalization to more complex sampling designs such as a three-level nested sample, for example if students are nested in classrooms that are nested in schools, or a cross-nested sample, where respondents are nested in more than one group.