British Journal of Mathematical and Statistical Psychology (2016)

The British Psychological Society

www.wileyonlinelibrary.com

Expert Tutorial

A tutorial on how to do a Mokken scale analysis on your test and questionnaire data

Klaas Sijtsma[1]* and L. Andries van der Ark[2]

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

Over the past decade, Mokken scale analysis (MSA) has rapidly grown in popularity among researchers from many different research areas. This tutorial provides researchers with a set of techniques and a procedure for their application, such that the construction of scales that have superior measurement properties is further optimized, taking full advantage of the properties of MSA. First, we define the conceptual context of MSA, discuss the two item response theory (IRT) models that constitute the basis of MSA, and discuss how these models differ from other IRT models. Second, we discuss dos and don'ts for MSA; the don'ts include misunderstandings we have frequently encountered with researchers in our three decades of experience with real-data MSA. Third, we discuss a methodology for MSA on real data that consist of a sample of persons who have provided scores on a set of items that, depending on the composition of the item set, constitute the basis for one or more scales, and we use the methodology to analyse an example real-data set.

1. Introduction

Mokken scale analysis (MSA) provides a set of statistical tools for constructing scales for measuring persons and items with respect to attributes from the personality and cognitive domain, health-related quality of life, sociology, marketing, and several other areas in which multi-item scales are used. The method uses relatively liberal assumptions that imply the ordering of persons on a scale by means of persons' total scores on a set of items. MSA has rapidly grown in popularity among applied researchers. The authors applaud the widespread application of the method but are concerned about its applications to real data. They devoted a modest literature search to substantiate their concern.

On 1 March 2016, using the date restriction 2015–present, the authors fed the search term 'Mokken scale' into Google Scholar to retrieve recent papers on MSA. Of the 176 hits, 85 (48%) were discarded because they were unavailable online, published before 2015, not written in English, did not report data analysis, or reported MSA without numerical results. Most selected articles (97%) reported scalability coefficients, but far fewer reported results from other MSA methods: the automated item selection procedure (AISP; 38%), monotonicity (32%), invariant item ordering (22%), and Mokken's reliability coefficient (22%). Other method results were rare or absent. The authors concluded that

*Correspondence should be addressed to Klaas Sijtsma, Department of Methodology and Statistics, TSB, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands (email: k.sijtsma@uvt.nl).

DOI:10.1111/bmsp.12078

most researchers equated MSA with the computation of scalability coefficients and did not perform additional analyses or did not report the results. By failing to do additional analyses, researchers fail themselves; and by failing to report results, researchers fail their colleagues. We provide this tutorial to help researchers further optimize the construction of scales having desirable measurement properties, taking full advantage of MSA.

The outline of the tutorial is as follows. First, we define the conceptual context of MSA, discuss the two item response theory (IRT) models on which MSA is based, and discuss how these models differ from other IRT models. Second, we discuss dos and don'ts for MSA; the don'ts include misunderstandings we have frequently encountered with researchers in our three decades of experience with real-data MSA. Third, we discuss a methodology for MSA on real data that consist of a sample of persons who have provided scores on a set of items that, depending on the composition of the item set, constitute the basis for one or more scales. We use the methodology to analyse an example, real-data set with respect to Type D personality (Denollet, 2005). When appropriate, we refer to other measurement models but concentrate on MSA, which is the topic of this tutorial.

2. Theory of Mokken scale analysis

2.1. Conceptual context: Definitions and notation

A real-data set suited for MSA consists of a sample of N persons who have provided scores on a set of J_{start} items. Items are indexed j, such that $j = 1, \ldots, J_{\text{start}}$. For simplicity, we assume that all items have the same scoring format, so that all items have equal weight. A common scoring format is the well-known Likert scoring, allowing persons to indicate for each item how much they agree with a particular statement about, for example, their personality, how they cope with a particular disease, how they experience their religiosity, and their preferences as a consumer. Likert items frequently have five ordered response categories, and the person chooses one category to express the degree to which they agree with the statement. This results in scores that run from 0, indicating little or no agreement, to 4, indicating high or perfect agreement; the labelling of response options depends on choices the researcher makes. Researchers may choose to have fewer or more ordered categories and item scores. Saris and Gallhofer (2007, chap. 5) discuss feasible numbers of answer categories, but in practice one rarely encounters more than, say, seven ordered answer categories. The smallest number equals two, frequently encountered in the measurement of maximum performance as in cognitive and educational measurement when answers are either incorrect or correct.

The description so far expresses preferences found in practice, but there is nothing in principle that would prevent dichotomous disagree/agree scoring from replacing Likert scoring, and ordered, polytomous scoring expressing the degree to which the solution a person provided approaches the ideal solution from replacing incorrect/correct scoring. Let the score on item j be denoted by random variable X_j with realization x_j; then item scores equal $x_j = 0, \ldots, M$. By definition, a higher item score expresses a higher attribute level.

Remarks

(1) In test and questionnaire construction, researchers sometimes use different scoring formats for different items. This amounts to weighting the items differently. Two justifications may be encountered. First, based on a theory about the attribute, it may be argued that one particular aspect of the attribute covered by item A may better characterize the attribute than another aspect covered by, say, item B. This may

justify giving a larger weight to item A, for example, by assigning higher scores to the answer categories or using more ordered answer categories, thus producing a greater score range. To our knowledge, theories about attributes rarely provide this level of detail, and thus provide little if any guidance for differential item weighting. Second, one may estimate item weights from the data without the guidance of a substantive theory. However, different samples for item analysis often come from different populations that differ from each other and also across time, hence rendering the sample used for weight estimation rather coincidental, also leaving the weights coincidental and thus rather meaningless.

(2) A principled argument against differential item scoring is that it produces meaningless total scores. For example, when a three-point Likert item is scored 0, 1, 2, and a five-point Likert item is scored 0, 1, 2, 3, 4, then, assuming both items are positively formulated, "agree" yields 2 credit points for the first item and 4 credit points for the second item. Here, technical considerations about the numbers of answer categories for different items determine differential weighing and cannot produce useful total scores. Even if one scores the first item, say, 0, 2, 4, the absence of two scores (1 and 3) still impairs total-score interpretation. The same problem appears when one collapses sparse score categories *ad hoc*. For example, transforming a five-point Likert item scored 0, 1, 2, 3, 4 to a dichotomously scored item raises the question whether the scoring should be 0, 1 or 1, 4, or something else. No compelling logic exists for how to do this and the result is arbitrary, not leading to well-justified total scores.

(3) Different item-score schemes are possible. For example, one may ask persons to express their level of agreement on a line segment running from 0 to 100% (Saris & Gallhofer, 2007, pp. 114–116). This may be a useful approach with other scaling methods, but unless one divides the percentage scores into a limited number of ordered categories, cross-tables are too sparse to be of practical use in MSA. Data such as response times (Van der Linden, 2006), scores resulting from asking persons to determine their mental distance from a stimulus (i.e., preferences; Andrich, 1989), and scores resulting from direct comparison of stimuli (i.e., paired comparisons; Cattelan, 2012) require different measurement models for constructing scales.

2.2. The models

MSA uses two different non-parametric IRT models to construct scales. Non-parametric models employ less restrictive assumptions about the data than most other, often parametric, IRT models, and they typically focus on detailed model fit investigation and data exploration to understand the test, its items and the population of interest in more depth (Junker & Sijtsma, 2001). The non-parametric models imply ordinal scales for persons and items based on observable test scores, defined by $X_+ = \sum_{j=1}^{J} X_j$, and item mean scores, respectively. Ordinal scales limit the models' use for equating and adaptive testing that are more typical of parametric IRT, but the latter models often produce worse model fit in particular applications, hence challenging their justification. For more information about non-parametric IRT, see Ramsay (1997) and Stout (2002).

2.2.1. A model for ordering persons

The relevance of the monotone homogeneity model (Mokken, 1971; Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002, 2016) is that it implies an ordinal scale for person

measurement using the observable test score. The three assumptions of the model are as follows:

(1) *Unidimensionality*. The single attribute that all items in a scale measure is quantified by means of a latent variable denoted θ.

(2) *Monotonicity*. As θ increases, the probability of scoring at least x_j on item j increases or remains constant, but cannot decrease; that is, the more one possesses of the attribute, the more likely one obtains scores representative of responses typical of the higher attribute level. The relationship between the probability of obtaining at least score x_j and latent variable θ, $P(X_j \geq x_j|\theta)$, is known as the item step response function (ISRF), and is defined for $x_j = 1, \ldots, M$. For $x_j = 0$, by definition $P(X_j \geq 0|\theta) = 1$, which is uninformative about the relation between the item score and the latent variable. If the item has two scores, one ISRF remains, then called the item response function (IRF), $P(X_j = 1|\theta)$. Figure 1 shows monotone non-decreasing ISRFs for two items j and k, each having four ordered scores, hence three ISRFs (solid and dashed curves).

(3) *Local independence*. Items measuring the same attribute correlate positively when people vary with respect to θ. That is, because, compared to people lower on θ, people higher on θ are expected to obtain higher item scores on each item measuring θ, the scores on different items measuring θ covary and are positively related. However, if one removes this source of variation, for example, by selecting a subgroup of people having the same θ value, the relationship between the items vanishes. If the scale is unidimensional, θ is the only source of variation and conditioning on θ renders the items independent. Statistically, local independence means that

$$P(X_1 = x_1, \ldots, X_J = x_J|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta),$$

meaning that the joint distribution of the J item scores equals the product of the J marginal distributions of the separate item scores. This property implies that the conditional covariance between items equals 0; for items j and k, $\text{Cov}(X_j, X_k|\theta) = 0$, also known as weak local independence (Stout, 1990).

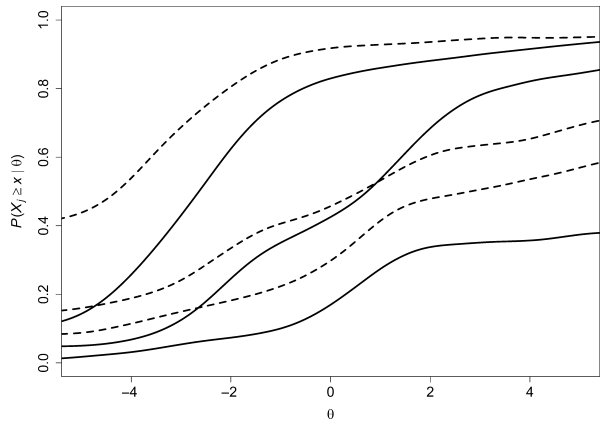

Figure 1. Three monotonically increasing ISRFs for item j (solid curves) and item k (dashed curves).

Remarks

(1) Typical of the monotone homogeneity model is that it does not allow estimation of θ but that it justifies and uses test score X_+ to order persons on latent variable θ (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997; Van der Ark & Bergsma, 2010). Hence, a monotone homogeneity model that fits the data enables ordering people on θ by means of their test scores X_+.

(2) Unidimensionality makes sense because measurement instruments in principle intend to measure one attribute at a time, just as a thermometer measures temperature and nothing else. However, in social and behavioural science measurement multidimensionality frequently appears in two ways. First, an attribute may consist of two or more sub-attributes, rendering it necessary to distinguish scales for each of the sub-attributes. For example, Denollet (2005) hypothesized that Type D personality consists of negative affectivity and social inhibition. In this case, one may investigate whether one needs one or two scales; see the real-data example. Second, in addition to the intended attribute, sets of unbidden attributes drive responses to items and threaten to dilute the targeted single-attribute measurement. For example, language skills influence responses to rating-scale items for personality measurement. In this case, one may investigate whether one scale suffices or whether some responses are so heavily laden with unwanted language influences that the items involved should be deleted, or whether it makes sense, for example, to distinguish two scales, one measuring the intended attribute relatively free of language influences and the other including such influences. MSA uses an AISP to separate items into different scales and identify deviating items.

(3) Figure 2 shows the three ISRFs of an item, two of which show non-monotonicities. For function $P(X_j \geq 2|\theta)$, the non-monotonicity extends along the high end of the latent variable θ scale, and deviations are shallow in the vertical direction. For function $P(X_j \geq 1|\theta)$, one non-monotonicity extends between $\theta = -1$ and $\theta = 2$ and is deep. The latter violation of the monotonicity assumption probably is more damaging to the degree to which the ordering of people by means of X_+ reflects their ordering on θ. MSA allows estimation of the ISRFs and assessment of the non-monotonicities.

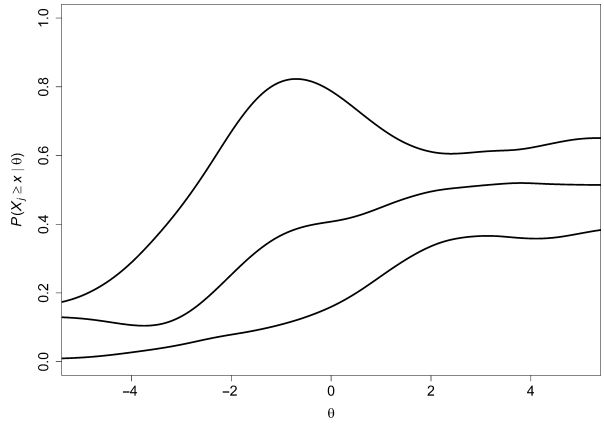

Figure 2. Three ISRFs: $P(X_j \geq 3|\theta)$ is monotonically non-decreasing in θ, $P(X_j \geq 2|\theta)$ is slightly decreasing for the extreme positive values of θ, and $P(X_j \geq 1|\theta)$ shows a sharp decrease between $\theta = -1$ and $\theta = 2$.

(4) Given that measurement of attributes is not likely to be purely unidimensional, conditioning on θ is not enough to secure local independence because within groups having the same θ value people will still vary on other latent variables, hence producing covariation between items. Another source of local dependence is that measurement quality often varies across different populations, meaning that scale properties are often different for different gender, cultural, and education groups. For example, for two persons having the same θ level who are members of two different groups, the same item may have different response probabilities. This is the phenomenon of differential item functioning (Holland & Wainer, 1993). MSA allows scale properties in different groups to be assessed.

2.2.2. A model for ordering persons and items

The double monotonicity model is a special case of the monotone homogeneity model. In addition to being an ordinal person-measurement model, the double monotonicity model implies the ordering of items by means of mean item scores. Several intelligence tests present items by descending mean score, so that respondents start with the easiest items. Hence, they can take some time to overcome possible test anxiety, whereas excessively difficult starting items do not discourage other respondents. For this to work, one has to establish that the item ordering is equal for different-ability respondents. The double monotonicity model implies such an invariant item ordering. It uses the same assumptions as the monotone homogeneity model, and adds the fourth assumption of non-intersecting IRFs.

(4) *Non-intersecting IRFs.* We define the IRF as follows,

$$\varepsilon(X_j|\theta) = \sum_{x_j=1}^{M} P(X_j \geq x_j|\theta).$$

For $M = 1$, we have $\varepsilon(X_j|\theta) = P(X_j = 1|\theta)$. The double monotonicity model assumes that the IRFs of the J items do not intersect. For three items, Figure 3 shows the non-intersecting IRFs, $\varepsilon(X_j|\theta)$. Items whose IRFs do not intersect have an invariant item ordering (Sijtsma & Hemker, 1998; Sijtsma & Junker, 1996); that is, an

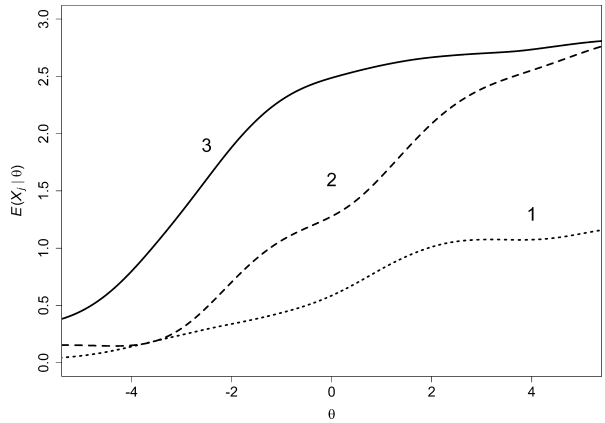

Figure 3. Three IRFs enumerated 1, 2, and 3, exhibiting an invariant item ordering.

ordering that, except for possible ties, is same for each θ value. Algebraically, invariant item ordering is defined as follows:

(a) Let the mean item score be the expectation, $\varepsilon(X_j) = \int \varepsilon(X_j|\theta)\mathrm{d}G(\theta)$, where $G(\theta)$ is the cumulative distribution of latent variable θ. Based on a standard normal θ, for the three items in Figure 3 their means are $\varepsilon(X_1) \approx 0.59$, $\varepsilon(X_2) \approx 1.28$ and $\varepsilon(X_3) \approx 2.42$.

(b) The items in Figure 3 have already been ordered such that item 1 has the smallest mean and item 3 the greatest mean, but in real data one has to first order the J items by their sample means and number the items such that (for simplicity, assuming absence of ties)

$$\bar{X}_1 < \bar{X}_2 < \ldots < \bar{X}_J,$$

which estimates the unknown population ordering (which can be different, but we use the same item indices for simplicity)

$$\varepsilon(X_1) < \varepsilon(X_2) < \ldots < \varepsilon(X_J).$$

(c) Assume that if, for any pair of items $j < j+1$, for at least one θ value we know that $\varepsilon(X_j|\theta) < \varepsilon(X_{j+1}|\theta)$, then, assuming that IRFs do not intersect, for all J items it follows that

$$\varepsilon(X_1|\theta) \leq \varepsilon(X_2|\theta) \leq \ldots \leq \varepsilon(X_J|\theta), \text{ for all } \theta.$$

In Figure 3, the three IRFs do not intersect but the IRFs of items 1 and 2 touch in the interval $-4 < \theta < -3$; hence, they exhibit an invariant item ordering.

Items having an invariant ordering facilitate the interpretation of the scale, because for each θ value item 1 has the smallest mean, depending on the measurement context rendering it the least popular or the most difficult item, followed by item 2, and so on, until item J, which is the most popular or the easiest. Hence, the relation between total score X_+ and the individual items lends the former more meaning.

Remarks

(1) IRFs for real data often do not exhibit an invariant item ordering. In fact, many intersections may occur simply because J functions, even when they are monotone non-decreasing, can cross one another in many ways, implying that item orderings can vary greatly across different θ values. Meijer and Egberink (2012) critically discuss invariant item ordering for clinical scales covering a limited range of symptoms.

(2) If, for each person, the same item is the least popular or the most difficult, the same item is the second least popular or second most difficult, and so on, this greatly facilitates the interpretation of the test scores. For example, for dichotomously scored items (0/1 scoring), a higher X_+ score implies that one not only expects that a person answered the same items correctly as a person having a lower X_+ score, but also one or more additional, more difficult items. For polytomously scored items (e.g., 0/1/2/3/4 scoring), the higher the X_+ score, the more a person endorses the individual items and endorsement is stronger as items are more popular. This logic lends meaning to test performance in addition to only having more items correct or more often having endorsed items as X_+ grows without knowing which items. Typically, the monotone homogeneity model allows test-score interpretation of the

latter, quantitative kind, while the double monotonicity model allows test score interpretation of the former, more qualitatively interesting kind.

(3) Molenaar (1997) originally defined the double monotonicity model for non-intersecting ISRFs rather than IRFs (Ligtvoet, Van der Ark, Te Marvelde, & Sijtsma, 2010). One may notice that, by definition, the M ISRFs of an item cannot intersect, but the sets of M ISRFs of different items can intersect (Figure 1). Molenaar's original model thus required a total of $J \times M$ ISRFs not to intersect. However, in real data is it likely that many ISRFs of different items intersect and, consequently, that the model shows misfit. More important, theoretically, a set of $J \times M$ non-intersecting ISRFs do not imply an invariant item ordering; that is, an invariant ordering of the J IRFs, $\varepsilon(X_j|\theta)$ (Sijtsma, Meijer, & Van der Ark, 2011). Hence, requiring $J \times M$ ISRFs not to intersect does not imply useful measurement properties and thus does not serve a useful purpose. By definition, non-intersection of the J IRFs implies an invariant item ordering, and we focus on IRFs.

2.3. A practical tool for selecting scales from item sets

2.3.1. IRT models and item quality

IRT models restrict the data distribution only partially, leaving much opportunity for weak items to slip into a scale. This is the reason why one must make decisions about item inclusion even if one has ascertained that the IRT model fits the data well. The issue is the following. Focusing on non-parametric IRT approaches, these models require the IRF slopes to be non-negative (see also Stout, 1990), which implies a non-negative correlation between an item and the latent variable, but the correlation is allowed to freely range between 0 and 1, the data determining the magnitude of the correlations. Non-parametric model fit addresses IRF monotonicity, but not IRF steepness related to the test-score distribution and expressed by an index that relates steepness to the test-score distribution. This information bears directly on the item's practical usefulness (Mokken, Lewis, & Sijtsma, 1986). One may notice that the same issue exists in parametric IRT.

Focusing on non-parametric IRT, one thus needs a statistical tool that separates items having low or high quality in relation to the test-score distribution. In MSA, the scalability coefficients play this role. Scalability coefficients are used to assess item quality in a given set of items or as an item selection tool in AISP (Sijtsma & Molenaar, 2002, chaps. 4 and 5).

2.3.2 Automated item selection procedure

The feature that probably has won MSA its popularity is the facility of AISP in the accompanying software (Molenaar & Sijtsma, 2000; Van der Ark, 2007, 2012). Conceptually, depending on the composition of the item set, AISP identifies K scales, indexed $k = 1, \ldots, K$; then the number of items in scale k is denoted J_k. The initial item set the researcher feeds to AISP is often experimental in the sense that the researcher has assembled the item set based on theory about the attribute of interest and is not yet sure whether all items have sufficient psychometric quality for selection in the final scales. For example, some items may not discriminate well between persons located low and persons located high on the attribute scale. These weakly discriminating items do not contribute to a reliable person ordering. AISP identifies such items and, depending on user-defined choices, may exclude them from a final scale. In addition, different sub-attributes may be distinguished or unbidden attributes different from the intended attribute may drive

responses to a subset of items or one or two items. AISP identifies subscales and deviating items. In addition, badly targeted items not belonging to any of the scales may remain behind; say, there are J_{unsc} of such unscalable items. Finally, the researcher may define a kernel of items that he considers key to the attribute and run AISP using the kernel as the point of departure. By definition, the kernel is in the scale and additional items that scale well with the kernel may complete the scale.

In its simplest form, AISP produces one scale that includes all the items from the initial set; that is, $K = 1$. A result found more often includes one or more scales and a couple of unscalable items. Then the items distribute across scales and an unscalable category, such that $J_{\text{start}} = \sum_{k=1}^{K} J_k + J_{\text{unsc}}$, and in the simplest case (all items together constitute one scale) $J_{\text{start}} = J_1$ or simply $J_{\text{start}} = J$.

AISP uses scalability coefficients for pairs of items, denoted H_{jk}, and individual items, denoted H_j (Mokken, 1971, pp. 148–153; Sijtsma & Molenaar, 2002, chap. 4). Let σ_{jk} denote the covariance between items j and k, and let $\sigma_{jk}^{\max}$ denote the maximum possible covariance given the marginal distributions of the two item scores; then

$$H_{jk} = \sigma_{jk} / \sigma_{jk}^{\max}.$$

To obtain H_j, we define the total score $R_{(j)}$ on $J - 1$ items except item j, and define

$$H_j = \sigma_{X_j R_{(j)}} / \sigma_{X_j R_{(j)}}^{\max}.$$

Thus, the item scalability coefficient is a normed corrected item–test covariance, but attains values much different from, for example, corrected item–test correlations.

The monotone homogeneity model implies that

$$0 \leq H_{jk} \leq 1 \text{ and } 0 \leq H_j \leq 1;$$

hence, negative values logically contradict the model and positive values tend to support the model, but especially smaller values do not rule out data that are multidimensional, locally dependent or non-monotone (e.g., Mokken *et al.*, 1986; Smits, Timmerman, & Meijer, 2012).

Technically, AISP selects items from an available set into one or more scales using a scale definition based on H_{jk} and H_j (Mokken, 1971, p. 184; Sijtsma & Molenaar, 2002, pp. 67–69). A set of items constitutes a scale if, for a suitable chosen positive constant c,

(a) for inter-item scalability coefficients $H_{jk} > 0$, for all $j, k, j \neq k$; and

(b) for item scalability coefficients $H_j \geq c > 0$, for all j.

Requirement (b) implies that the total-scalability coefficient, denoted H, and defined as

$$H = \frac{\sum_{j=1}^{J} \sigma_{X_j R_{(j)}}}{\sum_{j=1}^{J} \sigma_{X_j R_{(j)}}^{\max}},$$

also equals at least c; that is, $c \leq H \leq 1$ (maximum value equals 1).

The monotone homogeneity model is related to the scale definition by implying $H_{jk} > 0$ and $H_j > 0$, but not $H_j \geq c > 0$. The latter additional requirement forces the rejection of items discriminating weakly between different θ values, which is reflected by H_j values close to 0, and only accepts items that discriminate well, which is reflected by higher H_j values. AISP allows the researcher to control the value of c, letting him decide what he finds reasonable. As an aid, software packages use the default $c = .3$. The researcher may use the following rules of thumb: $.3 \leq H < .4$ constitutes a weak scale;

$.4 \leq H < .5$ a medium scale; and $H \geq .5$ a strong scale; a set of items for which $H < .3$ is considered unscalable.

Although experience has shown that the default $c = .3$ is useful in real-data analysis, following Hemker, Sijtsma, and Molenaar (1995) we recommend running AISP 12 times consecutively using $c = 0, .05, .10, \ldots, .55$, and looking for one of the following two typical outcome patterns:

(1) In unidimensional data, as c increases one subsequently finds
   (a) most or all items in one scale;
   (b) one smaller scale; and
   (c) one or a few small scales and several unscalable items.
   Take the result in stage (a) as final.

(2) In multidimensional data, as c increases one subsequently finds
   (a) most or all items in one scale;
   (b) two or more scales; and
   (c) two or more smaller scales and several unscalable items.
   Take the result in stage (b) as final.

The messy structure of real data may obscure these ideal outcome patterns, requiring researchers to draw their own conclusion.

Remarks

(1) Van Abswoude, Vermunt, Hemker, and Van der Ark (2004) and Brusco, Köhn, and Steinley (2015) suggested alternative algorithms for item selection that are interesting but have not been implemented in MSA software. The genetic algorithm of Straat, Van der Ark, and Sijtsma (2013) is an explicit attempt to improve upon AISP. The R package mokken (Van der Ark, 2007, 2012) includes the genetic algorithm. Because of its close relationship to AISP, we used the genetic algorithm in our real-data example and compared it to AISP.

(2) Researchers sometimes incorrectly assume that AISP selects scales that have an invariant item ordering. In fact, the defining feature of the double monotonicity model, non-intersecting IRFs, does not mathematically restrict H; that is, sets of intersecting and non-intersecting IRFs can have the same H values (Sijtsma, *et al.*, 2011). This means that H values do not distinguish sets of intersecting IRFs from sets of non-intersecting IRFs, and higher H values thus do not necessarily produce scales having an invariant item ordering.

(3) Unscalable items have low quality in the context of the item set and the population in which one collected the data, but, unless these items were badly constructed, they may function well in other scales or different populations. For example, a calculus item may show misfit in an arithmetic test but scale well in an advanced algebra test. It will also be out of place in a primary-school population but may scale well in a graduate population.

(4) For *a priori* defined scales that one wishes to keep intact as much as possible, one might first compute H_{jk}, H_j, and H coefficients for the complete scales. If scalability values are too low, one might run AISP on each scale or on the complete item set to study its dimensionality and obtain evidence for a new scale structure.

2.4. Misunderstanding the two measurement models and AISP

During the past two decades in which MSA has become quite popular with scaling practitioners, we have seen many fascinating applications but also witnessed a number of

regularly recurring misunderstandings that tend to lead researchers to make incorrect decisions or use their scales for purposes for which they were not suited. We mention three general misunderstandings, and then go one to describe a methodology for MSA.

2.4.1. Confusion of the two models

Several researchers do not seem to be aware that MSA includes two IRT models, both implying a person ordering but only one, the double monotonicity model, implying an (invariant) item ordering. We have noticed that researchers often take the latter model as the default model and the only model. This would not really be a problem if researchers adequately assessed the goodness of fit of the double monotonicity model to the data, because the double monotonicity model is a special case of the monotone homogeneity model and has both ordering properties. However, we have noticed that researchers regularly do not even attempt to assess the double monotonicity model's goodness of fit but rather assess (aspects of) the monotone homogeneity model without seeming to realize they do. This suggests that researchers often are unaware that they assess (aspects of) the wrong model and assume they have assessed the other model. We will demonstrate how to assess each model.

2.4.2. Limiting the analysis to automated item selection

If available, AISP selects the initial J_{start}-item set into K scales, using the formal definition of a scale and default choices for some procedural features such as minimum item scalability ($H_j \geq c > 0$) and statistical significance level used for testing hypotheses about H coefficients. We have noticed, also in our limited literature search, that the availability of AISP makes life so easy for the researcher that as a rule they limit MSA to just running the procedure and forgetting to assess the monotonicity of the IRFs of the items selected in each scale, local independence and, if deemed desirable, whether items are invariantly ordered. Three comments seem to be in order:

(1) AISP does not assess monotonicity, but if an IRF shows gross violations of monotonicity, this will tend to lower the item's scalability coefficient H_j and AISP will likely not select the item in a scale. However, there is no guarantee that non-monotonicities always produce $H_j < c$, because this also depends on the θ distribution, and steep IRFs with smaller violations may produce $H_j \geq c$ and go unnoticed.

(2) AISP does not assess items based on statistical features of an invariant ordering. Invariant item ordering is a characteristic of an item set that has to be assessed separately (Sijtsma *et al.*, 2011). We have already seen that researchers tend to assume an invariant item ordering to hold without having checked the empirical evidence.

(3) In non-parametric IRT, local independence has received attention at the theoretical level (Ellis, 2014; Holland & Rosenbaum, 1986) and the data-analysis level (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996; Zhang & Stout, 1999). Straat, Van der Ark, and Sijtsma (in press) proposed a method for investigating local independence; see the next section.

2.4.3. Limiting the automated item selection to default choices

The availability of the user-friendly AISP, including default choices such as $c = .3$, unfortunately renders it a hit-and-run procedure, discouraging researchers from considering alternatives to the defaults, thus limiting the procedure's possibilities and failing

their goal. The real-data example will use the values $c = 0, .05, .10, \ldots, .55$, and take better advantage of the possibilities of AISP and MSA.

3. Mokken scale analysis in practice

A scale analysis is complex and can be done in many different ways, but here we propose a three-stage procedure, the stages being data examination, scale identification and scale properties, consisting respectively of three, four, and three analysis steps. We describe the ten steps in chronological analysis order and then illustrate the whole procedure using a real-data example.

3.1. Ten steps for doing a Mokken scale analysis

We describe the three stages containing ten steps for doing an MSA:

i. *Data examination*. Examine the data and take appropriate measures when particular data problems occur (steps 1–3):
   (1) *Recoding.* Recode scores of items that are negatively worded relative to the attribute of interest, so that for all items higher scores mean a higher position on the attribute scale.
   (2) *Inadmissible scores and missing data.* Treat inadmissible scores as missing values. Determine the total percentage of missing item scores in the data set, and the respondents that left open an unreasonable number of answers. Ask yourself why so many data are missing and why particular respondents produced so many missing values. If, say, more than 10% of the total data are missing, was there something wrong with the study design or the wording of particular items? If a respondent left open, say, at least 30% of the answers, did they take the task seriously? We impute missing item scores using two-way imputation (Bernaards & Sijtsma, 2000; Van Ginkel, Van der Ark, & Sijtsma, 2007), but many other possibilities are feasible. For handling completed data sets resulting from multiple imputation, see Van der Ark and Sijtsma (2005).
   (3) *Outliers.* Identify whether particular item-score patterns qualify as outliers, because many unpopular or difficult items received high scores while many popular or easy items received low scores. We advise performing the scale analysis separately on the complete data and on the data without the identified outliers. If the removal of a small number of outliers greatly influences the scaling results, removal seems to be justified, as one cannot accept that only few observations greatly determine scaling results. We used the number of Guttman errors, denoted by index G_+, in combination with Tukey's fences for outlier detection (Zijlstra, Van der Ark, & Sijtsma, 2011).

ii. *Scale identification*. Identify one or more scales that satisfy either the model of monotone homogeneity and, if deemed appropriate, check whether the scales also satisfy the more demanding model of double monotonicity (steps 4–7):
   (4) *Scalability.* If one wishes to assess *a priori* defined scales, compute H_{jk}, H_j and H coefficients for the complete scales. If one wishes to explore the item set for its dimensionality, perform AISP using $c = 0, .05, .10, \ldots, .55$, and use Hemker *et al.* (1995) to look for relevant outcome patterns. Depending on the data, the outcome patterns may have become clear before the highest c values are reached. AISP roughly sorts items in scales that order people using X_+ without making big mistakes but may miss a couple of nuances needed for making finer-grained decisions; see steps 5–7.

(5) *Local independence*. Investigate local independence using the conditional association procedure (Straat *et al.*, in press). The conditional association procedure involves two indices W_1 and W_3 flagging locally dependent item pairs.
(6) *Monotonicity*. Investigate monotonicity of IRFs using a non-parametric regression method for an item score on the total score on the other $J - 1$ items in the same scale (Junker & Sijtsma, 2000; Sijtsma & Molenaar, 2002, 2016). Graphical analysis provides an impression of the degree to which an observed curve can be considered monotone, and local deviations from monotonicity can be tested for statistical significance.
(7) *Invariant item ordering.* Researchers only wishing to construct a scale that orders persons on one dimension may skip this step. However, if one requires an invariant item ordering, one may use the search procedure suggested by Ligtvoet *et al.* (2010). The scalability coefficient H^T expresses the degree to which respondents order items invariantly (Ligtvoet *et al.*, 2010; Sijtsma & Meijer, 1992).

iii. *Scale properties*. Determine scale properties of the scales identified in the second stage (steps 8–10).
(8) *Reliability.* Use the Molenaar–Sijtsma (MS) method (Sijtsma & Molenaar, 1987) to estimate test-score reliability. The MS method assumes the double monotonicity model.
(9) *Norms.* Researchers requiring norm tables for the measurement of individuals may estimate norms and confidence intervals for the norms using a regression procedure (Oosterhuis, Van der Ark, & Sijtsma, 2016). In scientific research, one does not need norms for interpreting individuals' test performance; in this case, one may skip step 9.
(10) *Group comparison.* If the sample contains meaningful subgroups, one may investigate whether the composition of the scales and the scale properties (steps 4–9) can be generalized across the subgroups. If the scale composition and the scale properties vary across groups, knowledge of this variation may be helpful when interpreting individuals' test performance and in research where group characteristics such as mean scale scores and correlations of scale scores with other variables are of interest.

3.2. The ten-step procedure in action: an MSA of the DS14 Type D Scale

We used a sample ($N = 541$; 68 women and 473 men) of age ranging from 23 to 89 years ($M = 58.7$), also used by Denollet, Pedersen, Vrints, and Conraads (2013); see Straat, Van der Ark, and Sijtsma (2014) for sample size recommendations for MSA. Respondents were patients suffering from mild coronary disease, who were administered a battery of questionnaires, among them the DS14, a 14-item questionnaire measuring Type D personality (Denollet, 2005). Seven items measure negative affectivity (NA) and the other seven items measure social inhibition (SI); see Figure 4. The sum of the 14 item scores, referred to as the DS14 score, measures Type D personality. Items consist of a statement followed by a five-point rating scale, scored $x = 0, \ldots, 4$. For scale analysis, we used R Version 3.3.1 (R Core Team, 2016). The analysis script is available from the supplementary material website. All steps were conducted using the R package `mokken` (Van der Ark, 2007, 2012).

(i) Data examination

Step 1. *Recoding.* Scores of items 1 and 3 were recoded as $x^* = 4 - x$.

Step 2. *Inadmissible scores and missing data.* Inadmissible scores were absent, all scores on sex and age were present, but eight respondents had one missing item

Instruction:
Below are a number of statements that people often use to describe themselves. Please read each statement and then *circle* the appropriate *number* next to that statement to indicate your answer. There are no right or wrong answers: Your own impression is the only thing that matters.

0 = false, 1 = rather false, 2 = neutral, 3 = rather true, 4 = true

1. I make contact easily when I meet people	0	1	2	3	4
2. *I often make a fuss about unimportant things*	0	1	2	3	4
3. I often talk to strangers	0	1	2	3	4
4. *I often feel unhappy*	0	1	2	3	4
5. *I am often irritated*	0	1	2	3	4
6. I often feel inhibited in social interactions	0	1	2	3	4
7. *I take a gloomy view of things*	0	1	2	3	4
8. I find it hard to start a conversation	0	1	2	3	4
9. *I am often in a bad mood*	0	1	2	3	4
10. I am a closed kind of person	0	1	2	3	4
11. I would rather keep other people at a distance	0	1	2	3	4
12. *I often find myself worrying about something*	0	1	2	3	4
13. *I am often down in the dumps*	0	1	2	3	4
14. When socializing, I don't find the right things to talk about	0	1	2	3	4

Figure 4. Items from the DS14 Type D personality questionnaire. Text in italics refers to items measuring NA, text in roman refers to items measuring SI. Items 1 and 3 need to be reverse-coded.

score, and one respondent had two missing item scores (i.e., 0.13% missing in total). Five of the ten missing scores occurred with item 2, while other missing scores were scattered across items. We repeated two-way imputation (Bernaards & Sijtsma, 2000) 10,000 times, thus producing 10,000 completed data sets. Differences between inter-item correlations did not exceed 0.01. Hence, we used one completed data set for MSA.

Step 3. *Outliers.* Given the null situation in which outliers are absent and G_+ is normally distributed, based on the criterion values derived from Tukey's fences (i.e., $G_+ = 202.5$) we expected 0.35% outliers but we found 4.8% (i.e., 26 cases). Given an extremely skew G_+ distribution (Figure 5), we computed the adjusted boxplot (Hubert & Vandervieren, 2008) to accommodate the skewness. This produced a criterion value $G_+ = 632.1$, without suspicious item-score patterns. Index G_+ correlated weakly positive with the DS14 score (.28), the NA score (.32), and the SI score (.14).

(ii) Scale identification

Step 4. *Scalability.* Including the outlier had a negligible effect on the scalability results; hence, it was included. For the NA items, for item pairs $.41 < H_{jk} < .73$ ($.03 < SE < .05$) and for items $.49 < H_j < .62$ ($.02 < SE < .03$) (Table 1). Hence, the NA items formed a strong scale ($H = .55$, $SE = .02$). For the SI items, for item pairs $.40 < H_{jk} < .67$ ($.03 < SE < .05$), and for items $0.45 < H_j < .57$ ($.02 < SE < .03$) (Table 1). For the whole SI scale, $H = .52$ ($SE = .02$), but the standard error was too large to conclude that in the population $H > .5$. Hence, the SI scale had medium strength. For the 14 DS14 items, we found $H = .36$ ($SE = .02$) but the items did not constitute a weak Mokken scale because $H_{2,3}$ and $H_{3,5}$ were negative; see requirement (b) of the scale definition. Several other item-pair scalability coefficients were not significantly greater than 0. Taking standard errors into account, item scalability coefficients H_2, H_5, and H_{11} were significantly greater than the conventional lower-bound value of .3 (Table 1, column 14).

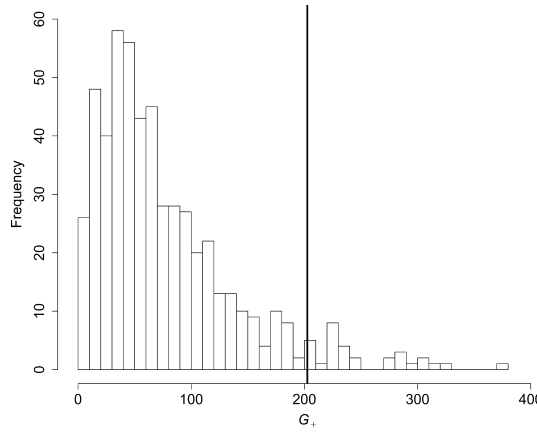

Figure 5. Distribution of item-pair based outlier score G_+ and the criterion value according to Tukey's fences (solid, vertical line).

We ran AISP for $c = 0, .05, .10, \ldots, .30$, and found one scale containing 13 items, all except item 3. For $c = .35$, items 2 and 5 fell out of the scale and together formed a second scale. For $c = .40$, we found the NA and SI scales. For $c = .45$, item 3 was unscalable, and $c > .45$ produced more than two scales and several unscalable items. For $c \geq .35$, the genetic algorithm version of AISP (Straat *et al.*, 2013) produced partitionings somewhat different and difficult to interpret. Thus, using AISP, we found support for NA and SI scales. Had it not been for item 3, the DS14 scale would have had medium strength.

Step 5. *Local independence*. For NA, the conditional association procedure using indices W_1 and W_3 did not flag any item. For SI, a large W_1 index suggested that the negatively worded items 1 and 3 were positively locally dependent. Index W_3 flagged item pair (3, 6), suggesting the items may be negatively locally dependent. Consistent with steps 3 and 4, this result suggested item 3 might be revised. Without item 3, index W_1 flagged item pairs (8, 14) and (10, 14) as positively locally dependent. For the whole DS14, the indices identified items 1 and 3 as positively locally dependent, and present in several negatively locally dependent item pairs with other items. Without items 1 and 3, the remaining data show no evidence of local dependence.

Step 6. *Monotonicity*. For NA, data analysis supported manifest monotonicity. For SI, one ISRF of item 6 showed one significant decrease, but the resulting IRF was not affected (Figure 6, top). For the whole DS14, the ISRFs of items 2 and 11 showed a significant decrease, but the effect on the IRF (Figure 6, middle and bottom, respectively) was minimal.

Step 7. *Invariant item ordering.* For NA, SI and DS14 items, visual inspection suggested that several IRFs were almost identical, so that establishing an invariant item ordering was difficult. The most rigorous method to investigate invariant item ordering (called *increasing in transposition*; Ligtvoet, Van der Ark, Bergsma, & Sijtsma, 2011) suggested that only four of the 14 items (NA: 4, 13; SI: 10, 14) did not show signs of violating invariant item ordering. Hence, for DS14 we did not estimate coefficient H^T.

Table 1. Descriptive statistics of the items (upper panel) and the scale (lower panel) for NA, SI, and DS14 scores

	NA					SI					DS14				
Item	*M*	*SD*	H_j	*SE*	*citc*	*M*	*SD*	H_j	*SE*	*citc*	*M*	*SD*	H_j	*SE*	*citc*
Si1						1.28	1.17	.57	.03	.71	1.28	1.17	.35	.03	.52
Na2	1.88	1.31	.48	.03	.56						1.88	1.31	.28	.03	.41
Si3						1.81	1.26	.45	.03	.53	1.81	1.26	.22	.03	.33
Na4	0.90	1.11	.56	.03	.68						0.90	1.11	.40	.03	.60
Na5	1.67	1.24	.50	.03	.60						1.67	1.24	.31	.03	.46
Si6						1.21	1.18	.49	.03	.62	1.21	1.18	.44	.02	.67
Na7	0.96	1.18	.59	.03	.72						0.96	1.18	.43	.02	.65
Si8						1.27	1.23	.57	.03	.73	1.27	1.23	.41	.02	.62
Na9	0.94	1.06	.51	.03	.62						0.94	1.05	.35	.03	.53
Si10						1.46	1.33	.55	.02	.69	1.46	1.33	.37	.02	.57
Si11						1.56	1.14	.49	.03	.60	1.56	1.14	.34	.03	.49
Na12	1.82	1.34	.56	.02	.67						1.82	1.34	.37	.02	.55
Na13	0.87	1.12	.62	.02	.71						0.87	1.12	.42	.02	.62
Na14						1.18	1.13	.52	.03	.64	1.18	1.13	.37	.02	.57
M			9.04	0.27				9.77	0.27				18.81	0.45	
SD			6.32	0.17				6.33	0.16				10.38	0.29	
H			.55	.02				.52	.02				.36	.02	
α			.87					.87					.88		
λ_2			.88					.87					.88		
MS			.88					.87					.88		

Note. *M*, mean; *SD*, standard deviation; H_j, item scalability coefficient; *SE*, standard error of item scalability coefficient; *citc*, corrected item–test correlation; α, Cronbach's alpha; λ_2, Guttman's lambda-2; MS, Molenaar–Sijtsma method.

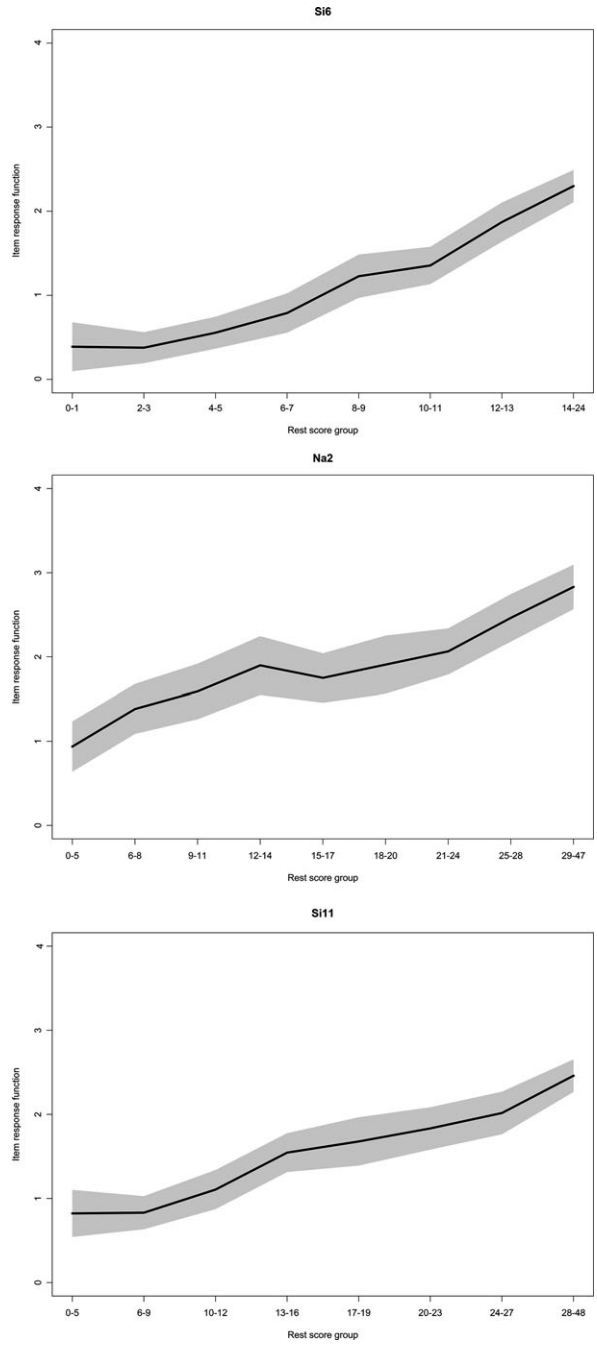

Figure 6. IRF of item 6 (Si6, SI; top) and items 2 (Na2, DS14; middle) and 11 (Si11, DS14; bottom).

Based on the scale identification stage, because the two negatively worded items 1 and 3 violated local independence and item 3 also appeared problematic in other analyses, these items should probably be revised. In addition, the locally dependent

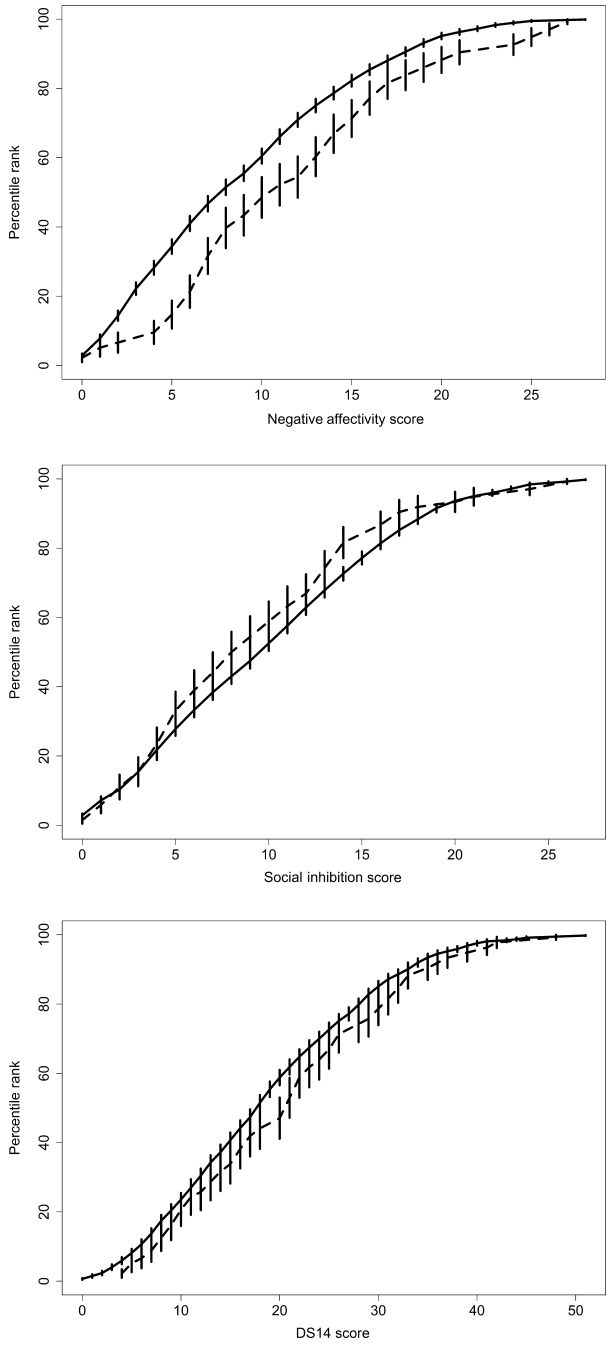

Figure 7. Percentile ranks for males (solid lines) and females (dashed lines) plus or minus one standard error (vertical lines) for NA (top), SI (middle), and DS14 (bottom).

items 8 and 14 also may be candidates for further scrutiny. We advocate using NA and SI scores both based on only seven items but leaving out item 3 when using the total score based on the longer DS14.

(iii) Scale properties

Step 8. *Reliability.* Table 1 (lower panel) shows the MS method reliability-estimate. Table 1 also provides coefficients α (Cronbach, 1951) and λ_2 (Guttman, 1945). All estimates are close to .9 and thus satisfactory. Reliability for DS14 was a little higher than for NA and SI. Without item 3, reliability grew by .01 units. The corrected item–test correlations (Table 1, upper panel, columns, 5, 15, and 20, *citc*) were satisfactory for all items.

Step 9. *Norms.* For men and women separately, Figure 7 shows rank percentiles for NA (top), SI (middle) and DS14 (bottom). For NA and DS14, the same test score results in a higher percentile rank for men than women. However, for DS14 the overlapping confidence intervals suggest that most score differences between men and women are not significant. For SI, women have consistently higher percentile scores but differences are not significant. Given the smaller sample size, women's confidence intervals are larger.

Step 10. *Group comparison.* We compared the scalability difference between men and women using the sample item ordering. For NA, item 5 showed the largest difference, being higher for men ($H_5 = .50$) than for women ($H_5 = .44$). Total scalability was equal for both groups. For SI, item 3 showed the largest difference (men: $H_3 = .45, SE = 0.03$; women: $H_3 = .19, SE = 0.11$). Total scalability was also higher for men (men: $H = .52$ ($SE = 0.02$); women: $H = .42$ ($SE = 0.07$)). For DS14, except for items 9 and 13, item scalability was higher for men. For women, $H_3 = -.01$ ($SE = 0.10$) and for men $H_3 = .22$ ($SE = 0.03$). Item 3 ("I often talk to strangers") seems to evoke a different response pattern in different gender groups.

4. Epilogue

We conclude with two take-away messages:

(1) Any scale analysis is complex, and so is MSA. In different analysis rounds, for varying sets of items and individual items the researcher has to assess the assumptions of measurement models, and provide quality indices such as scalability and reliability. Only considering scalability coefficients provides an incomplete picture and must be discouraged.

(2) The three-stage, ten-step procedure helps the researcher on his way but must be applied with judgement; that is, decisions to include or leave out items and assemble items in scales should not only be based on statistical considerations but also be evident from the item's content, preferably derived from theory or common practice.

Acknowledgement

The authors thank Johan Denollet for kindly providing the DS14 data that we used for the analysis example.

References

Andrich, D. (1989). A probabilistic IRT model for unfolding preference data. *Applied Psychological Measurement*, *13*, 193–216. doi:10.1177/014662168901300211

Bernaards, C. A., & Sijtsma, K. (2000). Influence of simple imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*, 321–364. doi:10.1207/S15327906MBR3503_03

Brusco, M. J., Köhn, H.-F., & Steinley, D. (2015). An exact method for partitioning dichotomous items within the framework of the monotone homogeneity model. *Psychometrika*, *80*, 949–967. doi:10.1007/s11336-015-9459-8

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, *27*, 412–433. doi:10.1214/12-STS396

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi:10.1007/BF02310555

Denollet, J. (2005). DS14: Standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine*, *67*, 89–97. doi:10.1097/01.psy.0000149256.81953.49

Denollet, J., Pedersen, S. S., Vrints, C. J., & Conraads, V. M. (2013). Predictive value of social inhibition and negative affectivity for cardiovascular events and mortality in patients with coronary artery disease: The Type D personality construct. *Psychosomatic Medicine*, *75*, 873–981. doi:10.1097/PSY.0000000000000001

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika*, *79*, 303–316. doi:10.1007/s11336-013-9341-5

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392. doi:10.1007/BF02294219

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282. doi:10.1007/BF02288892

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*, 337–352. doi:10.1177/014662169501900404

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331–347. doi:10.1007/BF02294555

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, *14*, 1523–1543. doi:10.1214/aos/1176350174

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, *52*, 5186–5201. doi:10.1016/j.csda.2007.11.008

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65–81. doi:10.1177/01466216000241004

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*, 211–220. doi:10.1177/01466210122032028

Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika*, *76*, 200–216. doi:10.1007/s11336-010-9199-8

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*, 578–595. doi:10.1177/0013164409355697

Meijer, R. R., & Egberink, I. J. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement*, *72*, 589–607. doi:10.1177/0013164411429344

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

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417–430. doi:10.1177/014662168200600404

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to 'The Mokken scale: A critical discussion'. *Applied Psychological Measurement*, *10*, 279–285. doi:10.1177/014662168601000306

Molenaar, I. W. (1997). Nonparametric models for polytomous items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York, NY: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. A program for Mokken scale analysis for polytomous items*. Groningen, The Netherlands: iecProGAMMA.

Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (2016). Computing standard errors and confidence intervals for norm statistics. *Psychometrika*, Advance online publication. doi:10.1007/s11336-016-9535-8

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 381–394). New York, NY: Springer.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: Wiley.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*, 183–200. doi:10.1007/BF02294774

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79–105. doi:10.1111/j.2044-8317.1996.tb01076.x

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*, 149–157. doi:10.1177/014662169201600204

Sijtsma, K., Meijer, R. R., & Van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*, 31–37. doi:10.1016/j.paid.2010.08.016

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, *52*, 79–97. doi:10.1007/BF02293957

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

Sijtsma, K., & Molenaar, I. W. (2016). Mokken models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume One: Models* (pp. 303–321). Boca Raton, FL: Chapman & Hall/CRC.

Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken scale analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement*, *36*, 516–539. doi:10.1177/0146621612451050

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika*, *55*, 293–326. doi:10.1007/BF02295289

Stout, W. F. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*, 485–518. doi:10.1007/BF02295128

Stout, W. E., Habing, B., Douglas, J., Kim, H., Roussos, L., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354. doi:10.1177/014662169602000403

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*, 72–99. doi:10.1007/s00357-013-9122-y

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement*, *74*, 809–822. doi:10.1177/0013164414529793

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (in press). Using conditional association to identify locally independent item sets. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*.

Van Abswoude, A. H. H., Vermunt, J. K., Hemker, B. T., & Van der Ark, L. A. (2004). Mokken scale analysis using hierarchical clustering procedures. *Applied Psychological Measurement*, *28*, 332–354. doi:10.1177/0146621604265510

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

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

Van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika*, *75*, 272–279. doi:10.1007/S11336-010-9147-7

Van der Ark, L. A., & Sijtsma, K. (2005). The effect of missing data imputation on Mokken scale analysis. In L. A. Van der Ark, M. A. Croon & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 147–166). Mahwah, NJ: Erlbaum.

Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204. doi:10.3102/10769986031002181

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research*, *42*, 387–414. doi:10.1080/00273170701360803

Zhang, J., & Stout, W. E. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249. doi:10.1007/BF02294536

Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, *36*, 186–212. doi:10.3102/1076998610366263

Received 29 March 2016; revised version received 29 August 2016