Goodness-of-fit methods for nonparametric IRT models

Sijtsma, K.; Straat, J.H.; van der Ark, L.A.

DOI
10.1007/978-3-319-19977-1_9

Publication date
2015

Document Version
Final published version

Published in
Quantitative Psychology Research

Chapter 9
Goodness-of-Fit Methods for Nonparametric IRT Models

Klaas Sijtsma, J. Hendrik Straat, and L. Andries van der Ark

Abstract This chapter has three sections. The first section introduces the unidimensional monotone latent variable model for data collected by means of a test or a questionnaire. The second section discusses the use of goodness-of-fit methods for statistical models, in particular, item response models such as the unidimensional monotone latent variable model. The third section discusses the use of the conditional association property for testing the goodness-of-fit of the unidimensional monotone latent variable model. It is established that conditional association is well suited for assessing the local independence assumption and a procedure is proposed for identifying locally independent sets of items. The procedure is used in a real-data analysis.

Keywords Conditional association • Goodness-of-fit methods • Local independence • Robustness of conclusions when models fail • Unidimensional monotone latent variable model

---

Paper presented at the International Meeting of the Psychometric Society 2014, Madison, Wisconsin, July 21st until July 25th, 2014.

K. Sijtsma (✉)
Department of Methodology and Statistics, TSB, Tilburg University,
Warandelaan 2, 5037 AB, PO Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: k.sijtsma@tilburguniversity.edu

J.H. Straat
Cito Arnhem, Amsterdamseweg 13, 6814 CM Arnhem, The Netherlands
e-mail: hendrik.straat@cito.nl

L.A. van der Ark
University of Amsterdam, Room D7.15, Nieuwe Achtergracht 127,
1018 WS Amsterdam, The Netherlands
e-mail: L.A.vanderArk@uva.nl

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 140, DOI 10.1007/978-3-319-19977-1_9

9.1 Introduction to the Unidimensional Monotone Latent Variable Model

We discuss the unidimensional monotone latent variable model (UMLVM), which is a nonparametric item response theory (IRT) model also known as the monotone homogeneity model (Sijtsma and Molenaar 2002). Next, we discuss the issues of assessing the goodness-of-fit (GoF) of IRT models and the UMLVM in particular to the data and problems that GoF investigation of IRT models typically encounters. Finally, we propose a new GoF procedure for the UMLVM that selects one item set or several item subsets consistent with the UMLVM's local independence assumption from an initial item set that may or may not be consistent with local independence.

Let θ denote the latent variable, and let X_j denote the random variable for the score on item j ($j = 1, \ldots, J$; J is the number of items in the test). The three assumptions on which the UMLVM is based are the following.

- Unidimensionality (UD): latent variable θ is unidimensional;
- Local independence (LI): the item scores are independent conditional on θ; that is,

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

  LI implies Weak LI, for covariances between items defined as

$$\sigma(X_j, X_k \Big| \theta) = 0,$$

  and which proves to be useful in this chapter. It may be noted that Weak LI is a weaker property than LI: LI $\Rightarrow$ Weak LI, but Weak LI $\not\Leftarrow$ LI;
- Monotonicity (M): The J IRFs are monotone nondecreasing in θ; that is, expectation $E(X_j \Big| \theta)$ is nondecreasing in θ.

The essential difference with parametric IRT models, such as the 1, 2, and 3-parameter logistic models, the (generalized) partial credit model and the graded response model, is that in nonparametric IRT models, such as the UMLVM, the IRFs are not parametrically defined by means of, for example, logistic functions, but are only subjected to order restrictions. For example, let us consider the logistic IRF of the 1-parameter logistic model (Van der Linden and Hambleton 1997a), in which δ_j denotes the item's location or difficulty parameter and 0/1 scoring for example denotes incorrect/correct scoring, so that

$$P(X_j = 1 \Big| \theta) = E(X_j \Big| \theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)}.$$

The latent variable θ and the latent item parameter δ_j can be estimated by means of maximum likelihood procedures. However, nonparametric IRT models such as the UMLVM only impose assumption M on the IRFs but do not parametrically define the IRFs, and in the absence of parametric IRFs such as the logistic, nonparametric IRT models do not enable estimating the latent variable θ and the latent item parameter δ_j (but see Mokken and Lewis 1982, for an alternative approach). However, the nonparametric UMLVM is a useful model because it does imply an ordinal scale for person measurement that is suited in most practical testing applications. Next, we discuss the properties of the ordinal scale.

For dichotomous items, the UMLVM implies stochastic ordering of latent variable θ by total score $X_+ = \sum_{j=1}^{J} X_j$ (SOL; Grayson 1988; Hemker et al. 1997). Let C and K be values of X_+, such that $0 \leq C < K \leq J$. Then for any t, SOL is defined as

$$P(\theta > t \Big| X_+ = C) \leq P(\theta > t \Big| X_+ = K).$$

SOL refers to the ordering of the conditional, cumulative distributions of θ. For the means of these distributions, SOL implies that an increasing total score X_+ produces an increasing mean latent variable θ. Hence, SOL means that X_+ orders persons on θ, and this allows making decisions about relative attribute levels.

For polytomous items, mathematically the UMLVM does not imply SOL but using an extensive simulation study, Van der Ark (2005) demonstrated that SOL holds by approximation and that person reversals with respect to θ due to the use of X_+ usually concern adjacent X_+ values. Thus, rare ordering errors do not seem to cause serious and far-reaching decision errors. In addition, SOL implies weak SOL, defined as

$$P(\theta > t \Big| X_+ < x_+) \leq P(\theta > t \Big| X_+ \geq x_+),$$

and Van der Ark and Bergsma (2010) proved that the UMLVM implies Weak SOL: SOL $\Rightarrow$ Weak SOL, but Weak SOL $\not\Rightarrow$ SOL. The dichotomization $X_+ < x_+$ and $X_+ \geq x_+$, typical of using cut scores, orders persons on θ, and allows the use of total score X_+ for assignment of individuals to the categories failure and success in educational testing, rejection and selection in job assessment, and ineligible and eligible for therapy or treatment in clinical settings.

The conclusion is that the UMLVM implies an ordinal scale on θ by means of total score X_+. An interesting note often ignored is that all the parametric models that are mathematical special cases of the UMLVM imply the use of X_+ as an ordinal estimator of θ, thus justifying the use of the much more accessible total score in all applications where this might prove convenient. That is, the 1, 2, and 3-parameter logistic models and their normal-ogive counterparts imply SOL, and the (generalized) partial credit model and the graded response model imply Weak SOL. In the 1-parameter logistic model and its polytomous-item generalization, the partial credit model, total score X_+ is a sufficient statistic for the maximum likelihood

estimation of latent variable θ. In other parametric IRT models, this relationship is absent and it is often assumed incorrectly that X_+ has no place in the application of such models. However, it has as an ordinal estimator of the θ scale, and when reasons to resort to the θ scale are absent one can use the ordinal X_+ scale instead.

9.2 Goodness-of-Fit Research for the UMLVM

A good fit of an IRT model to the data is essential for establishing the model's measurement properties for the particular application envisaged. Without a well-fitting model, the measurement specialist and the researcher cannot know whether the measurement properties, in case of the UMLVM an ordinal scale, hold for the test of interest, and the measurement practitioner cannot know whether conclusions about people based on the scale are valid. An important question is when to use the UMLVM as opposed to parametric IRT models. The answer is: When parametric IRT models fail to fit the data well. This may seem to be a modest position, but model-fit failure is the rule rather than the exception and is frequently ignored implicitly assuming that the misfitting parametric IRT model can be used in practice anyway; thus, the UMLVM may be useful in many applications to obtain an IRT model that fits better than a parametric IRT model. We first discuss GoF in general and then address the question of why researchers tend to neglect GoF investigation.

Like any model, IRT models are idealizations of reality and, consequently, they cannot describe the data structure perfectly well. Thus, a GoF investigation will always suggest at least some degree of model misfit. We distinguish three possible outcomes of a GoF investigation. First, one may find that an IRT model provides a reasonable approximation to the data and accept the model as a description of the data. Second, one may conclude that the IRT model shows serious misfit and decide that, for example, the item set should be divided into different subsets each measuring a different attribute or misfitting items should be removed from the item set hoping the IRT model to fit to the remaining item subset. The second outcome may be a reasonable approximation but in principle the result is always some degree of misfit. The third outcome is that the misfit is hopeless and nothing can be done to save the test; that is, as long as one sticks to the IRT model selected to model the data. In each case, in particular when misfit appears hopeless (i.e., option 3) but also when items are rejected because their IRFs are not logistic or have slopes deviating from the majority of the IRF slopes (i.e., option 2) may one resort to an alternative IRT model based on weaker assumptions, such as the UMLVM.

In test construction, GoF investigation appears to be somewhat neglected despite the availability of GoF methods for several IRT models (e.g., Glas and Verhelst 1995; Sijtsma and Molenaar 2002; Van der Linden and Hambleton 1997b). One can only speculate about the reasons for the more general neglect. One reason may be that GoF investigation is complex. First, GoF methods never address the whole model simultaneously but only one model assumption or a pair of model assumptions. For example, several GoF methods exist that assess the combination

of UD and LI or only LI, and other methods assess M possibly including a particular parametric shape, but methods that simultaneously assess all assumptions of a model, say, the 1-parameter logistic model or the graded response model, to our knowledge do not exist. Second, GoF methods may be global, assessing the GoF of all items with respect to one or two assumptions simultaneously, or they may be local, assessing whether pairs of items are locally independent or whether the IRF of one particular item is monotone. Third, splitting the item set in subsets or removing an item from the item set produces a smaller data set and affects the GoF results in the next analysis round, possibly causing initially fitting items to show misfit. Combining these different aspects of a GoF investigation is difficult and may easily discourage researchers; De Koning et al. (2002) and Sijtsma et al. (2011) suggest how to consistently perform a complex GoF investigation.

Another reason for GoF neglect is that several GoF methods check a particular observable consequence of a model, following the logic that negative results imply that the IRT model cannot have generated the data. While the logic is correct, it remains unknown which assumption or assumptions were violated in particular. For example, the UMLVM and all its special cases including many parametric IRT models imply positive inter-item correlations but the presence of negative correlations among several positive correlations usually does not inform the researcher which assumption or which assumptions have been violated, only that the model did not generate the data. Hence, the diagnostic value of negative inter-item correlations appears limited.

A comprehensive GoF investigation based on the combination of different methods assessing different assumptions, for all items simultaneously and for individual items and pairs of items, and possibly also considering GoF methods providing little diagnostic information, may produce additional problems. First, a comprehensive GoF procedure typically involves many decisions as the procedure moves along thus introducing results that increasingly capitalize on chance, calling for cross validating the end result. Second, a GoF investigation typically produces a plethora of results that need to be combined so as to enable the researcher to draw a conclusion about the fit of his IRT model to the data. Little research has been done with respect to the question of how to combine the detailed results into one conclusion about GoF.

In the third section of this chapter, we discuss a new GoF method that, when the data are inconsistent with the UMLVM, has two apparent problems that we try to solve: The method does not inform the researcher unequivocally which assumption—UD, LI, or M—is violated and moreover produces an avalanche of detailed results. We investigate which assumption is violated when the method indicates model misfit and we suggest a solution to the problem of multiple detailed results. Many GoF methods assess the UMLVM; for UD assessment see Mokken (1971) and Straat et al. (2013); for LI assessment see Zhang and Stout (1999) and Douglas et al. (1998); and for M assessment see Rossi et al. (2002) and Tijmstra et al. (2013). Sijtsma and Molenaar (2002) and Sijtsma and Meijer (2007) provide overviews.

9.3 Conditional Association

We studied conditional association (CA; Holland and Rosenbaum 1986), which is an observable consequence of the UMLVM, as a potential method for assessing whether the data are consistent with the model's assumption of LI. Let the vector of J item-score variables be denoted by $\mathbf{X}$, and let $\mathbf{X}$ be divided into two mutually exclusive and exhaustive item subsets $\mathbf{Y}$ and $\mathbf{Z}$, so that $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. Also, let f_1 and f_2 be nondecreasing functions and let h be any function. Holland and Rosenbaum proved that the UMLVM implies CA,

$$\sigma[f_1(\mathbf{Y}), f_2(\mathbf{Y}) \Big| h(\mathbf{Z}) = \mathbf{z}] \geq 0.$$

CA implies that in particular subgroups defined by $h(\mathbf{Z}) = \mathbf{z}$, the covariance between nondecreasing functions $f_1(\mathbf{Y})$ and $f_2(\mathbf{Y})$ is non-negative. Examples (Sijtsma 2003) of CA are:

- $\sigma(X_j, X_k) \geq 0$; all inter-item covariances/correlations are non-negative;
- $\sigma(X_j, X_k \Big| X_l = x_l) \geq 0$; within item-score subgroups, all inter-item covariances are non-negative; and
- $\sigma(X_j, X_k \Big| R_{(jk)} = r) \geq 0$, rest score $R_{(jk)} = \sum_{i \neq j,k} X_i$, R has realizations r; within rest-score subgroups, all inter-item covariances are non-negative.

These covariances are used separately in several GoF methods for the UMLVM, but here we will investigate whether they can be used for investigating LI.

CA provides the means for testing the GoF of the UMLVM to the data as follows:

- If the covariances are negative, then the UMLVM did not generate the data; and
- If the covariances are positive, then one has found support for the UMLVM (but not proof, which is impossible in sample data).

A drawback for this sort of GoF research is the many covariances generated, among them perhaps negative covariances due to serious model violations but others due to only minor violations and sampling fluctuation, thus rendering it difficult to draw straightforward conclusions about GoF. For example, assume one has 20 items and 5 ordered item scores per item; then, drawing conclusions about GoF would involve a complete and possibly confused inspection that assesses and combines the results from

- 190 covariances $\sigma(X_j, X_k)$;
- 5700 covariances $\sigma(X_j, X_k \Big| X_l = x_l)$; and
- 13,680 covariances $\sigma(X_j, X_k \Big| R_{(jk)} = r)$.

9.3.1 How to Use CA Failure to Identify UMLVM Misfit?

How are the three cases of CA related to violations of UD, LI, and M? Suppose, we need a multidimensional $\boldsymbol{\theta}$ to explain the associations between the items; then, conditioning on one latent variable θ violates LI and also weak LI, and may also cause non-monotone IRFs reflected by negative (conditional) inter-item covariances. We distinguish two violations of weak LI: positive local dependence (PLD), $\sigma(X_j, X_k \big| \theta) > 0$, and negative local dependence (NLD), $\sigma(X_j, X_k \big| \theta) < 0$ (Rosenbaum 1988). What one needs to know is whether, for example, $\sigma(X_j, X_k \big| X_l = x_l) < 0$ is due to items j and k being PLD or NLD, or whether the negative covariance is due to non-monotonicity of the items' IRFs. We used mathematical results provided by Holland and Rosenbaum (1986) and Rosenbaum (1988) and a computational study to find an answer to questions like these when the three cases of CA provide negative values in sample data (Straat et al. 2014).

The mathematical results showed that even when the UMLVM fails to hold, particular observable covariances are positive; hence, such covariances are useless to assess UMLVM fit. For the other observable covariances, a computational study was used to mimic PLD or NLD for particular item pairs or IRF non-monotonicity for particular items, and to estimate the proportion by which a particular conditional covariance for the corresponding items was negative. Reversely, we argued that the higher the proportion, the higher the power of a particular covariance to identify item pairs that were PLD or NLD, or items that had non-monotone IRFs.

The results of the computational study were the following. Conditional covariances had insufficient power to detect IRF non-monotonicity; hence, they are not suitable for this purpose. The next three types of covariances are suitable for identifying PLD and NLD; that is, they are suited to identify violations of LI. Let a and b be two items from item subset $\mathbf{Y}$, and let c be an item from $\mathbf{Z}$; j indexes any item from the union of both subsets, $\mathbf{X}$. PLD(a, b) means that items a and b are PLD, and NLD(a, b) that both items are NLD. Further, s denotes sample covariance. The next three results appear consistently across different choices of item parameters:

PLD: 1. If PLD(a, c) is investigated, then $s(X_a, X_j \big| X_c = x_c) < 0$ indentifies PLD;

2. If PLD(a, j) is investigated, then $s(X_a, X_b \big| R = r) < 0$ identifies PLD;
Note: item b may replace item a; formally, nothing changes.

NLD: 3. If NLD(a, b) is investigated, then $s(X_a, X_b \big| R = r) < 0$ identifies NLD.

These results show that only a limited number of observable conditional covariances have enough power to be useful in GoF research. The other covariances often have positive values if LI is violated, that is, when the UMLVM fails to fit the data.

9.3.2 Usefulness of CA Failure for Identifying Locally Dependent Items

Straat et al. (2014) proposed a methodology that uses the three specific conditional covariances above for identifying locally dependent items from a larger set, and which therefore are candidates for removal from the test represented by vector $\mathbf{X}$. For each of the three covariance results discussed in the previous section, the authors defined unique indices denoted $W^{(1)}$, $W^{(2)}$, and $W^{(3)}$, respectively, that quantify the degree to which the item is suspected to belong to locally dependent pairs. For a set of J items, each of the $J(J-1)$ indices $W^{(1)}$ is a weighted count of negative covariances defined in Result 1 in the previous section [i.e., $s(X_a, X_b \big| X_c = x_c) < 0$, $j = 1, \ldots, J; j \neq a, b$]; each of the J indices $W^{(2)}$ is a weighted count of negative covariances defined in Result 2 in the previous section [i.e., $s(X_a, X_j \big| R = r) < 0$, $j = 1, \ldots, J; j \neq a; r = 1, \ldots, R$]; and each of the $J(J-1)$ indices $W^{(3)}$ is a weighed count of negative covariances defined in Result 3 in the previous section [i.e., $s(X_a, X_b \big| R = r) < 0; r = 1, \ldots, R$]. Each index is the sum of probabilities that a sample conditional covariance s is negative under the null hypothesis that the population covariance σ is non-negative. After a Fisher Z-transformation, sample covariances are assumed to be normally distributed, and the sum of the areas under the normal curve that correspond to the negative scale region on the abscissa defines the value of a W index. A larger negative sum, that is, a larger positive W value, suggests a stronger case for local dependence and thus removing the item from $\mathbf{X}$.

Tukey's fences were used to determine whether a W index has a negative value high enough to remove the item from $\mathbf{X}$. The authors adjusted a procedure Ligtvoet et al. (2010) used in another context for item selection to their purpose, which was to identify and then remove locally dependent items from $\mathbf{X}$. Straat et al. (2014) called the adjusted procedure the CA procedure. In a simulation study, the authors found that CA procedure had a specificity—the proportion of correctly identified LI items or item pairs that were kept in $\mathbf{X}$—equal to 89.5 %. The CA procedure's sensitivity was defined for single items and pairs of items and assessed for different versions of local dependence, and varied from approximately 42–90 %.

9.3.3 Real-Data Example: The Adjective Checklist

We analyzed data from the Adjective Checklist (Gough and Heilbrun 1980), which are available in the R package "`mokken`" (Van der Ark 2007, 2012). The data consisted of the scores of 433 students on 218 items from a Dutch version of the Adjective Checklist. Each item is an adjective having five ordered answer categories (0 = completely disagree, 1 = disagree, 2 = neither agree nor disagree, 3 = agree, 4 = completely agree). The respondents were instructed to consider whether an adjective described their personality, and mark the answer category that fitted best

Table 9.1 Item means, item-scalability coefficients, and total-scalability coefficient (standard errors between parenthesis) for two ACL scales

Achievement				Nurturance			
Item	Mean	H_j	(s.e.)	Item	Mean	H_j	(s.e.)
Active	2.471	0.408	(0.030)	Kind	2.771	0.266	(0.036)
Alert	2.395	0.337	(0.036)	Aloof*	2.312	0.190	(0.031)
Ambitious	2.448	0.410	(0.030)	Helpful	2.624	0.264	(0.034)
Thorough	2.259	0.322	(0.036)	Intolerant*	2.998	0.247	(0.034)
Energetic	2.460	0.423	(0.032)	Sympathetic	2.778	0.265	(0.036)
Unambitious*	2.734	0.367	(0.033)	Snobbish*	3.044	0.196	(0.032)
Quitting*	2.811	0.321	(0.036)	Affectionate	2.972	0.207	(0.037)
Determined	2.499	0.384	(0.036)	Hostile*	3.307	0.337	(0.027)
Industrious	2.067	0.372	(0.034)	Friendly	2.806	0.317	(0.032)
Persevering	2.298	0.433	(0.032)	Distrustful*	2.700	0.221	(0.031)
Total scale		0.378	(0.026)	Total scale		0.247	(0.024)

Note: An asterisk indicates adjectives that are negative with respect to the attribute. Tabulated results are based on recoded item scores

to this description. The 218 items constitute 22 scales. For illustration purposes we selected two 10-item scales: Achievement, having item-scalability H_j-values (Sijtsma and Molenaar 2002, chap. 4) greater than 0.3 for all items, and Nurturance, having rather low item-scalability coefficients (Table 9.1). We used the R package "`mokken` (Van der Ark 2007, 2012) to compute the scalability coefficients of the items, and we used the R package "`CAprocedure`" (available from the third author upon request) for the CA procedure. The R-code is provided in the Appendix.

The UMLVM implies that item-pair scalability coefficients (Sijtsma and Molenaar 2002, chap. 4) and item-scalability coefficients are non-negative. For both scales, we found that all item-pair scalability coefficients (not tabulated) and all item-scalability coefficients indeed were positive, lending support to the fit of the UMLVM.

For Achievement, the CA procedure flagged only item pair (Ambitious, Unambitious) for possible PLD (Index $W^{(1)}$). The 10 items produced 90 indices $W^{(1)}$. Based on these 90 indices, Tukey's upper fence was equal to 11.433. For item pair (Ambitious, Unambitious), index $W^{(1)}$ equaled 12.194. For all other item pairs, the $W^{(1)}$ values did not exceed Tukey's upper fence. None of 10 indices $W^{(2)}$ and none of the 45 indices $W^{(3)}$ exceeded the corresponding Tukey's upper fences. The result can be explained by noticing that the reversed scores of Unambitious were used to compute the results, and reversal of the scores renders the items similarly worded, so that a flag for PLD seems reasonable. Because Unambitious had the lower item-scalability value, this item is a candidate for removal.

For Nurturance, the CA procedure flagged item-pair (Hostile*, Distrustful*) for PLD ($W^{(1)} = 18.189$, Tukey's upper fence $= 15.777$), and the items Aloof* ($W^{(2)} = 71.047$, Tukey's upper fence $= 70.741$) and Snobbish* ($W^{(2)} = 74.924$,

Tukey's upper fence = 70.741) for being in a PLD item pair. Aloof* had the lowest item-scalability coefficient and was removed first, followed by Distrustful*, and Snobbish*. After these three items were removed, Intolerant* ($W^{(2)} = 33.027$, Tukey's upper fence = 30.650) was flagged for being in a PLD item pair, and was also removed, leaving six items in the scale. Except for Hostile*, all negatively worded items were removed. An explanation for the large number of flagged items is that the negatively worded items formed a separate dimension.

9.4 Discussion

Conditional association offers possibilities for LI assessment in goodness-of-fit studies of the UMLVM. Given the variable results for CA procedure's sensitivity, it seems worthwhile to study how the procedure can be improved so as to increase its sensitivity. A comparison with alternative procedures assessing LI is useful and should be conducted. In a broader context, we noticed that GoF methods for any statistical model hardly ever address the complete model but target particular assumptions or sets of intimately related assumptions. For nonparametric IRT models the picture is no different but fortunately a large array of GoF methods assessing nonparametric IRT assumptions is available. The assessment of LI seems to be the least well developed. This chapter discussed a contribution to LI assessment. From the researcher's viewpoint a sound methodology that combines the best GoF methods so as to obtain a comprehensive picture of a model's fit to the data with respect to UD, LI and M is another topic we intend to pursue.

A.1 Appendix

R code we used for the real-data example.

```
R> library("CAprocedure")
R> library("mokken")
R> data(acl)
R> # Achievement
R> Ach <- acl[, 11 : 20]
R> coefH(Ach)
R> apply(Ach, 2, mean)
R> CAP(Ach, TRUE)
R> # Nurturance
R> Nur <- acl[, 61 : 70] #
R> coefH(Nur)
R> apply(Nur, 2, mean)
R> CAP(Nur, TRUE)
```

References

De Koning, E., Sijtsma, K., & Hamers, J. H. M. (2002). Comparing four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement, 26*, 302–320.

Douglas, J., Kim, H., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129–151.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Their foundations, recent developments and applications* (pp. 69–96). New York, NY: Springer.

Gough, H. G., & Heilbrun, A. B. (1980). *The Adjective Check List, manual 1980 edition*. Palo Alto, CA: Consulting Psychologists Press.

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

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.

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

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.

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

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

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349–359.

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics, 27*, 291–317.

Sijtsma, K. (2003). Developments in practical nonparametric IRT scale analysis. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 183–190). Tokyo, Japan: Springer.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics*. Vol. 26, Psychometrics (pp. 719–746). Amsterdam, The Netherlands: Elsevier.

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.

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

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.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2014). *Using conditional association to identify locally independent item sets* (Manuscript submitted for publication).

Tijmstra, J., Hessen, D. J., Van der Heijden, P. G. M., & Sijtsma, K. (2013). Testing manifest monotonicity using order-constrained statistical inference. *Psychometrika, 78*, 83–97.

Van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*, 283–304.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–10.

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1–27.

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.

Van der Linden, W. J., & Hambleton, R. K. (1997a). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). New York, NY: Springer.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997b). *Handbook of modern item response theory*. New York, NY: Springer.

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