Using scalability coefficients and conditional association to assess monotone homogeneity

Straat, J.H.

Using Scalability Coefficients and Conditional Association to Assess Monotone Homogeneity

Printed by Ridderprint BV, Ridderkerk
ISBN/EAN: 978-90-5335-598-5

Using Scalability Coefficients and Conditional Association to Assess Monotone Homogeneity

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op

vrijdag 23 november 2012 om 14.15 uur

door

Johannes Hendrikus Straat

geboren op 24 juli 1983 te Eindhoven

Promotiecommissie:

Promotor:	prof. dr. K. Sijtsma
Copromotor:	dr. L. A. van der Ark
Overige leden:	prof. dr. J. K. L. Denollet
	prof. dr. W. J. Heiser
	prof. dr. R. R. Meijer
	dr. A. A. Béguin
	dr. W. H. M. Emons

Contents

Chapter 1

Introduction

In the social and behavioral sciences, researchers commonly use tests and questionnaires to measure attributes such as cognitive abilities including aspects of intelligence, personality traits, and attitudes. Measurement of these attributes is liable to more random measurement error than measurement of, for example, temperature or distance and also to systematic but undesirable influences, such as social desirability, tiredness, and cheating. Consequently, a single item does not provide a reliable and valid measurement of the attribute and researchers have to construct multiple items to control for random measurement error and to cover different aspects of the attribute well. The responses to the items contain information about the attribute that a researcher intends to measure. If the items have been constructed adequately, test takers with a higher attribute score (e.g., being more verbally intelligent, more extravert, or having a more positive attitude towards euthanasia) are expected to score higher on an item than test takers with lower attribute scores.

Researchers are predominantly interested in the positions of test takers on the scale for the attribute of interest rather than their scores on single items. Psychometric measurement models summarize a pattern of item scores into a score on a latent variable that represents the attribute. In validity research, researchers have to evaluate whether the latent variable is a valid representation of the attribute of interest and whether the latent variable covers all relevant aspects of the attribute. Measurement models restrict the relation of the item scores and the latent variable. A general class of measurement models is item response theory (IRT; Embretson & Reise, 2000; Van der Linden & Hambleton, 1997). The evaluation of the nonparametric IRT model known as the monotone homogeneity model (Mokken, 1971; Molenaar, 1997) in real data is the central theme of this study. The monotone homogeneity model is also known as the unidimensional monotone latent variable model (Holland & Rosenbaum, 1986).

Three assumptions define the monotone homogeneity model: unidimensionality, local independence, and monotonicity. The unidimensionality assumption posits that the items in the test measure only one latent variable. This assumption reflects the ideal that items should measure one attribute so as to simplify the test performance's interpretation. The local independence assumption posits that the unidimensional latent variable is the only source of association between the items, so that the multivariate, distribution of the item scores conditional on the latent variable equals the product of the univariate, conditional distributions. The local independence assumption implies that for test takers with the same score on the latent variable, the item scores are independent, hence local independence. In combination with unidimensionality, local independence ascertains that the test measures one latent variable and nothing else. The monotonicity assumption encompasses the intuitively appealing idea that the expected item score is a monotone nondecreasing function of the latent variable. This means that as the latent-variable value increases, for each item the expected score remains the same or increases.

Many well-known parametric IRT models assume unidimensionality and local independence and assume a parametric function such as the logistic to describe the relation between the expected item score and the latent variable. These parametric IRT models are special cases of the nonparametric monotone homogeneity model that only restricts the relation between the expected item score and the latent variable to be nondecreasing. Examples of parametric IRT models for dichotomous items are the Rasch model (Rasch, 1960), the 2-parameter logistic model and the 3-parameter logistic model (Birnbaum, 1968), and examples of parametric IRT models for polytomous items are the partial credit model (Masters, 1982), and the graded response model (Samejima, 1969). The parametric IRT models reject items that have monotone relations with the latent variable that are not logistic. Because the monotone homogeneity model is less stringent with respect to the relation between the item scores and the latent variable, if different models could select items from a large pool then the monotone homogeneity model would select more items than parametric IRT models that are special cases of the former model. Including many items in a scale may be considered to be a desirable property of a measurement model. Moreover, because the unidimensional parametric IRT models are special cases of the monotone homogeneity model, assessment of the fit of the monotone homogeneity model to the data also provides information about a parametric model's data fit.

An important question is how the latent variable can summarize a person's pattern of item scores. The monotone homogeneity model justifies the use of the easily interpretable total score, which is the unweighed sum of the item scores, as an ordinal estimator of the latent variable. Thus, test takers with a higher total score on average have a higher score on the latent variable than test takers with a lower total score. Grayson (1988; also, Huynh, 1994) showed that the ordering of the latent variable by the total score holds for dichotomous items, but Hemker, Sijtsma, Molenaar, and Junker (1997) proved that for polytomous items the total score strictly is not an ordinal estimator of the latent variable. However, Van der Ark (2005) demonstrated that violations of ordinal measurement of the latent variable are rare. Moreover, Van der Ark and Bergsma (2010) proved that for polytomous items a weaker form of the ordering of the latent variable by the total score holds. Hence, the monotone homogeneity model suffices as a measurement model when the measurement purpose requires ordinal measurement, for example, to identify the most capable applicants in personnel selection (ranking of total scores) or patients who need a particular treatment more than others (dichotomization of the total-score scale).

This thesis discusses methods that use observable consequences of the monotone homogeneity model to assess the fit of the model to the data and the measurement quality of items. Observable consequences provide necessary but not sufficient conditions for the measurement model. Hence, observable consequences are particularly useful to investigate whether one or more items are inconsistent with the monotone homogeneity model. The property of conditional association (Holland & Rosenbaum, 1986; Rosenbaum, 1984) defines a large set of observable consequences of the monotone homogeneity model. Let the set of all items under consideration be partitioned into two disjoint sets of items. One set contains the items for which covariances are computed and the other set divides the total group of test takers in one or more subgroups. Conditional association implies that all covariances between any two nondecreasing functions of the items in the first set are nonnegative for any subgroup based on any function of the items in the second set. For example, the covariance between two item scores from the first set conditional on the total score on the items in the second set must be nonnegative. Because the number of partitionings of all items into two sets, the number of nondecreasing functions of the first item set, and the number of functions of the second item set are large even for a small number of items, conditional association can be specialized by means of a large number of special cases

Mokken scale analysis (Mokken, 1971, chap. 5; Sijtsma & Molenaar, 2002, chap. 5) is a nonparametric IRT method that uses special cases of conditional association to investigate the fit of the monotone homogeneity model. Mokken scale analysis evaluates two normed covariances that have values between 0 and 1 given that the monotone homogeneity model is the correct model; the normed covariance between two items, known as the H_{ij} coefficient, and the normed covariance between item j and the total score on the other items excluding item j, known as the H_j coefficient. Then, a scale (Mokken, 1971, p. 184) is defined by two criteria: (1) for all item pairs (i, j), $H_{ij} > 0$ (formally, Mokken, 1971, p. 184, used $\rho_{ij} > 0$, where ρ is the product-moment correlation), and (2) for all items j, $H_j \geq c$, where c is a user-specified lower bound (by default $c = .3$). Strictly, the monotone homogeneity model implies that the H_j values are nonnegative; that is, $0 \leq H_j \leq 1$. Hence, a positive lower bound c is not necessary for a set of items to be consistent with the monotone homogeneity model. However, it is desirable that the items have sufficient discrimination power (Van der Ark, Croon, & Sijtsma, 2008); that is, the items should distinguish test takers scoring relatively low on the latent variable and test takers scoring relatively high on the latent variable. Because an H_j value reflects the relation between the item score and the total score ordinally representing the latent variable (Van Abswoude, Van der Ark, & Sijtsma, 2004), a higher lower bound c expresses the minimally required measurement quality of the items. In addition to the H_{ij} and H_j coefficients, the total-scalability coefficient H expresses the accuracy by which the total score orders test takers on the latent variable. Mokken (1971, p. 185) provided practical rules of thumb for interpreting H values: $H < .3$ means that the set of items is unscalable; $.3 \leq H < .4$ means the scale is weak; $.4 \leq H < .5$ means the scale is medium; and $H \geq .5$ means the scale is strong.

An automated item selection procedure (Mokken, 1971, pp. 190-192) is available to partition a set of items into one or more Mokken scales. The automated item selection procedure is a bottom-up algorithm; that is, it starts with two items and adds items one by one as long as the criteria for a scale are satisfied. The procedure first selects the two items from the set that have the highest, significantly positive H_{ij} value, and in each subsequent item selection step the procedure adds the item for which $H_j \geq c$ and which has the highest H value with respect to the already selected items. The procedure stops the selection of the first scale when all items are selected or when the unselected items do not satisfy the Mokken scale criteria with respect to the selected items. Then, if possible, from the unselected items the procedure selects a second scale, then a third scale, and so on. The automated selection of items finishes when fewer than two items remain unselected or when the unselected items do not satisfy the Mokken scale criteria (Mokken, 1971, pp. 190-192; Sijtsma & Molenaar, 2002, pp. 71-72). Alternatively, if researchers have prior beliefs about which items belong in the same scale, they can investigate their beliefs by computing the H_{ij}, H_j, and H coefficients for each cluster of items to determine whether the items indeed form a scale based on the criteria that $H_{ij} > 0$ and $H_j \geq c$, and use H to interpret the strength of the scale (Mokken, 1971, pp. 189-190).

In this thesis, we discuss methods that use observable consequences of the monotone homogeneity model to evaluate the fit of the monotone homogeneity model to the data collected by means of a test or a questionnaire in a sample drawn from a particular population. We proposed an alternative formalization of the automated item selection procedure for Mokken scale analysis, investigated a new procedure using conditional covariances for the identification of locally independent item sets, determined the minimally required sample size for item selection in Mokken scale analysis, and applied the nonparametric IRT methods to psychological data from two different questionnaires frequently used in the context of clinical, health, and medical psychology.

1.1 Outline of the thesis

In Chapter 2, we proposed a genetic algorithm as an alternative for the bottom-up item selection method used in Mokken scale analysis. Mokken's automated item selection procedure has two problems: due to its bottom-up formulation, the procedure sometimes selects items that are inconsistent with the definition of a Mokken scale, and the bottom-up selection procedure may result in a local maximum with respect to Mokken's (1971) objective of partitioning items into one or more scales. In this study, we compared the performance of Mokken's bottom-up procedure and the genetic algorithm with respect to Mokken's scaling objective, and applied both versions of Mokken's automated item selection method to the communality items of the Adjective Checklist (Gough & Heilbrun, 1980).

In Chapter 3, we applied Mokken scale analysis to the Type-D Scale 14 (Denollet, 2000, 2005), a psychological questionnaire measuring the personality traits of negative affectivity and social inhibition. Previous studies obtained three different factor models describing the internal structure of the Type-D Scale 14. We used Mokken's automated item selection procedure, its genetic-algorithm version, exploratory factor analysis, and confirmatory factor analysis to investigate which of the three factor models best described the internal structure of the Type-D Scale 14.

In Chapter 4, we evaluated the dimensionality of the Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983). Recent studies criticized the Hospital Anxiety and Depression Scale because the dimensionality results seemed to depend heavily on the statistical method used and the population investigated. We showed that Mokken scale analysis can explain why the statistical methods obtain different dimensionality results and used Mokken scale analysis to identify items that are inconsistent with the monotone homogeneity model.

In Chapter 5, we investigated minimum sample-size requirements for the use of the bottom-up item selection procedure and its genetic-algorithm version in Mokken scale analysis. We determined which factors affect the minimally required sample size for accurate automated item selection in Mokken scale analysis and whether the minimally required sample size differed for the two versions of the automated item selection procedure.

In Chapter 6, we used three special cases of conditional association to assess the fit of the monotone homogeneity model to test and questionnaire data. Thus IRT theorists largely ignored the potential of conditional association for model-fit assessment. In this study, we combined three special cases of conditional association into a procedure for the identification of locally independent item sets, compared the new procedure with automated item selection in Mokken scale analysis and DETECT with respect to their potential to identify violations of local independence and monotonicity, and applied the new procedure to the Type D Scale 14.

In Chapter 7, we discussed future research that the results obtained in this PhD thesis suggested, and the ideas that came up during the research but for which time needed to put them into action ran out.

Chapter 2

Comparing Optimization Algorithms for Item Selection in Mokken Scale Analysis*

Abstract

Mokken scale analysis uses an automated bottom-up stepwise item selection procedure that suffers from two problems. First, when selected during the procedure items satisfy the scaling conditions but they may fail to do so after the scale has been completed. Second, the procedure is approximate and thus may not produce the optimal item partitioning. This study investigates a variation on Mokken's item selection procedure, which alleviates the first problem, and proposes a genetic algorithm, which alleviates both problems. The genetic algorithm is an approximation to checking all possible partitionings. A simulation study shows that the genetic algorithm leads to better scaling results than the other two procedures.

*This chapter has been accepted for publication as: Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (in press). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*.

2.1 Introduction

Tests and questionnaires — tests, for short — are used as measurement instruments in psychological, educational, sociological, marketing, and medical and health research. The aim is to measure the respondents' level on a scale for the attribute of interest, such as introversion (psychology), arithmetic ability (education), religiosity (sociology), service-quality demands (marketing), and health-related quality of life (medicine and health). To constitute a scale, the items in the test must meet the requirements of a measurement model (Section 2.2). Most measurement models assume a unidimensional scale, which facilitates the interpretation of the measurements.

Measurements may be used for making decisions about individual respondents, for example, in job selection and clinical assessment, for ordering or classifying respondents, or for comparing group means or correlating the scale scores with other interesting variables. The degree to which these measurement applications are successful is determined by the quality of the items and the number of items, denoted J. Tests that are used for making decisions about individuals require a large number of items to have enough measurement precision — that is, a relatively small standard error of measurement for the true score or a small standard error for the latent variable — and tests that are used in research that only addresses group characteristics may contain fewer items (Mellenbergh, 1996).

Item quality is often related to the degree to which an item can precisely distinguish respondents with low measurement values from respondents with high measurement values. Items that distinguish more sharply are said to have higher discrimination. The degree to which the whole test rather than the individual items distinguishes respondents depends on the test length; given fixed item discrimination, the longer the test, the more accurately the scale distinguishes respondents. The total score is the sum of the scores on the J individual items in the test, and it is often used for measuring the respondents. If item discrimination is high, then fewer items are needed to obtain a precise total score but Emons, Sijtsma, and Meijer (2007) showed that when fewer well-discriminating items are used one still needs a relatively large number of items to make precise individual decisions.

For the development of a new test, researchers usually start by creating a large pool of items that they believe contains enough items that are good indicators of the attribute. In general, several of the items may cover the intended attribute well but other items may also cover other attributes or different aspects of the same attribute, and a dimensionality analysis of the test data is done to remove the deviating items or divide the whole item pool in different clusters. Exploratory dimensionality methods are factor analysis and principal component analysis for continuous item scores and cluster techniques tailored to the discreteness of the item scores. Typically, the clustered items range from weak to strong discrimination. The purpose of Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002), which is central in this study, is to select as many sufficiently-discriminating items as possible in each cluster. The researcher defines what (s)he considers “sufficient” discrimination. Items that have sufficient discrimination relative to a total score that estimates a latent variable or a conglomerate of latent variables measure much in common and tend to be unidimensional. Deviations from unidimensionality expressed by local dependence within a cluster are rare (Straat, Van der Ark, & Sijtsma, 2012a).

Mokken scale analysis is used in almost all measurement areas in which researchers construct stand-alone tests and questionnaires for ordinal measurement. An exception is large-scale educational testing in which equating of different scales is paramount. Researchers prefer using parametric item response theory (IRT) models that allow metric scales for this purpose.

Mokken scale analysis includes a sequential clustering algorithm (Hemker, Sijtsma, & Molenaar, 1995; Mokken, 1971, pp. 191-193), which is known as the *automated item selection procedure* (AISP; Sijtsma & Molenaar, 2002, chap. 5). AISP aims at selecting from a given pool of items the largest subset of items that measure the same attribute and satisfy particular scaling criteria (Mokken, 1971, pp. 189-190). Such an item subset is a *Mokken scale*. The idea is that a test should contain as many sufficient-quality items as possible. Items left unselected may measure a different attribute, and AISP next tries finding the largest second Mokken scale in the set of remaining items, then a largest third Mokken scale, et cetera. Thus, AISP partitions a given set of items into mutually exclusive, unidimensional clusters that contain sufficiently-discriminating items.

Interestingly, AISP does not precisely formalize Mokken's goal of selecting as many sufficiently-discriminating items as possible in each cluster, so that sometimes item clusters have a different composition (Mokken, 1971, p. 193; Sijtsma & Molenaar, 2002, chap. 5). We propose a procedure that explicitly formalizes Mokken's goal, and thus is expected to produce better results than AISP. In addition, AISP selects items one by one in consecutive steps, so that an item that is selected in the beginning of the procedure may no longer satisfy the formal criteria for inclusion later on when other items also have been selected (Mokken, 1971, p. 193; Sijtsma & Molenaar, 2002, chap. 5). Mokken suggested that the researcher should exclude such misfitting items afterwards. The newly proposed procedure does not have the problem of selecting items that show misfit in hindsight. Moreover, we propose a version of AISP that guarantees that the end result does not contain misfitting items.

Other dimensionality methods also partition a given set of items into unidimensional clusters but use different definitions of dimensionality, different algorithms for finding the dimensionality, and different criteria for deciding on the final solution. Moreover, they ignore item discrimination. Examples are DETECT (Zhang, 2007; Zhang & Stout, 1999a, 1999b), which finds subsets of items that are locally independent within subsets but locally dependent between subsets, and HCA/CCPROX (Roussos, Stout, & Marden, 1998), which is a hierarchical cluster analysis method that uses a proximity measure based on conditional covariances for finding a limited number of locally optimal item clusters approximating local independence within clusters and local dependence between clusters. DIMTEST (Nandakumar & Stout, 1993) can be used to test the unidimensionality of the separate dimensions identified by DETECT or HCA/CCPROX. Van Abswoude, Van der Ark, & Sijtsma (2004), Balàsz, Hidegkuti, & De Boeck (2006), and Van Abswoude, Vermunt, & Hemker (2007) studied these and other item selection methods. Hattie (1985) and Tate (2003) discuss multiple methods for investigating test-data dimensionality for many different IRT models.

This paper is organized as follows. In Section 2.2, we discuss nonparametric IRT and AISP. In Section 2.3, we introduce an objective function for item selection methods which formalizes the ideas of Mokken, and use this objective function to define alternatives for AISP, which are a modified AISP and a genetic algorithm

(Michalewicz, 1996). In Section 2.4, we use a simulation study to compare AISP, the modified AISP, and the genetic algorithm. In Section 2.5, we apply the three item selection methods to real test data.

2.2 Mokken's Nonparametric Item Response Theory

2.2.1 Monotone Homogeneity Model

Mokken (1971, chap. 4) proposed the monotone homogeneity model (MHM) for dichotomously scored items. The MHM is defined by the assumptions of a unidimensional latent variable denoted θ, local independence of the J item score variables X_j $(j = 1, \ldots, J)$ given θ, and monotonicity of the expected item score as a function of θ (Mokken & Lewis, 1982; Sijtsma and Molenaar, 2002, pp. 18-20). This function is the item response function. Let the sum score X_+ be defined as $X_+ = \sum_{j=1}^{J} X_j$. Grayson (1988) showed that the MHM implies that X_+ stochastically orders θ. Thus, sum score X_+ can be used for ordering persons on θ. This result justifies the use of the sum score X_+ (Sijtsma & Molenaar, 2002, p. 22). For polytomous items, Hemker, Sijtsma, Molenaar, & Junker (1997) proved that, theoretically, X_+ sometimes fails to stochastically order θ, but Van der Ark (2005) found that violations of stochastic ordering are rare if the number of items exceeds 5, and Van der Ark and Bergsma (2010) proved that for polytomous items a weaker form of stochastic ordering holds.

Fit of the MHM to the data can be investigated in several ways (Sijtsma & Meijer, 2007). Mokken (1971, pp. 182-184) proposed to use scalability coefficients for selecting items and assessing the quality of the scale. These coefficients are discussed next.

Scalability Coefficients

Let the covariance between two items X_i and X_j be denoted by $Cov(X_i, X_j)$, and the maximum covariance given marginal item-score distributions by

$Cov_{max}(X_i, X_j)$. The scalability coefficient H_{ij} for item pairs is defined as

$$H_{ij} = \frac{Cov(X_i, X_j)}{Cov_{max}(X_i, X_j)}.$$

Coefficient H_{ij} is the normed covariance between items i and j, which has the desirable property that its maximum equals 1 irrespective of the item-score distributions.

For each item j, we define a rest score $R_{(j)} = X_+ - X_j$. Then, coefficient H_j for individual items is defined as

$$H_j = \frac{Cov(X_j, R_{(j)})}{Cov_{max}(X_j, R_{(j)})}.$$

Similar to sum score X_+, rest score $R_{(j)}$ is an ordinal estimator of θ (Junker, 1991), and one may argue that H_j reflects the association of item j with latent variable θ: the higher H_j, the stronger the association. Hence, coefficient H_j may be interpreted as index of item discrimination in a group characterized by a particular distribution of the sum score X_+ or the rest score $R_{(j)}$ (Mokken, Lewis, & Sijtsma, 1986).

Coefficient H is a weighted average of the J coefficients H_j (Mokken & Lewis, 1982), and is defined as

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

Coefficient H expresses the accuracy by which sum score X_+ orders persons on θ. Mokken (1971, pp. 148-153) proved that, given the MHM, $0 \leq H_{ij} \leq 1$, $0 \leq H_j \leq 1$, and $0 \leq H \leq 1$, and he (ibid, p. 185) provided practical rules of thumb for interpreting H values: $H < 0.3$ means that the set of items is unscalable; $0.3 \leq H < 0.4$ means a weak scale; $0.4 \leq H < 0.5$ a medium scale; and $0.5 \leq H \leq 1$ a strong scale.

Definition of a Mokken Scale

Mokken (1971, p. 184; Sijtsma & Molenaar, 2002, p. 68) defined a scale as a set of items satisfying two criteria: (1) for all item pairs, the product-moment correlation is positive; that is, for all (i, j) pairs, $\rho_{ij} > 0$, and (2) for a user-specified positive value of c, for all items j, scalability coefficient $H_j \geq c$. By default, $c = 0.3$, but researchers are free to choose different values for c so as to express the item quality they prefer. Items satisfying the two criteria by definition constitute a "Mokken scale". The MHM implies Criterion 1 but, given the bounds for coefficient H_j, the MHM implies Criterion 2 only if $c = 0$ (Sijtsma & Molenaar, 2002, pp. 58-59). Thus, any positive H_j is consistent with the model, and the reason to require a higher positive c value is that under the MHM such values imply a higher discrimination power (Van der Ark, Croon, & Sijtsma, 2008). Highly discriminating items are desirable in a test because they contribute to a reliable ordering of persons on latent variable θ by means of X_+. Finally, if $H_j \geq c$ for all j, then $H \geq c$ (Sijtsma & Molenaar, 2002, p. 58, Eq. 4.9).

Automated Item Selection Procedure

Based on the definition of a scale, AISP partitions J items into one or more Mokken scales, and possibly one or more items that are left unscalable. The first step of AISP is to select from all item pairs (i, j) the pair (i.e., the start set) with the greatest H_{ij} value that is significantly greater than 0 and exceeds lower bound c. Suppose, at a given step in AISP, J_s items ($J_s \geq 2$) have been selected. In the next step, from the unselected items the $(J_s + 1)$st item is selected if it: (1) correlates positively with each of the J_s selected items (Criterion 1); (2) has an H_{J_s+1} coefficient with respect to the J_s selected items that is significantly greater than 0 and also exceeds lower bound c (Criterion 2); and (3) produces the greatest H value with the J_s selected items, given all candidate items for selection. An item satisfying these criteria is selected, and this step is repeated for the $J - J_s - 1$ unselected items. AISP stops when there are no items left that satisfy the three criteria. If at least two items remain after the formation of the first scale, AISP tries to construct a second scale from these items, then a third scale, and so on. AISP terminates when no more than one item is left, or when the items left do not satisfy the scaling criteria.

AISP always produces a partitioning. However, because it is a bottom-up algorithm, which selects an item only once without the possibility of revoking the assignment later on, AISP does not consider all possible partitionings. In practice, AISP may find a first scale that is smaller than one or more subsequent scales; for example, see Sijtsma and Molenaar (2002, p. 84). In this study, we arbitrarily considered the longest scale found by AISP as the first scale, the second longest scale as the second scale, and so on. Mokken's (1971, p. 190) objective was to select as many items as possible into the first scale, then as many items as possible into the second scale, and so on. If this result was attained, we had found the optimal partitioning.

2.3 New Item Selection Methods

2.3.1 Objective Function

First, we introduce some notation. Often a set of J items can be partitioned in different ways into one or more Mokken scales. The total set of different partitionings into Mokken scales (and for each partitioning, possibly unscalable items) is denoted $\mathcal{M}$. From this set we seek the partitioning that best represents Mokken's objective. For a particular partitioning from $\mathcal{M}$, a scale indicator vector denoted $\mathbf{v} = (v_1, ..., v_J)$ describes the partitioning of the items, such that $v_j = 0$ denotes that item j was unscalable, $v_j = 1$ that item j was assigned to the first scale, $v_j = 2$ that item j was assigned to the second scale, and so on. Let set $\mathcal{M}$ contain F partitionings, which are indexed f, so that $f = 1, \ldots, F$. Within a partitioning, the number of Mokken scales is denoted K_f ($K_f \leq \frac{J}{2}$). Mokken scales in partitioning f are indexed k, so that $k = 1, \ldots, K_f$, and the number of items in scale k is denoted by J_{fk}.

Following Mokken's intention, given the definition of a scale, the first selected cluster contains the maximum number of items; if items remain unselected, the second cluster contains the maximum number of items; and so on, until there are no items left that constitute a Mokken scale. Hence, for partitioning f the objective function should reflect that one extra item in scale k is more important than any number of items in the subsequent shorter scales $k+1$ through K_f. An objective function satisfying this requirement, and to be used for the evaluation of partitioning f, is

$$O(\mathbf{v}_f) = \sum_{k=1}^{K_f} J^{-k} \times J_{fk}. \tag{2.1}$$

Objective function $O(\mathbf{v}_f)$ weights the number of items in scale k, J_{fk}, by J^{-k}, and then adds these products. Because the scales are ordered by the number of items they contain, $O(\mathbf{v}_f)$ assigns the greatest weight, J^{-1}, to the scale with the largest number of items, the second greatest weight, J^{-2}, to the scale with the second largest number of items, and so on. The argument of the objective function, $\mathbf{v}_f$, does not appear on the right-hand side of Equation 2.1 but K_f, J, and J_{fk} are functions of $\mathbf{v}_f$, which renders $\mathbf{v}_f$ a valid argument. By definition, $O(\mathbf{v}_f) = 0$ if one or more scales in partitioning f do not satisfy the criteria of a Mokken scale. In Appendix 1, we prove that the objective function $O(\mathbf{v}_f)$ indeed realizes Mokken's intention.

The global maximum for a particular data set is described by the scale indicator vector $\mathbf{v}_f$ that yields the highest possible value for objective function $O(\mathbf{v}_f)$, and hence is defined as $\operatorname{argmax} O(\mathbf{v}_f)$. The partitioning that represents the global maximum is denoted by $\mathbf{v}_*$.

2.3.2 Modification of AISP

The AISP is a bottom-up algorithm, so that the scale(s) produced by AISP may contain one or more items for which $H_j < c$ (i.e., a violation of Criterion 2; see also Sijtsma & Molenaar, 2002, pp. 79-80). We propose an adjusted version of AISP that does not have this problem. This version is denoted *AISP-modified*, and was modelled after stepwise regression analysis. It allows item j to leave the scale after a new item has been selected and, as a result of that, H_j has dropped below lower bound c.

2.3.3 Genetic Algorithm

We used a genetic algorithm (GA; R package `mokken`, version 2.0 and beyond; Van der Ark, 2007), which evaluates several partitionings simultaneously throughout the procedure. GA maximizes the objective function subject to the side condition that each selected cluster is a Mokken scale. Thus, the end result consists of scales that are consistent with the definition of a Mokken scale without necessarily having the highest H values possible but scales have maximum length according to the objective function. This is consistent with Mokken's definition of a scale. GA starts with an initial set of P randomly chosen partitionings, indexed p ($p = 1, ..., P$) and denoted by $\mathbf{v}_{1(0)}, ..., \mathbf{v}_{P(0)}$. Partitionings $\mathbf{v}_{1(0)}, ..., \mathbf{v}_{P(0)}$ constitute the initial population, which is denoted by $\mathcal{P}_0$. The subscript between parentheses in $\mathbf{v}_{p(t)}$ indicates the population obtained at iteration t, and $t = 0$ indicates the 0th iteration resulting in $\mathcal{P}_0$.

The probability that the global maximum $\mathbf{v}_*$ is included in $\mathcal{P}_0$ is small when J is large because P is generally much smaller than the number of partitionings F in set $\mathcal{M}$. Therefore, GA mimics an evolutionary process to find $\mathbf{v}_*$. The partitionings of $\mathcal{P}_0$ are changed in an iterative process such that after a large number of iterations, denoted by T, implying a large number of changes, at least one partitioning $\mathbf{v}_{p(T)}$ approaches $\mathbf{v}_*$. Appendix 2 discusses the details of GA. As an aside, we compared GA with an algorithm that examined all possible partitionings for limited numbers of items, and found that GA yielded the same solution and was always faster.

2.4 Comparing Three Item Selection Methods

We did a simulation study to compare AISP, AISP-modified, and GA with respect to (1) the frequency in which local maxima are found; and (2) the degree to which each procedure retrieves the true dimensionality in the data. Even though imposing restrictions on item quality may stand in the way of an optimal dimensionality retrieval, it is of interest to know to what degree the procedures can do this for particular design choices.

2.4.1 Method

Simulation Model

A two-dimensional graded response model (GRM; Samejima, 1969) was used for data simulation. Let item j have scores $0, \ldots, m$. A two-dimensional latent variable, $\boldsymbol{\theta} = (\theta_1, \theta_2)$, explains the association between the item scores. The item difficulty parameters are $\delta_{j1}, \ldots, \delta_{jm}$, and the item discrimination parameters are $(\alpha_{j1}, \alpha_{j2})$. The two-dimensional GRM gives the probability of a score of at least x ($x = 1, \ldots, m$) by means of

$$P(X_j \geq x|\, \boldsymbol{\theta}) = \frac{\exp[\sum_{l=1}^{2} \alpha_{jl}(\theta_l - \delta_{jx})]}{1 + \exp[\sum_{l=1}^{2} \alpha_{jl}(\theta_l - \delta_{jx})]}.$$

Design of the Study

The next four characteristics were fixed: (1) number of latent variables: 2; (2) sample size: $N = 1,000$; (3) lower bound: $c = 0.3$; and (4) number of replications in each design cell: # repl. = 100 (Table 2.1). A pilot study showed that the number of items J, the range of item difficulty parameters δ, and the item discrimination parameters α affected the frequency with which local maxima were found. These three factors and three other factors were varied as follows (Table 2.1).

Table 2.1: Fixed Design Characteristics and Independent Variables

Fixed characteristics	Value
Number of θs	2
Sample size	1,000
Lower bound c	0.3
Number of replications	100

Independent variables	Levels
Correlation between θs	0, 0.35, 0.7, 1
Item format	2, 5
Test length	10, 20, 40
Range of δ	[-1.5,1.5], [-3,3]
Mean item discrimination	1, 1.25, 1.5
Item selection procedure	AISP, AISP-modified, GA

Correlation between latent variables. Latent variables had a bivariate standard normal distribution. The correlation between the latent variables was either 0 (zero correlation), 0.35 (medium correlation), 0.7 (strong correlation), and 1 (unidimensional latent variable). We simulated a simple structure. The first half of the items had discrimination parameters $\alpha_{j1} = \alpha_j$ and $\alpha_{j2} = 0$, and the second half had discrimination parameters $\alpha_{j1} = 0$ and $\alpha_{j2} = \alpha_j$.

Item format. Items were either dichotomous ($m + 1 = 2$) or polytomous ($m + 1 = 5$).

Test length. The number of items was either 10, 20, or 40. We expected that more local maxima were found as J increased, because more items lead to more and longer scales thus increasing the probability of finding suboptimal partitionings.

Range of item difficulty parameters. Item difficulties δ were drawn from a continuous uniform distribution on an interval equal to either $[-1.5, 1.5]$ or $[-3, 3]$; choices were loosely based on Thissen and Wainer (1982).

Item discrimination. Item discrimination parameters were drawn from a normal distribution with mean 1, 1.25, or 1.5, and a standard deviation equal to 0.1 (Mokken et al., 1986). A pilot study showed that in combination with a standard normal θ, these α values produced ample suboptimal partitionings but lower and higher values were not effective because either no items were selected (all $H_j \ll c$) or all items were selected (all $H_j \gg c$), respectively. Mean 1 might be too low for having good quality of measurement but we emphasize that some measurement areas may typically be characterized by items having modest discrimination (e.g., inductive reasoning; de Koning, Sijtsma & Hamers, 2003). Choosing a lower bound smaller than $c = 0.3$ accommodates this situation.

Item selection procedure. AISP, AISP-modified and GA were included. Appendix 2 provides the optimal choices for the quantities that influence the efficiency of the algorithm.

Item selection procedure was a within-subject variable, and the other five independent variables were between-subject variables. Thus, AISP, AISP-modified, and GA were used to evaluate each data set.

The first dependent variable was the frequency with which an item selection procedure finds the best partitioning, abbreviated *best partitioning* and defined as follows. For a simulated data set, the partitioning $\mathbf{v}_f$ that produced the highest $O(\mathbf{v}_f)$ value out of the three values obtained for each of the item selection procedures is the best partitioning. This need not be the global maximum. For each data set we recorded for each item selection procedure whether it resulted in the best partitioning ($Y = 1$) or not ($Y = 0$). AISP-modified and GA by definition produce Mokken scales but AISP may fail because after completion of the procedure $H_j < c$ for one or more items, thus violating the second criterion for inclusion in a scale. Following Mokken (1971, p. 193), we removed such items "by hand" so that partitionings resulting from AISP also constituted Mokken scales.

We used *best partitioning* (i.e., variable Y) in a logistic regression on the design factors. Effect sizes on best partitioning per procedure were expressed by a transformation (Tabachnick & Fidell, 2007, p. 463) of an odds ratio into Cohen's (1988, p. 281) η^2, using

$$\eta^2 = \frac{[\ln(\text{odds ratio})/1.81]^2}{[\ln(\text{odds ratio})/1.81]^2 + 4}.$$

For categorical predictor variables this is a useful transformation for interpreting the effect size of logistic regression coefficients. Guidelines for interpretation are (Cohen, 1988, pp. 284-288): $\eta^2 < 0.01$ is a negligible effect; $0.01 \leq \eta^2 < 0.06$ a small effect; $0.06 \leq \eta^2 < 0.14$ a medium effect; and $0.14 \geq \eta^2$ a large effect.

The second dependent variable quantifies whether GA outperforms AISP ($Y = 1$) or not ($Y = 0$) with respect to finding the true dimensionality. True dimensionality meant either all items are in one scale ($\rho = 1$) or the first half of the items is in one scale and the second half in the other scale ($\rho < 1$). This variable was operationalized using MIN (Van der Ark & Sijtsma, 2005), which expresses the degree to which the item selection procedure misrepresented the true dimensionality. The MIN value counted the number of items incorrectly assigned, either to the wrong scale or no scale at all. For each data set, we recorded whether the MIN value of GA was smaller than the MIN value of AISP ($Y = 1$) or not ($Y = 0$). We used Y in a logistic regression on the design factors. The same measure of effect size was used as for the first dependent variable.

2.4.2 Results

The results for $\rho = 0.35$ and $\rho = 0.7$ were similar to the results for $\rho = 0$. Results for $\rho = 0.35$ and $\rho = 0.7$ were included in the statistical analyses but not tabulated to prevent tables from becoming very large. Table 2.2 shows for all conditions the frequency that AISP, AISP-modified, and GA found the best partitioning. GA always performed at least as well as AISP and AISP-modified, in particular when $\alpha = 1$ and $\alpha = 1.25$, and in many cells differences are large. For $\alpha = 1.5$, AISP, AISP-modified, and GA almost always produced a partitioning with the same objective function value, thus performing equally well. Only for 243 out of 14,400 data sets (i.e., 1.7%) did AISP and AISP-modified find different partitionings. Because their results were so similar, we report only logistic regression results for

AISP. GA almost always found the best partitioning; hence, logistic regression did not yield interesting results.

Logistic Regression Effects on Best Partitioning

For AISP, Table 2.3 shows main effects and important two-way interaction effects on best partitioning. Logistic regression produced the effects for the conditions with $\alpha = 1$ and $\alpha = 1.25$. A forward selection procedure (Miller, 2002, pp. 39-42) was used to add possible main and interaction effects. The Hosmer-Lemeshow (Hosmer & Lemeshow, 1989) statistic showed that the resulting logistic regression model fitted acceptably given the large sample size ($\chi^2 = 16.136$, $df = 8$, $p = .04$). All effects were significant, and varied from small to large. Only the medium and large effects are discussed.

As concerns main effects, the probability of finding the best partitioning decreased as the correlation between the latent traits increased; for unidimensional data ($\rho = 1$) the effect was medium ($\eta^2 = 0.158$). The probability of finding the best partitioning was larger for polytomous items than dichotomous items ($\eta^2 = 0.245$), and decreased as test length increased ($\eta^2 = 0.272$ for $J = 20$ and $\eta^2 = 0.604$ for $J = 40$). Four 2-way interactions involved item discrimination. As α increased from 1 to 1.25, the main effects of item format ($\eta^2 = 0.135$), test length ($\eta^2 = 0.069$ for $J = 20$ and $\eta^2 = 0.351$ for $J = 40$), range of δ ($\eta^2 = 0.080$), and correlation ($\rho = 1$; $\eta^2 = 0.092$) were weaker.

Logistic Regression Effects on Misrepresentation of True Dimensionality

Table 2.4 shows the average MIN value over 100 replications in each design cell. When interpreting the entries, it is most important to realize again that all three methods counter-balance finding unidimensional scales with selecting items for which $H_j \geq c$. Hence, for lower bound $c = 0.3$, item discrimination $\alpha_j = 1$ produces many wrong item assignments (first, second, and third panels), whereas for $\alpha_j = 1.5$ assignment is almost flawless (seventh, eighth, and ninth panels). If c were lowered, thus accepting lower item quality, all table entries would go down, and if c were raised all entries would go up. Thus, Table 2.4 shows (1) that

Table 2.2: Number of Replications Out of 100 in Which the Best Partitioning was Found in Each of 216 Design Cells (AISP-m Stands for AISP-modified).

				Unidimensional			Two dimensional		
α[1]	J	$m+1$	Range of δ	AISP	AISP-m	GA	AISP	AISP-m	GA
1	10	2	[-1.5,1.5]	54	55	99	82	82	99
			[-3,3]	46	46	86	70	72	84
	10	5	[-1.5,1.5]	86	86	99	100	100	99
			[-3,3]	80	80	98	96	96	100
	20	2	[-1.5,1.5]	17	19	100	43	43	98
			[-3,3]	8	8	98	26	27	86
	20	5	[-1.5,1.5]	66	66	99	86	86	99
			[-3,3]	40	38	98	61	62	99
	40	2	[-1.5,1.5]	2	2	98	3	3	98
			[-3,3]	0	0	100	1	1	99
	40	5	[-1.5,1.5]	15	17	99	44	45	100
			[-3,3]	4	6	98	18	18	100
1.25	10	2	[-1.5,1.5]	71	73	100	83	83	100
			[-3,3]	76	79	97	85	85	93
	10	5	[-1.5,1.5]	86	86	100	87	87	99
			[-3,3]	85	85	100	90	90	100
	20	2	[-1.5,1.5]	52	53	99	59	62	98
			[-3,3]	58	61	93	48	49	93
	20	5	[-1.5,1.5]	65	67	100	69	71	99
			[-3,3]	75	75	99	81	84	97
	40	2	[-1.5,1.5]	40	41	96	38	41	97
			[-3,3]	43	44	90	46	49	86
	40	5	[-1.5,1.5]	45	47	100	56	58	100
			[-3,3]	57	60	99	64	70	98
1.5	10	2	[-1.5,1.5]	99	99	100	99	99	100
			[-3,3]	94	96	98	96	96	98
	10	5	[-1.5,1.5]	100	100	100	100	100	100
			[-3,3]	100	100	100	100	100	100
	20	2	[-1.5,1.5]	100	100	100	100	100	100
			[-3,3]	95	96	99	98	99	99
	20	5	[-1.5,1.5]	100	100	100	100	100	100
			[-3,3]	100	100	100	100	100	100
	40	2	[-1.5,1.5]	100	100	100	100	100	100
			[-3,3]	82	82	100	86	91	99
	40	5	[-1.5,1.5]	100	100	100	100	100	100
			[-3,3]	100	100	100	100	100	100

[1] α = average value of discrimination parameters.

Table 2.3: Effects on Number of Replications in Which AISP Found Best Partitioning.

	Effect[1]	β	SE	η^2
Main effects				
	Intercept	1.909		
	Discrimination	-0.465	0.177	0.016
	Test length			
	- $J = 20$	-2.214	0.155	0.272
	- $J = 40$	-4.467	0.180	0.604
	Item format	2.063	0.130	0.245
	Range of δ	-1.060	0.161	0.079
	Correlation			
	- $\rho = 0.35$	-0.307	0.220	0.007
	- $\rho = 0.7$	-0.585	0.216	0.025
	- $\rho = 1$	-1.568	0.210	0.158
Interaction effects				
	Discr. × Test length			
	- $J = 20$	0.983	0.132	0.069
	- $J = 40$	2.660	0.160	0.351
	Discr. × Item format	-1.432	0.133	0.135
	Discr. × Range of δ	1.070	0.216	0.080
	Discr. × Correlation			
	- $\rho = 0.35$	0.169	0.207	0.002
	- $\rho = 0.7$	-0.310	0.208	0.007
	- $\rho = 1$	1.154	0.206	0.092

All effects were significant; $p < 0.001$.

[1] Reference categories were $\alpha = 1$, $J = 10$, $m = 1$, range of $\delta = [-1.5, 1.5]$, and $\rho = 0$.

indeed the methods react as predicted to the design features but also (2) which method obtains the best results with respect to finding true dimensionality. The results are the following.

AISP and AISP-modified often resulted in the same partitioning. When the methods did not result in the same partitioning, AISP represented true dimensionality better than AISP-modified. In general, the partitioning GA obtained represented the true dimensionality best.

We performed a logistic regression to model the probability of finding a

Table 2.4: Average MIN values over 100 Replications in Each of 216 Design Cells (AISP-m stands for AISP-modified).

				Unidimensional			Two dimensional		
α[1]	J	$m+1$	Range of δ	AISP	AISP-m	GA	AISP	AISP-m	GA
1	10	2	[-1.5,1.5]	7.04	7.10	6.86	6.17	6.18	6.13
			[-3,3]	6.35	6.45	6.02	5.10	5.17	4.98
	10	5	[-1.5,1.5]	7.94	7.93	7.89	7.26	7.26	7.26
			[-3,3]	7.42	7.44	7.35	6.68	6.68	6.65
	20	2	[-1.5,1.5]	16.10	16.21	15.52	14.73	14.76	14.30
			[-3,3]	13.89	14.17	12.86	12.52	12.64	11.74
	20	5	[-1.5,1.5]	17.35	17.35	17.09	16.02	16.01	15.98
			[-3,3]	15.77	15.83	15.25	14.83	14.85	14.62
	40	2	[-1.5,1.5]	33.80	33.91	32.34	31.85	31.98	30.39
			[-3,3]	29.67	30.01	27.93	27.99	28.45	26.03
	40	5	[-1.5,1.5]	35.69	35.70	35.00	34.27	34.28	33.89
			[-3,3]	33.27	33.27	32.07	32.06	32.13	31.17
1.25	10	2	[-1.5,1.5]	2.40	2.54	2.33	2.71	2.72	2.55
			[-3,3]	1.68	1.89	1.70	2.08	2.72	2.03
	10	5	[-1.5,1.5]	2.24	2.28	2.15	2.10	2.12	2.05
			[-3,3]	1.10	1.18	1.13	1.53	1.62	1.50
	20	2	[-1.5,1.5]	4.47	4.60	4.28	4.53	4.66	4.22
			[-3,3]	3.04	3.34	3.01	3.65	4.04	3.54
	20	5	[-1.5,1.5]	3.86	3.93	3.82	4.42	4.50	4.27
			[-3,3]	1.88	2.03	1.93	2.24	2.38	2.28
	40	2	[-1.5,1.5]	8.08	8.21	7.95	8.67	8.88	8.40
			[-3,3]	6.28	6.67	6.32	6.21	6.72	6.29
	40	5	[-1.5,1.5]	8.23	8.36	8.20	7.35	7.49	7.39
			[-3,3]	3.94	4.05	3.93	3.91	4.18	4.03
1.5	10	2	[-1.5,1.5]	0.07	0.07	0.07	0.21	0.21	0.21
			[-3,3]	0.46	0.48	0.42	0.72	0.75	0.69
	10	5	[-1.5,1.5]	0	0	0	0	0	0
			[-3,3]	0.01	0.01	0.01	0.01	0.01	0.01
	20	2	[-1.5,1.5]	0.02	0.02	0.02	0.05	0.05	0.05
			[-3,3]	0.31	0.36	0.33	0.64	0.71	0.71
	20	5	[-1.5,1.5]	0	0	0	0	0	0
			[-3,3]	0	0	0	0	0	0
	40	2	[-1.5,1.5]	0.02	0.02	0.02	0.07	0.07	0.07
			[-3,3]	0.92	0.98	0.80	1.44	1.25	1.09
	40	5	[-1.5,1.5]	0.01	0.01	0.01	0.01	0.01	0.01
			[-3,3]	0.01	0.01	0.01	0.01	0.01	0.01

[1] α = average value of discrimination parameters.

Table 2.5: Effect of Design Factors on GA Representing the True Dimensionality Better Than AISP (Score 1) or Otherwise (Score 0).

	Effect[1]	β	SE	η^2
Main effects				
	Intercept	-3.200		
	Discrimination	0.707	0.185	0.026
	Test length			
	- $J = 20$	1.909	0.176	0.218
	- $J = 40$	3.446	0.177	0.476
	Item format	-1.629	0.121	0.168
	Range of δ	0.871	0.116	0.055
	Correlation			
	- $\rho = 0.35$	0.619	0.228	0.028
	- $\rho = 0.7$	1.012	0.217	0.072
	- $\rho = 1$	1.499	0.219	0.146
Interaction effects				
	Discr. × Test length			
	- $J = 20$	-1.120	0.151	0.087
	- $J = 40$	-2.759	0.156	0.367
	Discr. × Range of δ	-1.132	0.111	0.089

All effects were significant; $p < 0.001$.

[1] Reference categories were $\alpha = 1$, $J = 10$, $m = 1$, range of $\delta = [-1.5, 1.5]$, and $\rho = 0$.

partitioning using GA that better represented the true dimensionality than the partitioning found by AISP. Because GA and AISP performed equally well for $\alpha = 1.5$, we only did the regression analyses for $\alpha = 1$ and $\alpha = 1.25$. Table 2.5 shows the logistic regression main effects and important 2-way interaction effects on whether GA produced a smaller MIN value than AISP (i.e., $Y = 1$ vs. $Y = 0$) for $\alpha = 1$ and $\alpha = 1.25$. A forward selection procedure (Miller, 2002, pp. 39-42) was used to add possible main and interaction effects. However, the Hosmer-Lemeshow statistic showed that the logistic regression model did not fit well ($\chi^2 = 25.620$, $df = 8$, $p = .001$). All effects were significant and varied from small to large. Only the medium and large effects are discussed.

Three large and medium main effects were found. The probability of $Y = 1$ was greater for dichotomous items than polytomous items ($\eta^2 = 0.168$), and increased as correlation ($\eta^2 = 0.072$ for $\rho = 0.7$; $\eta^2 = 0.146$ for $\rho = 1$) and test length ($\eta^2 = 0.218$ for $J = 20$; $\eta^2 = 0.476$ for $J = 40$) increased. Two interaction effect involved item discrimination. As α increased from 1 to 1.25, the main effect of test length decreased ($\eta^2 = 0.087$ and $\eta^2 = 0.367$, respectively) and the probability of $Y = 1$ decreased for range of δ ($\eta^2 = 0.089$).

2.5 Real Data Example

To study the partitionings produced by AISP and GA ($c = 0.3$), we used data 433 respondents provided who answered the first ten items of the Dutch translation of the Adjective Checklist (Gough & Heilbrun, 1980). The ten items 2.6 measure the trait communality. Communality may be interpreted as a response style rather than a personality trait. The scale consists of items that are either extremely popular or extremely unpopular. For example, the item "cruel" is extremely unpopular as a self-descriptive adjective. The unpopular items (indicated by an asterisk) were reversely coded. Respondents that have a high score on communality are particularly good at giving responses that are commonly accepted. This phenomenon is called *satisficing* (Krosnik, 1991). Each item consisted of an adjective, and respondents used five ordered answer categories to express the degree to which the adjective applied to them.

Table 2.6 shows that AISP produced two 4-item scales, whereas GA found one 7-item scale. Both AISP and GA did not select the items *unintelligent** and *unscrupulous**. The main difference was that AISP selected the adjective *honest* as the third item in the first 4-item scale, whereas GA left this item out of the longer 7-item scale. The detailed results for the first scale are that AISP first selected *dependable* and *reliable* ($H_{\text{dependable,reliable}} = 0.72$), then *honest* ($H_{\text{honest}} = 0.54$), and last *deceitful* ($H_{\text{deceitful*}} = 0.31$). The example shows neatly that AISP selected the third item due to its highest H_j value with respect to the start pair but the GA result shows that this locally optimal decision leads to a suboptimal final result.

Table 2.6: AISP and GA Results for Communality Items from Adjective Check List.

Communality	AISP	GA
reliable	1	1
honest	1	0
unscrupulous*	0	0
deceitful*	1	1
unintelligent*	0	0
obnoxious*	2	1
thankless*	2	1
unfriendly*	2	1
dependable	1	1
cruel*	2	1

* = reversely coded items

2.6 Discussion

GA found the best partitioning more often than AISP and AISP-modified. AISP and AISP-modified usually found the same partitionings. GA and AISP found the true dimensionality of the data more often than AISP-modified, and GA beat AISP but not always. In general, GA seems to be the best method for automated item selection in the context of Mokken scale analysis. Table 2.7 provides an overview of the advantages and the disadvantages of AISP, AISP-modified, and GA. Evaluation criteria are ordered by importance given Mokken's objectives of item selection. An option could be to start GA with the AISP solution but some trials showed that this did not improve results compared to starting with a random partitioning of the item set.

First, AISP-modified and GA were developed such that a violation of Criterion 2 ($H_j \geq c > 0$) is impossible. Second, GA finds the global maximum more easily than AISP and AISP-modified because GA is a stochastic algorithm, which moves the population away from local optima using convenient choices for the quantities that influence the efficiency of the algorithm; see Appendix 2. In contrast to GA, AISP and AISP-modified are deterministic, always producing the same partitioning in a particular data set, making it

Table 2.7: Evaluation Criteria for AISP, AISP-modified, and GA.

	Item Selection Procedure		
	AISP	AISP-modified	GA
All items satisfy the Mokken scale criteria	-	+	+
Procedure is able to find the global maximum for any data set	-	-	+
The item selection procedure is best capable of finding the true dimensionality	-	-	+
Insight in the item selection process	+	+	-
Reasonable computation time	+	+	-

impossible to fix selection errors. In the simulation study, we found effects of item format, item discrimination, test length, range of δ and correlation between θs on the probability that AISP found a local maximum. These effects are probably due to several H_js that were close to c, which caused AISP to more likely find a local maximum. Our results show that GA outperforms AISP especially when the item-scalability coefficients H_j are close to the lower bound c. This result has an important consequence. Hemker, Sijtsma, and Molenaar (1995; also see Sijtsma and Molenaar, 2002, pp. 80-82) advocated investigating the dimensionality structure of an item set by conducting Mokken scale analysis for several increasing values of c; for example $c = 0$, 0.1, 0.2, 0.3, and 0.4. When following this method, one of the investigated lower bounds is almost surely close to the values of some of the item-scalability coefficients H_j and this is likely to produce a local maximum. Hence, for the Hemker et al. (1995) method, GA is always preferred over AISP.

Third, GA more often found the true dimensionality than AISP and AISP-modified. Thus, GA found better partitionings with respect to the objective function and these partitionings better represented the true dimensionality. Fourth, because AISP and AISP-modified are deterministic, they provide exact insight in the item selection process. The genetic algorithm generates a huge number of random processes, which do not provide intelligible information on item quality. Sometimes, AISP finds a better partitioning than GA. We recommend to include both AISP and GA in the same software package (see Van der Ark, 2007), use the same objective function for both, and adopt the solution that generated the highest objective function value. Next, for each Mokken scale resulting from the best partitioning algorithm the scalability coefficients and model assumptions should be investigated. Fifth, for the default setting for 40 items GA takes approximately 15 minutes to complete. Computation time increases as test length increases but computers are rapidly becoming faster, so that this problem may be obsolete before long.

Appendix 1

Suppose that two partitionings, $\mathbf{v}_1$ and $\mathbf{v}_2$, have the same number of items in the first $k-1$ scales; that is, $J_{1i} = J_{2i}$ for $i = 1, \ldots, k-1$. Henceforth, i is used as scale index. Further, suppose that $\mathbf{v}_1$ has a items more in scale k than $\mathbf{v}_2$; that is, $J_{1k} = J_{2k} + a$. Finally, suppose that nothing is known about the remaining scales $k+1, k+2, \ldots$. The function values of $\mathbf{v}_1$ and $\mathbf{v}_2$ are

$$O(\mathbf{v}_1) = \sum_{i=1}^{k-1} J^{-i} J_{1i} + J^{-k} J_{1k} + \sum_{i=k+1}^{K_1} J^{-i} J_{1i} \tag{2a}$$

and

$$O(\mathbf{v}_2) = \sum_{i=1}^{k-1} J^{-i} J_{1i} + J^{-k}(J_{1k} - a) + \sum_{i=k+1}^{K_2} J^{-i} J_{2i}. \tag{2b}$$

Under these conditions, the smallest possible value of $O(\mathbf{v}_1)$ should always exceed the largest possible value of $O(\mathbf{v}_2)$; this is what we prove next.

The sums on the right-hand sides in equations 2a and 2b with respect to the first $k-1$ scales are equal and are replaced by symbol A_{k-1}. The minimum value of $O(\mathbf{v}_1)$ is obtained if there are no scalable items left after scale k; that is, if $\sum_{i=k+1}^{K_1} J^{-i} J_{1i} = 0$. The maximum value of $O(\mathbf{v}_2)$ is obtained for $a = 1$; that is, a minimal difference is obtained between $\mathbf{v}_1$ and $\mathbf{v}_2$ in the kth scale; and if all items that remain after scale k have been selected in the $(k+1)$st scale (i.e., $\sum_{i=k+1}^{K_2} J_{2i}$), which then receives the greatest possible weight, $J^{-(k+1)}$. It may be noted that this maximum value of $O(\mathbf{v}_2)$ is an upper bound, because scale $k+1$ cannot contain more items than scale k (i.e., given $a=1$, we have that $J_{2(k+1)} \leq J_{2k} = J_{1k} - 1$); so, no more than $J_{1k} - 1$ items can in fact receive the greatest weight, $J^{-(k+1)}$. This reduces equations 2a and 2b for the minimum value of $O(\mathbf{v}_1)$ and the maximum value of $O(\mathbf{v}_2)$ to

$$O(\mathbf{v}_1) = A_{k-1} + J^{-k} J_{1k} \tag{3a}$$

and

$$O(\mathbf{v}_2) \leq A_{k-1} + J^{-k}(J_{1k} - 1) + J^{-(k+1)} \sum_{i=k+1}^{K_2} J_{2i}. \tag{3b}$$

It follows from equations 3a and 3b that

$$\begin{aligned} O(\mathbf{v}_1) - O(\mathbf{v}_2) &\geq J^{-k} J_{1k} - \left[J^{-k}(J_{1k} - 1) + J^{-(k+1)} \sum_{i=k+1}^{K_2} J_{2i} \right] \\ &= J^{-k} - J^{-(k+1)} \sum_{i=k+1}^{K_2} J_{2i}. \end{aligned} \tag{4}$$

The difference between $O(\mathbf{v}_1)$ and $O(\mathbf{v}_2)$ is positive if

$$J^{-k} > J^{-(k+1)} \sum_{i=k+1}^{K_2} J_{2i}. \tag{5}$$

The right-hand side and the left-hand side of Equation 5 are equal if $\sum\limits_{i=k+1}^{K_2} J_{2i} = J$. Because some items have already been selected in scale k, it is always true that $\sum\limits_{i=k+1}^{K_2} J_{2i} < J$ and, therefore, Equation 5 is always true. This result completes the proof.

Appendix 2

The first iteration entails selection, crossover, and mutation of $\mathcal{P}_0$, which yields population $\mathcal{P}_1 = \mathbf{v}_{1(1)}, ..., \mathbf{v}_{P(1)}$. In iteration t, population

$\mathcal{P}_{t-1} = \mathbf{v}_{1(t-1)}, ..., \mathbf{v}_{P(t-1)}$ is changed into population $\mathcal{P}_t = \mathbf{v}_{1(t)}, ..., \mathbf{v}_{P(t)}$. Across the iterations, the partitioning yielding the largest value of the objective function is saved and used as the best partitioning. There is no guarantee that a genetic algorithm finds the global optimum but an increase in T and P produces an increase in the number of partitionings being evaluated and hence a higher probability of finding the global maximum.

The details of GA are the following. We discuss the steps in iteration 0 (i.e., steps 1 and 2) and the steps taken in each of the next iterations (i.e., steps 3 through 8). We also discuss the specific configuration of GA that we use in the present study. Table 2.8 provides an example of the process. The initial population and the first iteration of GA are described in Table 2.8 using $P = 4$, $J = 6$, and $c = 0.3$. Let K_{max} denote the maximum number of scales that can be selected from J items, π_{cross} the probability of a partitioning from the population to be selected for a crossover, and π_{mutate} the probability that an item in a partitioning in the population mutates.

Step 1: Initial population. Integers are randomly drawn from a discrete uniform distribution in the interval $[1, K_{max}]$ and assigned to each item j of each partitioning in the initial population. Table 2.8 (part 1a) shows a possible population consisting of random partitionings.

Step 2: Reparation and evaluation of the initial population. If the partitionings in the initial population do not satisfy the definition of a scale, the partitionings are repaired. In this reparation process, items that violate Criterion 2 ($H_j \geq c$) are removed. If the definition of a scale still is not satisfied, items that violate Criterion 1 ($\rho_{ij} > 0$) are removed. Finally, for each partitioning in the initial population, $\mathcal{P}_0$, objective function value $O(\mathbf{v}_{p(0)})$ is computed. Table 2.8 (part 1b) shows an example of an initial population.

Step 3: Selecting partitionings. From population $\mathcal{P}_{t-1}$, P partitionings are randomly drawn with replacement from a multinomial distribution with probabilities

$$\pi_{p(t-1)} = \frac{O(\mathbf{v}_{p(t-1)})}{\sum_{p=1}^{P} O(\mathbf{v}_{p(t-1)})},\ p = 1, ..., P.$$

For $t = 1$, Table 2.8 (part 1b, last column) shows the probabilities of being selected from population $\mathcal{P}_0$ into population $\mathcal{P}_1$. In general, the partitionings that are selected for the next population are denoted $v^s_{p(t)}$ (e.g., Table 2.8, part

Table 2.8: Example of GA. Numbers Having Changed Relative to the Former Subtable are in Bold Face.

Part 1: Initial population

1a: Random partitionings

Partitioning	Item number					
	1	2	3	4	5	6
1	1	1	1	3	2	2
2	3	2	2	1	3	1
3	2	2	1	1	3	3
4	2	3	2	2	2	2

1b: Repaired partitionings

Partitioning	Item number						$O(\mathbf{v}_{p(0)})$	$\pi_{p(0)}$
	1	2	3	4	5	6		
$\mathbf{v}_{1(0)}$	1	1	1	**0**	2	2	0.556	0.311
$\mathbf{v}_{2(0)}$	**0**	**1**	**1**	**0**	**0**	**0**	0.333	0.186
$\mathbf{v}_{3(0)}$	2	2	1	1	3	3	0.398	0.223
$\mathbf{v}_{4(0)}$	1	**0**	1	1	**0**	**0**	0.5	0.280

Part 2: An example of a possible first iteration

2a: Selected partitionings

Partitioning	Item number					
	1	2	3	4	5	6
$\mathbf{v}^s_{1(1)}$	1	0	1	1	0	0
$\mathbf{v}^s_{2(1)}$	1	1	1	0	2	2
$\mathbf{v}^s_{3(1)}$	2	2	1	1	3	3
$\mathbf{v}^s_{4(1)}$	1	1	1	0	2	2

2b: Crossover of partitionings

Partitioning	Item number					
	1	2	3	4	5	6
$\mathbf{v}^c_{1(1)}$	1	0	1	1	0	0
$\mathbf{v}^c_{2(1)}$	1	**2**	**1**	**1**	**3**	2
$\mathbf{v}^c_{3(1)}$	2	**1**	**1**	**0**	**2**	3
$\mathbf{v}^c_{4(1)}$	1	1	1	0	2	2

2c: Mutation of partitionings

Partitioning	Item number					
	1	2	3	4	5	6
$\mathbf{v}^m_{1(1)}$	**2**	0	1	1	0	0
$\mathbf{v}^m_{2(1)}$	1	2	**3**	1	3	2
$\mathbf{v}^m_{3(1)}$	2	1	1	0	2	3
$\mathbf{v}^m_{4(1)}$	1	1	1	**1**	2	2

2d: Repaired partitionings

Partitioning	Item number						$O(\mathbf{v}_{p(1)})$	$\pi_{p(1)}$
	1	2	3	4	5	6		
$\mathbf{v}_{1(1)}$	**0**	0	1	1	0	0	0.333	0.187
$\mathbf{v}_{2(1)}$	1	**0**	**2**	1	**2**	**0**	0.389	0.219
$\mathbf{v}_{3(1)}$	**0**	1	1	0	**0**	**0**	0.333	0.187
$\mathbf{v}_{4(1)}$	1	1	1	1	2	2	0.722	0.406

2a).

Step 4: Crossovers. The exchange of a subvector of partitioning $\mathbf{v}^s_{p(t)}$ with the corresponding subvector of partitioning $\mathbf{v}^s_{q(t)}$ is called a crossover. Each of the P partitionings $\mathbf{v}^s_{p(t)}$ has a probability π_{cross} of being selected for a crossover with another partitioning. The partitionings that are selected for a crossover are divided into pairs $(\mathbf{v}^s_{p(t)}, \mathbf{v}^s_{q(t)})$. For each pair of partitionings $(\mathbf{v}^s_{p(t)}, \mathbf{v}^s_{q(t)})$, two random numbers a and b are drawn from a discrete uniform distribution in the interval $[1, J]$. The starting point of the subvector of $\mathbf{v}^s_{p(t)}$ is denoted by a and the end point by b. This subvector of $\mathbf{v}^s_{p(t)}$ is then exchanged with the corresponding subvector of $\mathbf{v}^s_{q(t)}$. If $a < b$, $(v^s_{p(t)a}, ..., v^s_{p(t)b})$ is exchanged with $(v^s_{q(t)a}, ..., v^s_{q(t)b})$; if $a = b$, $v^s_{p(t)a}$ is exchanged with $v^s_{q(t)a}$; and if $a > b$, $(v^s_{p(t)1}, ..., v^s_{p(t)b})$ and $(v^s_{p(t)a}, ..., v^s_{p(t)J})$ are exchanged with $(v^s_{q(t)1}, ..., v^s_{q(t)b})$ and $(v^s_{q(t)a}, ..., v^s_{q(t)J})$, respectively. After crossover, the partitionings are denoted $\mathbf{v}^c_{p(t)}$. Table 2.8 (part 2b) shows examples of partitionings after crossover, where $\mathbf{v}^s_{2(1)}$ and $\mathbf{v}^s_{3(1)}$ were selected, and $a = 2$ and $b = 5$.

Step 5: Mutations. Mutation entails the random assignment of an item to another scale. Because the correct number of scales is unknown a priori, in this step a random process determines whether an additional scale $K_f + 1$ is formed in addition to the K_f existing scales. The item is assigned to either one of the K_f existing scales or the new scale $K_f + 1$. Table 2.8 (part 2c) shows examples of mutations.

Step 6: Reparation and evaluation of population t. Partitionings in population t that do not satisfy the definition of a scale are repaired. Items violating Criterion 2 are removed. If the definition of a scale is still dissatisfied, items that violate Criterion 1 are removed. Finally, for each partitioning in $\mathcal{P}_t$, objective function value $O(\mathbf{v}_{p(t)})$ is computed. Table 2.8 (part 2d) shows an example of the population after the first iteration.

Step 7: Storage of the best partitioning. Let $\mathbf{v}^{best}_{(t)}$ denote the best partitioning found in iteration t, and let $\mathbf{v}^{best}$ denote the best partitioning found in the first $t - 1$ iterations. At the end of iteration t, if $O(\mathbf{v}^{best}_{(t)}) > O(\mathbf{v}^{best})$, then the best partitioning of $\mathcal{P}_t$ is also the best partitioning of all former populations, and it is stored as the new best partitioning; that is, $\mathbf{v}^{best}$ becomes $\mathbf{v}^{best}_{(t)}$. If $O(\mathbf{v}^{best}_{(t)}) \leq O(\mathbf{v}^{best})$, then $\mathbf{v}^{best}$ remains unchanged. In the latter case, it is ascertained that $\mathbf{v}^{best}$ is contained in $\mathcal{P}_t$ by replacing the worst partitioning of $\mathcal{P}_t$ with respect to

$O(\mathbf{v}_{p(t)})$ by $\mathbf{v}^{best}$. This procedure, which always saves the best partitioning from the population, provides an example of an elitist model (Michalewicz, 1994, p. 61).

Step 8: Convergence of GA. GA stops when $O(\mathbf{v}^{best})$ has not changed for Q iterations. The literature does not provide a value for Q; hence, the researcher must specify this value.

We did a pilot study to find convenient values of P, π_{cross}, π_{mutate}, and Q for which GA most often found the global maximum. We found that $P = 20$ may be considered to represent a sufficiently large population size for GA to perform well. This was the value used in this study. The pilot study also showed that for $P = 20$, the combination of $\pi_{cross} = .5$ and $\pi_{mutate} = .1$ resulted most often in the global maximum. Hence, these values were used in this study. The choice of Q was less straightforward. The pilot study showed that for $J = 10$, $Q = 10,000$ is a reasonable choice, and that for $J = 20$ and $J = 40$, Q should at least be equal to 100,000 and 1,000,000, respectively. Hence, if J doubles, Q increases with a factor equal to 10. These results were used in the simulation study reported in this article.

Chapter 3

Multi-method analysis of the internal structure of the Type D Scale-14 (DS14)*†

Abstract

The Type D Scale-14 (DS14) measures distressed (also, Type D) personality by assessing the medium-level trait negative affectivity that encompasses the low-level traits dysphoria, anxiety, and irritability, and the medium-level trait social inhibition that encompasses low-level traits social discomfort, reticence, and lack of social poise. The literature discusses three different structural models of the DS14. The goal of this study was to investigate which of the three models best describes the internal structure of the DS14. We used three methods to investigate the internal structure of the DS14 items using data collected in representative samples from the Dutch general population ($N = 3,181$). The methods were exploratory factor analysis, confirmatory factor analysis, and Mokken scale analysis. Exploratory factor analysis suggested a two-factor structure without evidence of the low-level factors, and the other two methods showed evidence of a three-level structure including the low-level factors. A two-factor model with correlated errors for items defining low-level traits adequately describes the data. The results support the three-level hierarchical model as a conceptual model for Type D personality, and support the interpretation of DS14 scores on item subsets representing medium-level traits and low-level traits.

[*] This chapter has been published as: Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2012). Multi-method analysis of the internal structure of the Type D Scale-14 (DS14). *Journal of Psychosomatic Research, 72*, 258-265

[†] Acknowledgements: We thank Wobbe Zijlstra for making available an SPSS version of the DS14 data and Johan Denollet for critical comments on a previous draft of this chapter.

3.1 Introduction

Distressed personality (Denollet, 2000; Denollet, Schiffer, & Spek, 2010; Kupper & Denollet, 2007), Type D for short, is a psychological risk factor for morbidity and mortality in patients suffering from cardiovascular disease (De Jonge et al., 2007; Kupper & Denollet, 2007; O'Dell, Masters, Spielmans, & Maisto, 2011). Type D is a hierarchically structured (Reise, Waller, & Comrey, 2000) personality construct. The general Type D trait represents the high level of the hierarchy (Figure 3.1). At the medium level, two traits drive behavior: Negative affectivity (NA) involves the experience of negative emotions across time and situations, and social inhibition (SI) the suppression of emotions in social interaction. The inhibition to express negative emotions in social interactions—that is, high levels of both NA and SI—defines Type D. At the low level of the hierarchy, feelings of dysphoria, anxious apprehension, and irritability drive NA, and discomfort in social situations, reticence, and lack of social poise drive SI (Denollet 2005; Emons, Meijer, & Denollet, 2007).

Type D is much debated. Ferguson et al. (2012; also, Coyne et al. 2011; Grande et al. 2011) concluded that distressed personality more likely is a continuum reflecting degree than the more widely accepted categorization of individuals into Type D or non Type-D. Their position supports the three-level model as a theoretical candidate for the explanation of distressed personality. We compared the three-level model with a two-level model excluding the subtraits level and another two-level model allowing correlated errors to obtain better model fit.

Other controversies with respect to Type D are the following. Coyne et al. (2011) and Grande et al.(2011) did not find support that cardiac patients with Type D had a greater mortality risk, thus contradicting previous research (Aquarius et al. 2009; Schiffer, Smith, Pedersen, Widdershoven, & Denollet, 2011). Dannemann et al. (2010) concluded that Type D classification is unstable among cardiac patients before and after surgery. Williams, Curren, and Bruce (2011) concluded that Type D and alexithymia are correlated but separate traits but Grande, Glaesmer, and Roth (2010) found that the SI scale does not distinguish shyness and introversion. Hence, there are doubts about SI's uniqueness.

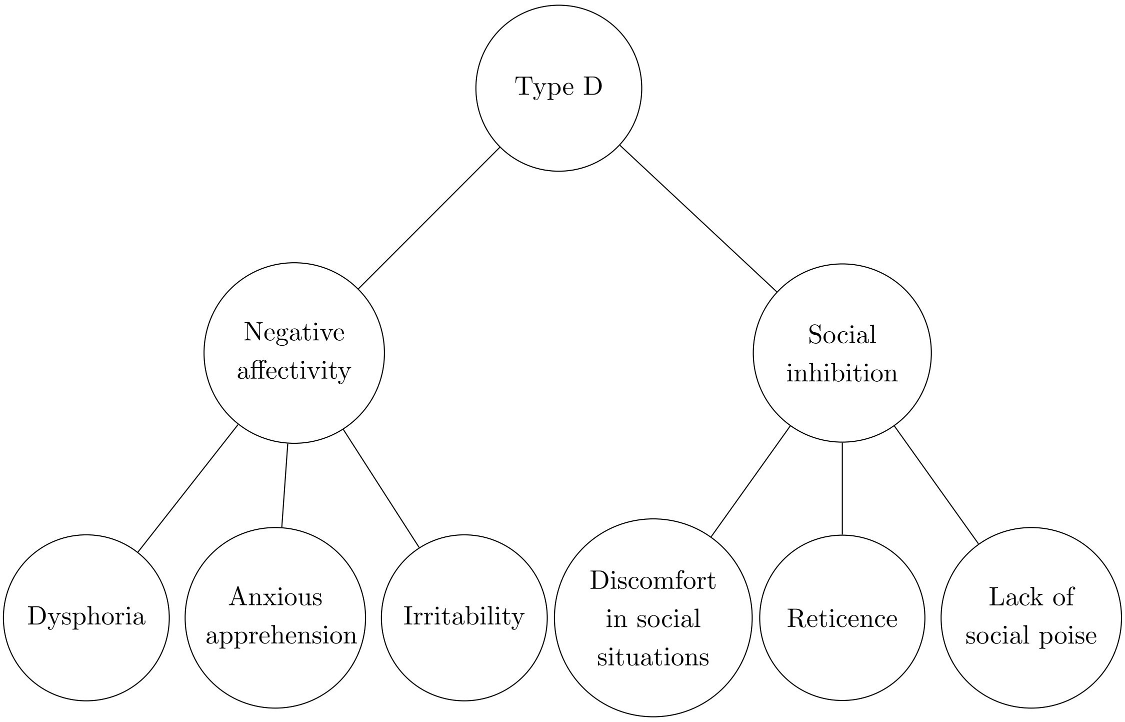


Figure 3.1: Hierarchy of the Type D construct.

The item structure of the Type D Scale-14 (DS14; Denollet et al, 2010) reflects the theoretical three-level hierarchy, and uses 14 items to assess Type D, NA (7 items) and SI (7 items), and the NA and SI subtrait triplets (Table 3.1). Different item subsets from the two seven-item sets assess the two low-level subtrait triplets. Each item statement is assessed on five ordered categories, scored 0 through 4. The NA-scale and the SI-scale yield two total scores, and if both scores are at least 10 points, the patient is diagnosed Type D (Emons et al., 2007). Thus, following the hypothesis that inhibition to express negative emotions in social interaction defines Type D, patients scoring in excess of particular cutoffs on both scales are diagnosed Type D. The dichotomy into Type D and non Type-D serves the practical purpose to determine a diagnosis.

Confirmatory factor analysis (CFA) of DS14 data revealed three different internal item structures, two of which suggest doubt about the correctness of the theoretical three-level hierarchy; see Figure 3.2. The "Two-factor Model" represents a two-level hierarchy with NA and SI factors that distinguish the

Table 3.1: Item Content, Medium-Level and Low-Level Scales for the Items of the DS14 (Denollet, 2005; Emons et al., 2007).

Item	Content	Position in DS14	Low-level scale
Negative affectivity scale			
N1	Often feels unhappy	4	Dysphoria
N2	Takes gloomy view of things	7	Dysphoria
N3	Is often down in the dumps	13	Dysphoria
N4	Worries about unimportant things	2	Anxious apprehension
N5	Often worries about something	12	Anxious apprehension
N6	Is easily irritated	5	Irritability
N7	Is often in a bad mood	9	Irritability
Social inhibition scale			
S1	Inhibited in social interactions	6	Discomfort in social situations
S2	Difficulties starting a conversation	8	Discomfort in social situations
S3	Does not find things to talk about	14	Discomfort in social situations
S4	Closed kind of person	10	Reticence
S5	Keeps others at a distance	11	Reticence
S6	Makes contact easily	1	Lack of social poise (reversed keyed)
S7	Often talks to strangers	3	Lack of social poise (reversed keyed)

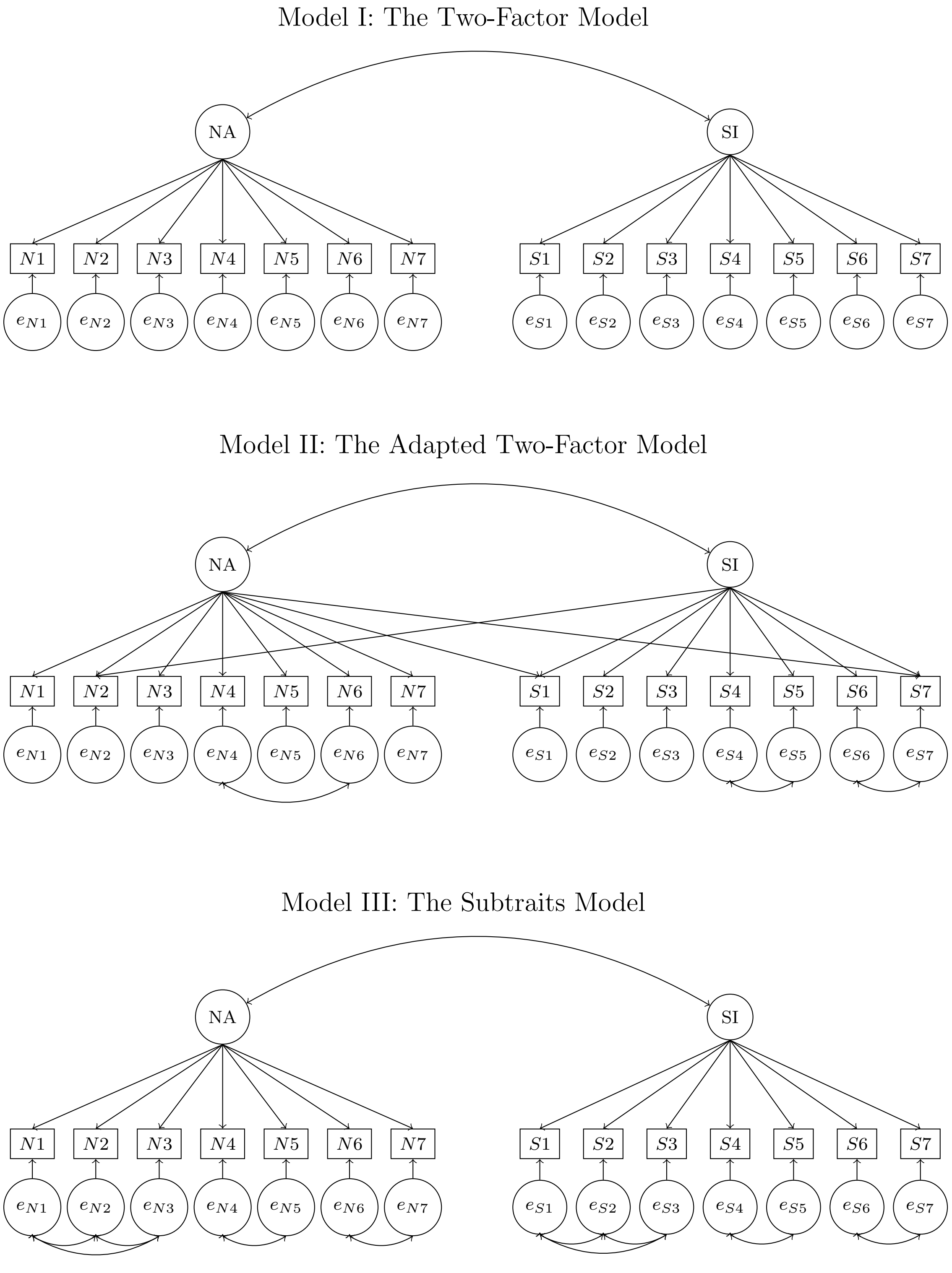
NA-scale and the SI-scale, but ignores the theoretical subtrait triplet structure (Grande, Romppel, Glaesmer, Petrowski, & Herrmann-Lingen, 2010; Lim et al., 2011; Spindler, Kruse, Zwisler, & Pedersen, 2009; Yu, Thompson, Yu, Pedersen & Denollet, 2010). The model does not explicitly incorporate a higher-order factor for modeling Type D but allows the two factors to correlate, thus suggesting an explanatory higher-order factor. The magnitude of the correlation between the factors suggests the degree to which a higher-order factor is plausible. The "Adapted Two-Factor Model" is based on modification indices of the Grande, Romppel, et al. (2010) Two-Factor Model, allowing cross-loadings and correlated error terms. The "Subtraits Model" (Svansdottir et al., 2011; Zohar, Denollet, Lev Ari, & Cloninger, 2011) represents the three-level hierarchy by means of a factor structure with positively correlated error terms that model the low-level subtraits and positively correlated factor scores that model the high-level Type D. The question is whether a careful analysis of DS14 data can provide more conclusive evidence of which theoretical model for the Type D construct is correct.

The goal of this study was to use three psychometric methods for assessing internal structure to compare the three factorial models for the DS14. The three methods provide different statistical perspectives. The methods are exploratory factor analysis (EFA), CFA and Mokken scale analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002); see Emons, Sijtsma, and Pedersen (2012) for a similar internal-structure study of the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983).

The outline of this article is as follows. First, we discuss research that used EFA and CFA to study the internal structure of the DS14. Second, we discuss MSA and how MSA may lead to results different from EFA and CFA. Third, we discuss the internal structure of the DS14 suggested by EFA, CFA, and MSA. Fourth, we discuss consequences of the results for the Type D structure and the practical use of the DS14.

3.2 Factor Analysis Results for Type D

Traditionally, EFA was the common method for assessing the internal structure of the DS14 in various populations (Denollet, 2005; Bergvik, Sørlie, Wynn, &



Model I: The Two-Factor Model

Model II: The Adapted Two-Factor Model

Model III: The Subtraits Model

Figure 3.2: A graphical representation of the three models investigated with CFA.

Sexton, 2010; Hausteiner, Klupsch, Emeny, Baumert, & Ladwig, 2010). Recently, CFA has become more popular (Grande, Romppel, et al., 2010; Svansdottir et al., 2011; Zohar et al., 2011). MSA in combination with EFA and CFA was used to analyze the Addiction Severity Index (Alterman, Cacciola, Habing, & Lynch, 2007), the HADS (Emons et al., 2012), the Minnesota Multiphasic Personality Inventory (Meijer & Baneke, 2004), and the Self-Concealment Scale (Wismeijer, Sijtsma, Van Assen, & Vingerhoets, 2008). We discuss studies that used EFA and CFA to assess the internal structure of the DS14.

3.2.1 Exploratory Factor Analysis

Denollet (2005), Svansdottir et al. (2011), Zohar et al. (2011), Bergvik et al. (2010), Hausteiner et al. (2010), and Yu, Zhang, and Liu (2008) used EFA to assess the internal structure of the DS14. EFA extracts the number of factors and the factor loadings from the data (Bollen, 1989, p. 228). Two rules determine the number of factors. The first rule equates the number of factors to the number of eigenvalues exceeding 1 but is vulnerable to chance capitalization, which leads to overestimation of the number of factors. Horn (1965) and Reise et al. (2000) proposed parallel analysis to correct for the overestimation. Parallel analysis compares the eigenvalues with eigenvalues generated from artificial data sets based on a multivariate normal distribution with zero correlation between the items, and maintains the eigenvalues that are "significantly" larger than 1. The second rule selects the eigenvalues to the left of the elbow in the scree plot (Reise et al., 2000) but decisions may be difficult if a sharp elbow does not appear.

The six studies concluded that a two-factor structure best described the data. Denollet (2005) found an interpretable, orthogonal two-factor structure, in which Item 6 (Table 3.1) had a cross-loading on the NA-scale. In a cardiovascular-patient group, Yu et al. (2008) found a cross-loading of Item 7 on the SI-scale, and in a control group they found cross-loadings of items 6, 10, and 14 on the NA-scale and items 7 and 13 on the SI-scale. Bergvik et al. (2010) found cross-loadings for items 6, 10, and 11 on the NA-scale but Zohar et al. (2011) and Hausteiner et al. (2010) did not find cross-loadings. Svansdottir et al. (2011) obtained an interpretable, oblique two-factor structure without cross-loadings larger than 0.30. The improved fit of the factor model after oblique rotation tentatively suggests that a higher-order factor explains the correlation between the factors. This is the “Two-Factor Model”.

3.2.2 Confirmatory Factor Analysis

Four studies used CFA to test the DS14 two-factor structure found in previous EFA analyses. Table 3.2 shows three fit indices used in the four studies; the root mean square error of approximation (RMSEA), the Tucker-Lewis index [TLI; also known as non-normed fit index (NNFI)], and the comparative fit index (CFI). A model fits the data if $RMSEA < 0.08$, $TLI > 0.90$, and $CFI > 0.90$ (Bentler, 1989; Browne & Cudeck, 1993). CFI is an incremental fit index, which compares the fit of the specified model to a nested baseline model (Hu & Bentler, 1999) but does not provide information about absolute fit. Except for $CFI = 0.98$ (Spindler et al., 2009), the other fit indices reported in the four studies produced similar conclusions. Because the four studies were unclear about the models they compared, one cannot meaningfully compare the CFI values.

Table 3.2: Fit Indices for the Two-Factor Model Using CFA.

Study	RMSEA	TLI	CFI
Grande, Romppel, et al. (2010)	0.09	0.89	0.91
Lim et al. (2011)	0.08	0.90	0.92
Spindler et al.(2009)	0.08	-	0.98
Yu et al. (2010)	0.08	0.91	0.93

Note: RMSEA is root mean square error of approximation, TLI is Tucker-Lewis index, and CFI is comparative fit index.

Hu and Bentler (1999) suggested that misspecified factor models are accepted too easily, and proposed to use the model selection criteria $RMSEA < 0.06$, $TLI > 0.95$, and $CFI > 0.95$. Grande, Romppel, et al. (2010) used these rules, but the other three studies used the traditional rules. Grande, Romppel, et al. (2010) rejected the Two-Factor Model. Using the new rules, Yu et al. (2010) and Lim et al. (2011) would have found that the fit of the Two-Factor Model was inadequate, whereas Spindler et al. (2009) might have raised doubts about the fit of the Two-Factor Model.

Grande, Romppel, et al. (2010) used modification indices to obtain an acceptably fitting model. The resulting model ($RMSEA = 0.06$, $TLI = 0.96$, $CFI = 0.97$) allowed cross-loadings for items 3 and 6 on the NA-factor and for Item 7 on the SI-factor, and correlations between the error terms of items 1 and 3, items 10 and 11, and items 2 and 5. This is the "Adapted Two-Factor Model". Modification indices are often used to obtain a model that fits the sample data without theoretical justification, thus inducing chance capitalization (Bollen, 1989, pp. 296 and 304).

Svansdottir et al. (2011) used CFA to investigate the three-level Type D model. This is the "Subtraits Model". The authors allowed correlating factors and correlating error terms of items representing low-level traits. This produced acceptable fit ($RMSEA = 0.06$ and $CFI = 0.95$). Zohar et al. (2011) investigated the Subtraits Model assuming zero correlation between NA and SI factors. This two-level model excludes a Type D personality trait, but produced worse fit ($RMSEA = 0.07$ and $CFI = 0.94$).

3.3 Mokken Scale Analysis

MSA evaluates whether a set of items is consistent with the monotone homogeneity model (Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002; Van Schuur, 2011) and thus constitutes a scale. A scale consistent with the monotone homogeneity model allows the ordering of persons by means of their total scores (Sijtsma & Molenaar, 2002, pp. 22-23). Because the DS14 uses total scores for the NA and SI scales, it is important to investigate whether the monotone homogeneity model is consistent with the data so as to justify the use of total scores.

The monotone homogeneity model is based on three assumptions. *Unidimensionality* and *local independence* together define the total score to reflect one trait. Within the context of CFA, unidimensionality means that all items load on the same factor and local independence means that the error terms are uncorrelated. The *monotonicity* assumption defines the regression of the mean item score on the scale score, also known as the latent variable, to be monotone nondecreasing. The regression is better known as the item response function (Sijtsma & Molenaar, 2002). Monotonicity can be investigated by inspecting whether the regression of the mean score of item j on the total score on the items except item j is a nondecreasing function (Sijtsma & Molenaar, 2002).

For item j, the item scalability coefficient H_j (Sijtsma & Molenaar, 2002) reflects the strength of the relationship of the item with the scale score based on the total score on $J - 1$ selected items except item j. Under the monotone homogeneity model, H_j values vary between 0 and 1. Higher H_j values imply that the item better discriminates low scale scores and high scale scores. MSA defines a scale consisting of J items as follows: (1) all inter-item correlations are positive; that is, for items j and k, and correlation ρ, $\rho_{jk} > 0$, for all item pairs; and (2) for a value c between 0 and 1 chosen by the researcher, all item scalability coefficients are at least as large as c; that is, $H_j \geq c > 0$, for all items. By default MSA uses $c = 0.3$ but researchers may choose a different value thus defining what they consider minimally acceptable discrimination. Additionally, a total-scalability coefficient H is provided with a guideline for the discrimination power of the whole scale (Mokken, 1971, p. 185): if $0.3 \leq H < 0.4$, the scale is weak; if $0.4 \leq H < 0.5$, the scale is moderate; if $H \geq 0.5$ the scale is strong; and if $H < 0.3$ the items are unscalable.

For MSA, two computationally different item selection methods (Straat, Van der Ark, & Sijtsma, in press) select as many items for which $H_j \geq c$ as possible into the same scale. The automated item selection procedure (AISP; Sijtsma & Molenaar, 2002, chap. 5) starts with the two best-scalable items and adds items one by one until no items remain that satisfy the criterion of $H_j > c$. Items are chosen such that in each selection step total-scalability coefficient H is maximized. From items remaining unselected, AISP selects as many as possible into a second scale, and so on. Finally, items may be left unscalable. Because in each step one item is selected, AISP considers a limited number of item combinations and the optimal scale may not be found. The genetic algorithm (GA; Straat et al., in press) seeks the optimal scale by smartly finding its way through all possible item subsets without having to consider each scale separately. AISP and GA may produce somewhat different scales, especially when items have H_j coefficients close to c (Straat et al., in press).

One could argue that the choice of lower bound c is arbitrary but running AISP and GA for different c values neutralizes this criticism. Sijtsma and Molenaar

(2002, pp. 80-82) recommended investigating the internal structure of an item set by running AISP for $c = 0$ and using increments of 0.05 until, say, $c = 0.55$. We used GA similarly. Across different AISP and GA analysis, as c increases the pattern of item clusters found suggests the internal structure of the item set. For example, if for Type D only the high-level trait and the two medium-level traits are active, then (a) for low c values, as all items are driven by the Type D trait they are all selected in one scale; (b) for higher c values, items that are also driven by the NA trait are selected in one scale and items also driven by SI in another scale; and (c) for the highest c values AISP and GA break down the NA and SI scales as the items do not have anything in common anymore that produces even higher H_js. If also the low-level traits are active, (a) and (b) produce the same results but (c) for the highest c values smaller scales are found, each reflecting a low-level trait.

MSA has two advantages over EFA and CFA (Emons et al., 2012; Wismeijer et al., 2008). First, MSA requires monotone nondecreasing relationships between items and the latent variable but EFA and CFA assume linear relationships, which is more restrictive. Thus, MSA facilitates a better fit to the data. Second, MSA is explicitly suited for discrete item scores such as the DS14 item scores. EFA and CFA assume that item scores are continuous and normally distributed for statistical testing (Bollen, 1989, p. 418) and for determining the number of factors (Tabachnick & Fidell, 2007, p. 613) but real item scores are discrete and by definition nonnormal. Several authors (e.g., Bernstein & Teng, 1989; Dolan, 1994; Olsson, 1979) investigated this misfit and concluded that for fewer than seven ordered item scores factor analysis may produce artifactual factors (so-called difficulty factors (McDonald & Ahlawat, 1974), biased factor loadings, and inflated chi-square statistics (Dolan, 1994; Lubke & Muthén, 2004). A possible solution for these problems is to use polychoric correlations.

3.4 Method

3.4.1 Participants

A local ethics committee at Tilburg University (protocol number: 2006/1101) approved of this study. The sample consisted of 3,181 participants from the

Dutch general population. Two gender levels and six age levels (20-29, ..., 60-69, 70-80) served as stratification criteria, and quota sampling produced twelve equal-sized groups. Research assistants approached participants personally or by phone, explained the study's purpose, handed over an informed consent form and a questionnaire, and participants returned both in closed envelopes to the research assistants (October 1, 2006—December 15, 2008). Returned questionnaires were coded by number for purposes of data collection tracking but were otherwise anonymous. Two-way imputation (Bernaards & Sijtsma, 2000; Van Ginkel, Van der Ark, & Sijtsma, 2007) was used to estimate missing item scores (0.39 %).

3.4.2 Analyses

Exploratory Factor Analysis

We used SPSS version 18 for EFA on product-moment correlations and polychoric correlations. Parallel analysis used 1,000 random data sets to determine the number of factors. Oblimin rotation was used to obtain an interpretable factor structure. We interpreted all factor loadings exceeding 0.3 (Tabachnick & Fidell, 2007, p. 649). Factor correlations greater than 0.4 were considered tentative support for a higher-order factor representing Type D. Cronbach's alpha was used to assess reliability.

Confirmatory Factor Analysis

We used AMOS version 19 for CFA on product-moment correlations and polychoric correlations. The Two-Factor Model, the Adapted Two-Factor Model, and the Subtraits Model were fitted; see Figure 3.2. Model fit was evaluated using the χ^2-statistic (Bollen, 1989, pp. 263-269) and RMSEA, TLI, and CFI. For CFI, we compared the fit of the models relative to an independence model. Factor correlations greater than 0.4 tentatively suggested a higher-order factor representing Type D.

Mokken Scale Analysis

We used the R package *mokken* (Van der Ark, 2007) for MSA, including AISP and GA, and a procedure to investigate manifest monotonicity. To investigate the the internal structure of the DS14, AISP and GA were run for $c = 0.00, 0.05, \ldots 0.80$ (maximum $c = 0.80$ rather than $c = 0.55$ so as not to miss the hierarchical data structure). Manifest monotonicity was investigated separately for the NA-scale and the SI-scale. Local decreases of item response functions were tested for significance.

3.5 Results

3.5.1 Exploratory Factor Analysis

EFA produced the same internal-structure results for both kinds of correlations; hence, we report results for product-moment correlations. Figure 3.3 shows a line connecting squares, which is the scree plot for the real data, and a line connecting circles, which is the scree plot for the average of the 1,000 random data sets. Two factors lie above the straight line. Oblimin rotation of the two-factor solution yielded factors that correlated 0.38. The factor loadings (Table 3.3, EFA heading) suggested the factors could be interpreted as NA factor (Cronbach's alpha = 0.86) and SI factor (Cronbach's alpha = 0.87). Consistent with Svansdottir et al. (2011), cross-loadings were absent.

3.5.2 Confirmatory Factor Analysis

Again, we report product-moment correlation results. Table 3.3 (CFA columns) shows the factor structure and the model-fit indices for the three estimated models, and Table 3.4 shows the correlations between the error terms. The three model-fit indices suggested that the Two-Factor Model did not fit well. The Adapted Two-Factor Model and the Subtraits Model showed acceptable fit. For the latter model the RMSEA was smaller and the TLI and CFI-values were larger (Table 3.3); hence, the Subtraits Model fitted best. The correlation between the factors was .51, tentatively suggesting evidence for a higher-order factor. The Adapted Two-Factor Model was defined by allowing three cross-loadings and three correlated error terms (Figure 3.2) but only three modifications improved the model fit. They were the cross-loading of item S1 on the NA factor (Table 3.3), the correlated error terms between item S4 and item

Table 3.3: Factor Loadings and Fit Indices for EFA and Three CFA Models.

				EFA		CFA					
						Two-Factor		Adapted Two-Factor		Subtraits	
	Low-level	Item	Item			Model		Model		Model	
Scale	Scale	Position	Label	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
NA	Dysphoria	4	N1	0.67	0.00	0.68	0	0.69	0	0.64	0
	Dysphoria	7	N2	0.77	0.05	0.80	0	0.78	0.04	0.79	0
	Dysphoria	13	N3	0.83	-0.03	0.84	0	0.84	0	0.81	0
	Anx Appr	2	N4	0.58	-0.05	0.54	0	0.64	0	0.52	0
	Anx Appr	12	N5	0.73	0.00	0.73	0	0.73	0	0.73	0
	Irritability	5	N6	0.61	0.03	0.69	0	0.59	0	0.58	0
	Irritability	9	N7	0.60	0.09	0.64	0	0.64	0	0.63	0
Cronbach's α				0.86							
SI	Discomfort	6	S1	0.28	0.54	0	0.75	0.22	0.59	0	0.73
	Discomfort	8	S2	0.06	0.73	0	0.77	0	0.80	0	0.77
	Discomfort	14	S3	0.12	0.68	0	0.75	0	0.77	0	0.77
	Reticence	10	S4	0.06	0.67	0	0.69	0	0.64	0	0.64
	Reticence	11	S5	0.13	0.62	0	0.67	0	0.62	0	0.63
	Lack Soc P	1	S6	-0.11	0.80	0	0.71	0	0.69	0	0.67
	Lack Soc P	3	S7	-0.17	0.68	0	0.57	-0.11	0.58	0	0.50
Cronbach's α					0.87						
$r(F_1, F_2)$				0.38		0.46		0.43		0.51	
χ^2						2118.19		966.38		795.80	
df						76		70		66	
RMSEA						0.092		0.063		0.058	
TLI						0.876		0.941		0.951	
CFI						0.897		0.955		0.963	

Note: RMSEA is root mean square error of approximation, CFI is comparative fit index, and TLI is Tucker-Lewis index. Item Position is the item number in the DS14.

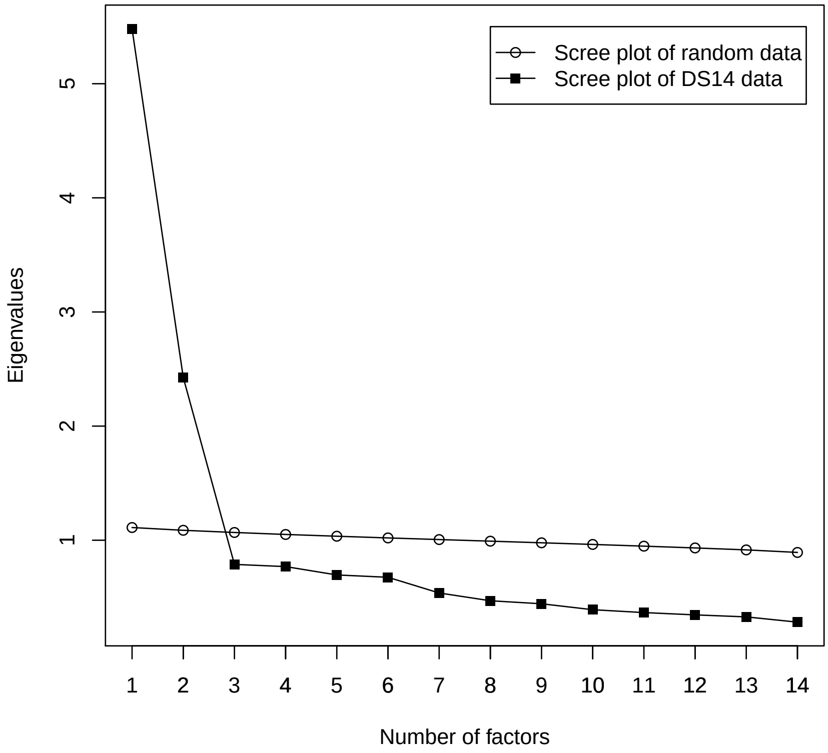


Figure 3.3: Results from parallel analysis for determining the number of factors representing the items of the DS14.

S5, and correlated error terms between item S6 and item S7 (Table 3.4). The other three modifications affected the model-fit indices (not tabulated) only marginally (changes were smaller than 0.005).

The Subtraits Model incorporated correlated error terms between items that together defined a low-level trait (Table 3.4). Three correlations (between items N1 and N2, N2 and N3, and S1 and S2) were so small that fixing them at zero left the values of the fit indices unchanged. Furthermore, the correlations (< 0.20) between the error terms of items S1 and S3, and S2 and S3 were too small to be meaningful. The remaining five error terms (items N1 and N3, N4 and N5, N6 and N7, S4 and S5, and S6 and S7) correlated high enough to considerably improve model fit.

Table 3.4: Error Correlations for Three CFA Models.

Correlation	Two-Factor Model	Adapted Two-Factor Model	Subtraits Model
$r(e_{N1}, e_{N2})$	-	-	0.08
$r(e_{N1}, e_{N3})$	-	-	0.22
$r(e_{N2}, e_{N3})$	-	-	0.10
$r(e_{N4}, e_{N6})$	-	0.11	-
$r(e_{N4}, e_{N5})$	-	-	0.25
$r(e_{N6}, e_{N7})$	-	-	0.22
$r(e_{S1}, e_{S2})$	-	-	0.01
$r(e_{S1}, e_{S3})$	-	-	-0.09
$r(e_{S2}, e_{S3})$	-	-	0.11
$r(e_{S4}, e_{S5})$	-	0.32	0.37
$r(e_{S6}, e_{S7})$	-	0.29	0.41

Note: A hyphen means that a correlated error term was fixed to zero. Not all correlations between error terms are shown. The correlations that are not in the table were fixed to zero.

3.5.3 Mokken Scale Analysis

All items satisfied the monotonicity assumption. Table 3.5 shows AISP and GA results for c values that produced a change in the composition of the scales but not for other c values. For low c values until $c = 0.30$ almost all items were assigned to one Type-D scale. At $c = 0.40$ the items were separated into two scales identifiable as NA-scale and SI-scale. For both scales, total-scale H coefficients were 0.51. As c further increased, the two scales scattered into several smaller scales that were each consistent with a low-level trait from the Subtraits Model. For the NA-scale, AISP and GA identified the low-level scale irritability (AISP at $c = 0.50$; GA at $c = 0.55$). GA found the subscales interpretable as anxious apprehension at $c = 0.50$, irritability at $c = 0.55$, and dysphoria at $c = 0.60$. The SI-scale scattered into subscales at higher values of c. AISP found the subscales interpretable as reticence at $c = 0.50$, lack of social poise at $c = 0.60$, and items S2 and S3 of discomfort in social situations at $c = 0.60$. GA produced all three low-level scales at $c = 0.55$.

Table 3.5: AISP and GA Results for Increasing *c*-Values.

				AISP						GA					
	Low-level	Item	Item	Lower bound						Lower bound					
Subscale	Scale	Position	Label	0.30	0.35	0.40	0.50	0.55	0.60	0.30	0.35	0.40	0.50	0.55	0.60
NA	Dysphoria	4	N1	1	1	1	1	1	0	1	1	1	0	1	1
	Dysphoria	7	N2	1	1	1	1	1	1	1	1	1	1	1	1
	Dysphoria	13	N3	1	1	1	1	1	1	1	1	1	1	1	1
	Anx Appr	2	N4	1	1	1	0	0	0	1	2	1	3	0	0
	Anx Appr	12	N5	1	1	1	1	1	1	1	1	1	3	1	0
	Irritability	5	N6	1	1	1	4	4	0	1	2	1	1	4	0
	Irritability	9	N7	1	1	1	4	4	0	1	1	1	1	4	0
SI	Discomfort	6	S1	1	1	2	2	0	0	1	1	2	2	2	0
	Discomfort	8	S2	1	2	2	2	2	3	1	1	2	2	2	3
	Discomfort	14	S3	1	2	2	2	2	3	1	1	2	2	2	3
	Reticence	10	S4	1	2	2	3	3	4	1	1	2	2	3	4
	Reticence	11	S5	1	2	2	3	3	4	1	1	2	2	3	4
	Lack Soc P	1	S6	1	2	2	2	2	2	1	1	2	2	5	2
	Lack Soc P	3	S7	0	2	2	0	0	2	0	0	2	0	5	2

Note: For each column, items with the same number indicate that these items are in the same scale; 1 means that the items are in scale 1, 2 means that the items are in scale 2, and so on. 0 indicates that the item is unscalable.

3.6 Discussion

Table 3.6 summarizes the results EFA, CFA, and MSA produced. CFA and MSA were more sensitive to the low-level scales than EFA. CFA showed an acceptable fit for the Adapted Two-Factor Model and the Subtraits Model, but for both models several effects were small. The Adapted Two-Factor Model is data driven and does not contribute to understanding the low-level traits in the DS14, whereas the Subtraits Model allows for the investigation of the proposed low-level scales. The good fit of the Subtraits Model supports the existence of the three-level structure. However, the hierarchical structure of the DS14 suggests that the error terms in the Subtraits Model are positively correlated, but the results in Table 3.4 showed that for low-level scales dysphoria and discomfort in social situations some correlated error terms were almost zero. This discrepancy suggests that the CFA results do not fully support the existence of dysphoria and discomfort in social situations. MSA provided additional evidence for the Subtraits Model. For default value $c = 0.3$, MSA produced one scale including 13 of the 14 items. For $c = 0.4$, MSA produced the strong NA and SI scales. For higher c, the scales scattered into smaller scales. Each of the six low-level traits was represented by one of the smaller scales found using higher c values.

Each method has its own strengths and is sensitive to different aspects of the internal structure of the DS14. The method versatility supported the Subtraits Model. The results of this study justify the use of the DS14 for assessment at three levels. At the high level, the DS14 assesses Type D as a psychological risk factor for morbidity and mortality in patients suffering from cardiovascular disease. At the medium level, researchers may use the questionnaire to assess NA and SI and to investigate how these traits interact to increase the risk for morbidity and mortality. At the low level, the traits NA and SI may be further scrutinized into the low-level traits. Researchers can investigate whether some of these traits affect the morbidity and mortality in cardiovascular patients more severely than others.

MSA and factor analysis may find different results with respect to the internal structure of the DS14 in other countries and clinical populations. For future research of the internal structure of the DS14 in different populations, we advise to use CFA for investigating the fit of the Subtraits Model and to use MSA

Table 3.6: Comparison of the Low-Level Structure From the Three Methods.

Subscale	Low-level Scale	Item Position	Item Label	EFA	CFA	MSA
NA	Dysphoria	4	N1	-	1	1
	Dysphoria	7	N2	-	-	1
	Dysphoria	13	N3	-	1	1
	Anx Appr	2	N4	-	2	2
	Anx Appr	12	N5	-	2	2
	Irritability	5	N6	-	3	3
	Irritability	9	N7	-	3	3
SI	Discomfort	6	S1	-	-	4
	Discomfort	8	S2	-	-	4
	Discomfort	14	S3	-	-	4
	Reticence	10	S4	-	4	5
	Reticence	11	S5	-	4	5
	Lack Soc P	1	S6	-	5	6
	Lack Soc P	3	S7	-	5	6

Note: Anx Appr = Anxious apprehension; Lack Soc P = Lack of social poise. For each method: Items having the same digit were found to be in the same low-level cluster. suggesting that the items represent a single low-level trait. A hyphen means that the item was not included in the low-level structure.

because (a) CFA was not conclusive about the choice between the Adapted Two-Factor Model and the Subtraits Model and (b) CFA did not identify all low-level scales.

The debate whether or not Type D is a single personality trait or originates from the interaction between NA and SI requires additional research. Our study supports the three-level model of distressed personality including Type D as the model's high-level trait. To reach a conclusive verdict about Type D requires the further development of the theory of Type D. This entails studying the cognitive and affective processes typical of the distressed personality, not only through correlational studies but also by means of experimentation including psychological and biological variables, and the way these processes affect morbidity and mortality in patients suffering from cardiovascular disease. A well-founded theory provides more explanatory power for relationships that are found to exist between Type D and relevant outcome variables such as life-expectancy and quality-of-life variables. We believe thus far Type D research has relied too much on correlational studies and has neglected theory development.

Chapter 4

Methodological Artifacts in Dimensionality Assessment of the Hospital Anxiety and Depression Scale (HADS)[*] [†]

Abstract

The Hospital Anxiety and Depression Scale (HADS) is a brief, self-administered questionnaire for the assessment of anxiety and depression in hospital patients. A recent review discussed the disagreement among different studies with respect to the dimensionality structure of the HADS, and concluded that the HADS must be abandoned. Our study argues that this disagreement is mainly due to a methodological artifact, and that the HADS needs revision rather than abandonment. We used Mokken scale analysis (MSA) to investigate the dimensionality structure of the 14 HADS items in a representative sample from the Dutch non-clinical population ($N = 3,643$) and compared the dimensionality structure to results Emons, Sijtsma, and Pedersen (2012) obtained in a Dutch cardiac-patients sample. We demonstrated how MSA can retrieve either one scale, two subscales, or three subscales, and that the result depends on the data structure but also on choices the researcher makes. Two 5-item scales for anxiety and depression seemed adequate. Four HADS items constituted a weak scale and contributed little to reliable measurement. MSA supported a 2-level hierarchical structure for ten HADS items, and suggested that four items should be discarded. At the first level, MSA suggested that ten items constitute one psychological distress scale; and at the second level MSA suggested an anxiety subscale (5 items) and a depression subscale (5 items). We argued that several psychometric methods only show one level of a hierarchical structure and that users of psychometric methods are often unaware of this phenomenon and miss information about other levels. In addition, we argued that a theory about the attribute may guide the researcher but that well-tested theories are often absent.

[*] This chapter has been submitted for publication.

[†] Acknowledgements: We thank Susanne Pedersen for providing the data.

4.1 Introduction

The Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983; Caci, Bayle, Dossios, Robert, & Boyer, 2003) is a brief, self-administered questionnaire for the assessment of the presence and the severity of anxiety and depression in physically ill patients. The HADS consists of two 7-item scales, one measuring anxiety and the other depression (Table 4.1). Somatic indicators of anxiety and depression are not part of the HADS because physical illness may interfere with somatic symptoms (Moorey et al., 1991). For the classification of individuals as anxious or depressed, researchers use the total scores on the 7-item Anxiety and Depression scales (Brennan, Worrall-Davies, McMillan, Gilbody, & House, 2010).

Table 4.1: Item Labels and Item Contents of the Hospital Anxiety and Depression Scale.

Item label	Item content
A1	I feel tense or wound up
A2	I get a sort of frightened feeling as if something awful is about to happen
A3	Worrying thoughts go through my mind
A4	I can sit at ease and feel relaxed
A5	I get a sort of frightened feeling like butterflies in my stomach
A6	I feel restless as if I have to be on the move
A7	I get sudden feelings of panic
D1	I still enjoy the things I used to enjoy
D2	I can laugh and see the sunny side of things
D3	I feel cheerful
D4	I feel as if I am slowed down
D5	I have lost interest in my appearance
D6	I look forward with enjoyment to things
D7	I can enjoy a good book or radio or TV program

Two literature reviews (Bjelland, Dahl, Tangen Haug, & Neckelmann, 2002; Herrmann, 1997) considered the HADS to be a psychometrically sound, 2-dimensional questionnaire for measuring anxiety and depression. More recently, Cosco, Doyle, Ward, and McGee (2012) found that many studies failed to replicate the HADS' expected 2-dimensional structure and, moreover, disagreed with the dimensionality structure of the HADS. Coyne and Van Sonderen (2012) concluded from this result that the HADS should be abandoned.

We discern three methodological reasons that help to better understand why so much disagreement exists with respect to the dimensionality structure of the HADS. First, different psychometric methods that are used to investigate the dimensionality structure of a set of items may produce different results. The reason is that different methods provide different perspectives on the data structure, select and amplify different aspects of the data structure, and produce different dimensionality structures. Second, the use of a particular method requires the user to make particular choices, and different choices may produce different dimensionality results. Third, in addition to a method effect, due to different psychological processes different populations and samples drawn from the populations may produce different dimensionality results (Cosco, Doyle, Ward, et al., 2012). We provide examples of each effect.

Examples of different methods producing different results for the HADS, are Rasch-model analysis (Rasch, 1960), which predominantly produced a 14-item psychological-distress scale (Gibbons et al., 2011; Pallant & Tennant, 2007); exploratory factor analysis, which usually confirmed the expected 2-dimensional structure; and confirmatory factor analysis that was used to test the expected 2-dimensional HADS structure, but often found support for a 3-dimensional structure, while different studies assigned different items to different factors (Caci et al., 2003; Dunbar, Ford, Hunt, & Der, 2000; Friedman, Samuelian, Lancrenon, Even, & Chiarelli, 2001).

Examples of subjective choices researchers have to make when they use one particular method are: For Rasch-model analysis, the goodness-of fit tests one uses to assess the fit of the model to the data (Molenaar, 1983; Glas & Verhelst, 1995); for exploratory factor analysis, the rotation method and the number of factors to be rotated (Reise, Waller, & Comrey, 2000); and for confirmatory factor analysis, the goodness-of-fit indices to be used and the model modifications to be executed based on statistical modification indices (Hu & Bentler, 1999).

Examples of population effect and sample effect are: A 2-dimensional structure that fitted best for non-clinical subjects, a 3-dimensional structure that fitted best for cardiac patients, and a varying dimensionality structure ranging from one to four dimensions in different cancer patient groups (Cosco, Doyle, Ward, et al., 2012).

Cosco, Doyle, Ward, et al. (2012) recommended using Mokken scale analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002) to study the HADS dimensionality structure. MSA is a scaling method that can be used for the assessment of Likert-items (Emons, Sijtsma & Pedersen, 2012; Straat, Van der Ark, & Sijtsma, 2012b; Wismeijer, Sijtsma, Van Assen, & Vingerhoets, 2008). MSA is a more flexible dimensionality assessment method than Rasch-model analysis, exploratory factor analysis, and confirmatory factor analysis. In a sample of Dutch cardiac patients, Emons et al. (2012) used MSA to study the dimensionality structure of the HADS. The authors used MSA in combination with exploratory factor analysis and confirmatory factor analysis, and found support for the Caci et al. (2003) 3-factor model in which items $A1$, $A2$, $A3$, $A5$, and $A7$ constitute a 5-item Anxiety scale, items $D1$, $D2$, $D3$, $D4$, and $D6$ a 5-item Depression scale, and items $A4$, $A6$, and $D7$ a 3-item Restlessness scale; Item $D5$ was unscalable. Based on MSA in a sample of Irish cardiac patients, Cosco, Doyle, Watson, Ward, and McGee (2012) concluded that one dimension best described the HADS dimensionality structure.

Coyne and Van Sonderen (2012) noticed that the highly varying results for the HADS dimensionality structure imply that the HADS is not dependable for the assessment of anxiety and depression in hospital patients. They concluded that the HADS must be abandoned in favor of instruments with a clearer dimensionality structure. In this study, we used MSA to provide evidence that the different dimensionality-structure results for the HADS probably are a methodological artifact, which may be explained from the HADS' hierarchical structure. Moreover, we identified four items having low measurement quality that may be removed from the HADS. The resulting ten HADS items constitute a strong basis for a revision of the questionnaire. Because previous studies (e.g., Andrea et al., 2004; Hunt-Shanks, Blanchard, Reid, Fortier, & Cappelli, 2010; Martin, Thompson, & Barth, 2008; Mykletun, Stordal, & Dahl, 2001) found different dimensionality structures in samples from a non-clinical population and a cardiac-patients population, we compared MSA results for samples from both populations so as to explain why studies investigating different populations produce different dimensionality-structure results.

This chapter contains the following information. First, we discuss the HADS' hierarchical dimensionality structure, which is also frequently found with other attributes (e.g., Straat et al., 2012b). Second, we use MSA to study the HADS' hierarchical structure in a sample from a non-clinical population and compare the results to MSA results Emons et al. (2012) obtained from a sample of cardiac patients. Third, we discuss the relation of the MSA results to previous findings from Rasch-model analysis, exploratory factor analysis, and confirmatory factor analysis. Finally, we discuss the consequences of the MSA results for the use of the HADS.

4.2 Hierarchical Structure of Psychological Attributes

Psychological attributes often have a hierarchical structure (Reise, Waller, & Comrey, 2000). In response to Coyne and Van Sonderen (2012), Norton, Sacker, and Done (2012) also made this point, based on the argument that researchers using the same dimensionality-assessment method usually found the same dimensionality structures for the HADS but researchers using different methods found different dimensionality structures. Their point thus is that different methods find different levels of the hierarchy but that the hierarchy does not become apparent when one does not use different methods or when one uses one method but fails to implement different modes of using the method. An example is the following: We assume that a hierarchical attribute structure is reflected in the structure of the item scores that constitute the data. A plausible structure would be that all items correlate positively but that several clusters contain items that correlate higher with one another than with the items from other clusters. This structure could suggest two levels, one on which the common denominator of all items is described and another on which the the different item clusters are identified. If clusters contain sub-clusters of items that share variance that other items do not share, one might even discern a third level.

How do different psychometric methods deal with the data structure just described? Rasch-model analysis has the Rasch model, which is a unidimensional scaling model, as the criterion for assessing the structure of the items. Given the formal prevalence for unidimensionality, a Rasch-model analysis tends to provide information on which items to retain in the scale and which items to remove but the end result tends to be one scale and one or more items that are not in the scale. The method thus tends to identify only the first level of the hierarchy. Meijer, Sijtsma, and Smid (1990) and De Koning, Sijtsma, and Hamers (2002) provided rather complex methodologies for identifying data multidimensionality using the Rasch-model analysis. Exploratory factor analysis is a typical dimensionality-reduction method that focuses on identifying a number of dimensions that each attract a number of distinct items and explain a reasonable amount of variance in the item scores. Unlike the Rasch-model analysis, exploratory factor analysis thus tends to identify the second level of the hierarchy. Like the Rasch-model analysis, confirmatory factor analysis is a confirmatory method and the researcher defines the dimensionality structure that serves as the hypothesis to be tested. The identification of the dimensionality structure that best reproduces the inter-item correlations yields the best model-data fit and this may entail preference for the third level in our example.

MSA is particularly useful to evaluate the different levels of the hierarchical structure. The researcher has to specify a numerical scaling criterion that controls the level at which the hierarchical structure is assessed. Usually, researchers rely on a default option that computer programs provide but Sijtsma and Molenaar (2002, chap. 5) recommend to try a range of criterion values. In the example, the lowest criterion values would produce one scale that includes most or all items, higher criterion values would produce the second-level subscales, and still higher criterion values would produce small sub-subscales whereas many items would not be included in scales anymore. Even higher criterion values would lead to the conclusion that the item set is unscalable. Thus, dimensionality assessment methods such as the Rasch-model method and exploratory and confirmatory factor analysis may find different dimensionality structures, but MSA may well find the hierarchy that the other methods miss. The hierarchical structure implies that there is not one “true” dimensionality but that the dimensionality depends on the level at which the hierarchy is assessed.

This was just one example, but different sets of outcomes are possible, depending on the structure of the attribute. For example, three cluster of items may exist such that within clusters inter-item correlations have approximately the same magnitude and between clusters inter-item correlations are zero. Rasch-model analysis will produce gross misfit and suggest to reject two thirds of the items and retain one short scale representing one cluster by approximation. Exploratory factor analysis will likely find the 3-cluster structure. Confirmatory factor analysis will likely produce that the three-factor solution with the items loading on the appropriate factors is the best-fitting model. For increasing criterion values MSA will continuously find the correct solution until suddenly all items appear unscalable.

4.3 Method

4.3.1 Participants

For the non-clinical sample, 3,708 Dutch participants were approached and 3643 (98.2 %) participants filled out the HADS. Two gender levels and six age levels (20-29, ..., 60-69, 70-80 years) served as stratification criteria, and quota sampling produced twelve equally sized groups. A local ethics committee at Tilburg University (protocol number: 2006/1101) approved this study. Research assistants approached participants personally or by phone. After having been explained the study's purpose, participants received an informed consent form and a questionnaire, and participants returned both documents in closed envelopes to the research assistants (between October 1, 2006—December 15, 2008). Returned questionnaires were coded by number for purposes of data collection tracking but were otherwise anonymous.

The sample consisted of 50% men. The mean age was equal to 50.12 and the standard deviation was equal to 16.31. For 68 respondents, one to three item scores were missing. Two-way imputation (Bernaards & Sijtsma, 2000; Van Ginkel, Van der Ark, & Sijtsma, 2007) was used to replace missing item scores by estimated scores.

4.3.2 Statistical Analyses

Mokken Scale Analysis

MSA assesses the fit of a measurement model known as the monotone homogeneity model (Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Meijer, 2007). First, the monotone homogeneity model assumes a single attribute, such as anxiety or depression, to capture the associations between the item scores; that is, the items do not measure any other attribute in common. Second, the monotone homogeneity model assumes a monotone nondecreasing relation between the scores on an item and the attribute. The first assumption ensures that the items measure only one attribute rather than a conglomerate of attributes that hinders a straightforward interpretation of test performance. The second assumption reflects the idea that the higher one scores on the attribute scale, the higher one is expected to score on each of the items in the test that are indicators of the attribute. Mokken (1971, chap. 4) and Sijtsma and Molenaar (2002, chaps. 2-5) provide technical details about the monotone homogeneity model. Let the total score be defined as the sum of the J item scores in the test. Then, if the data are consistent with the monotone homogeneity model, individuals with a higher total score are expected to also score higher on the attribute (Grayson, 1988; Van der Ark, 2005). Hence, a monotone homogeneity model that fits the anxiety-item data provides a justification for the use of the total score as a measure of anxiety; likewise for depression.

MSA uses item scalability coefficient H_j, which expresses the strength of the relation between the scores on item j and the attribute the total score measures (Van Abswoude, Van der Ark, & Sijtsma, 2004). A high H_j value implies that the item distinguishes well between low scores on the attribute and high scores on the attribute. Given the monotone homogeneity model, it can be shown that $0 \leq H_j \leq 1$. MSA aims at obtaining scales consisting of items with H_j values exceeding a lower bound c (default $c = .3$). The researcher can specify lower bound c to reflect the minimum required strength of the relation of an item with the attribute for the item to be admitted to the scale. Items for which $H_j < c$ are not admitted to the scale. The total-scale coefficient H reflects the discrimination power of the total scale. Mokken, Lewis, and Sijtsma (1986) suggested that H expresses the accuracy of a person ordering by means of the total score. Mokken (1971, pp. 148-153) suggested that $.30 \leq H < .40$ defines a weak scale, $.40 \leq H < .50$ a medium scale, and $H \geq 0.50$ a strong scale; $H < .3$ means that items are unscalable.

An automated item selection procedure (Straat, Van der Ark, & Sijtsma, in press) that is part of MSA partitions a set of items, such as the 14 HADS items, into one or more scales if the data permit. The two requirements for a scale are that (1) all inter-item correlations are positive and (2) each H_j value exceeds lower bound c; that is, $H_j \geq c$ (Mokken, 1971, p. 184; Sijtsma & Molenaar, 2002, p. 68). The automated item selection procedure (Mokken, 1971; pp. 190-193) is a bottom-up algorithm that starts with the two items i and j that have the highest, significantly positive H_{ij} value that exceeds lower bound c. In each consecutive step, the procedure adds one item that correlates positively with the already selected items, which has an H_j value that exceeds c, and that produces the highest H value with the items already selected in the previous steps, given all possible items that are candidates for selection in the present step. The item selection proceeds until there are no items left that satisfy the requirements for inclusion in the scale. If items remain unselected, from these items the procedure may select a second scale, a third scale, and so on, until there are no items left or the items left are unscalable.

Sometimes the procedure selects an item that after completion of the procedure does not satisfy the scale requirements anymore due to the items selected later in the procedure. Another problem is that the procedure does not always find the best possible partitioning. The first problem is circumvented and the second problem is almost always circumvented by the use of a genetic algorithm (Straat et al., in press) that obtains only partitionings that satisfy the scale requirements.

Lower bounds c serve as the criterion values that can be varied to study different levels of the hierarchical structure of a psychological attribute. We used the methodology that Hemker, Sijtsma, and Molenaar (1995; also, see Sijtsma & Molenaar, 2002, chap. 5) recommended to find different dimensionality structures, and that entails running the automated item selection procedure several times, starting with minimum $c = 0$, in each next run using a lower bound c that has increased by 0.05, and terminating with $c = 0.60$ or higher. We used R package mokken (Van der Ark, 2007) to run the automated item selection procedure and the genetic-algorithm version. To investigate the dimensionality of the HADS, both item selection procedures were run for $c = 0.00, 0.05, \ldots, 0.60$.

Item scalability coefficient H_j expresses the strength of the relationship of the item and the attribute but does not provide information on whether the relation between item j and the attribute measured by the total score on the items except item j, also called the rest score, is nondecreasing (including item j in the total score would produce an artifact; Junker & Sijtsma, 2000). The relationship between the item score and the rest score is locally decreasing if an increase of the rest score produces a decrease of the expected item score along a small range of rest scores. This decrease violates the monotonicity assumption of the monotone homogeneity model. We investigated for each item j whether the mean item score is a nondecreasing function of the rest score (Junker & Sijtsma, 2000). We used the R package mokken (Van der Ark, 2007) to investigate the monotonicity assumption for a single 14-item HADS scale, the 7-item Anxiety and Depression scales, and the three scales of the Caci et al. (2003) model.

Reliability

We used four methods to estimate the total-score reliability (Sijtsma, 2009). They were coefficient α, coefficient λ_2 (both computed using R package mokken; Van der Ark, 2007), the greatest lower bound to the reliability (glb; computed using R package psych; Revelle, 2012), and the Molenaar-Sijtsma method (Van der Ark, Van der Palm, & Sijtsma, 2011; computed using R package mokken; Van der Ark, 2007). Methods α, λ_2, and glb are lower bounds to the total-score reliability. Their mutual relationship is: $\alpha \leq \lambda_2 \leq$ glb. Coefficient α is the most frequently used estimate, but λ_2 and glb provide estimates closer to the population total-score reliability and may be preferred over α (Sijtsma, 2009). The Molenaar-Sijtsma method was developed in the context of MSA and is a reliability estimator with smaller bias than α and λ_2 (Sijtsma & Molenaar, 2002, p. 110; Van der Ark et al., 2011). We computed the reliability estimates for a unidimensional scale containing all 14 HADS items, the 7-item Anxiety and Depression scales, and the three scales of the Caci et al. (2003) model.

4.4 Results

4.4.1 Mokken Scale Analysis

For each lower bound c, the automated item selection procedure and its genetic-algorithm version yielded the same item partitionings. Table 4.2 shows the item partitionings for lower bounds c equal to $0, .3$, and .45; other c-values did not provide additional information. For $c = 0$, all items were selected in one scale. At lower bound $c = .3$, the automated item selection procedure produced a single scale containing 11 items. Items $A6$, $D5$, and $D7$ were unscalable. For higher c values, the automated item selection procedure found two distinct scales that resembled the shortened Anxiety and Depression scales of the Caci et al. (2003) 3-factor model. Based on this result, we used confirmatory MSA and computed the H_j and the H coefficients for the a priori identified Anxiety, Depression, and Restlessness scales (Table 4.2). Given that two out of three restlessness items had $H_j < .3$, the restlessness items were unscalable; hence, here only a 5-item Anxiety scale and a 5-item Depression scale were obtained. In none of the investigated scales – the 14-item scale, the 7-item Anxiety and Depression scales, and the three Caci et al. (2003) scales – did we find violations of monotonicity.

4.4.2 Reliability

Table 4.3 shows the four reliability estimates for the models with one 14-item scale, two 7-item scales measuring anxiety and depression, and the three scales based on the Caci et al. (2003) model. The 14-item scale had the highest reliability, and the 5-item Anxiety and Depression scales had higher α, λ_2, and MS than the 7-item Anxiety and Depression scales. The glb of the 5-item Anxiety scale was approximately equal to the glb of the 7-item Anxiety scale and the glb for the 5-item Depression scale was higher than the glb for the 7-item Depression scale. Hence, the reliability estimates suggest that items $A4$ and $A6$ do not contribute to the reliable measurement of anxiety and items $D5$ and $D7$ do not contribute to the reliable measurement of depression.

Table 4.2: Results from Exploratory and Confirmatory Mokken Scale Analysis in the Non-Clinical Sample.

	Exploratory MSA				Confirmatory MSA		
	$c = 0$	$c = .3$	$c = .45$				
Item label			Scale 1	Scale 2	Anx	Depr	Restl
A1	.37	.42	.51		.51		
A2	.35	.40	.52		.52		
A3	.37	.44	.56		.56		
A4	.33	.38					.38
A5	.34	.34	.48		.48		
A6	.26						.29
A7	.34	.38	.49		.49		
D1	.34	.38		.50		.49	
D2	.38	.43		.52		.52	
D3	.36	.40		.46		.44	
D4	.39	.43		.49		.47	
D5	.17						
D6	.30	.32				.45	
D7	.23						.29
H	.32	.39	.51	.49	.51	.47	.32

Note: Anx = Anxiety scale, Depr = Depression scale, and Restl = Restlessness scale.

4.4.3 Comparing the Non-Clinical and Cardiac-Patients Populations

Table 4.4 shows the exploratory and confirmatory MSA results that Emons et al. (2012) obtained. For exploratory MSA, the scales in the non-clinical sample (Table 4.2) and the cardiac-patients sample (Table 4.4) were comparable but for the cardiac-patients sample, the dimensionality structure remained intact for higher values of lower bound c. For confirmatory MSA, the H_j and H coefficients were also higher in the cardiac-patients sample than in the non-clinical sample. As a result, the Anxiety scale and the Depression scale are stronger scales in the cardiac-patients sample than in the non-clinical sample, and the Restlessness scale satisfied the Mokken scale criteria at the default lower bound of .3 in the cardiac-patients sample but not in the non-clinical sample.

Table 4.3: Coefficients α, λ_2, the glb, and the Molenaar-Sijtsma Method for One Scale, Two Scales, and Three Scales (Non-Clinical Sample).

	α	λ_2	glb	MS
One scale	.832	.836	.873	.840
Two scales				
Anxiety (7-item)	.773	.776	.823	.784
Depression (7-item)	.735	.739	.752	.734
Three scales				
Anxiety (5-item)	.780	.783	.801	.796
Depression (5-item)	.762	.766	.800	.771
Restlessness (3-item)	.560	.565	.634	.559

Note: glb is the greatest lower bound.

4.5 Discussion

Figure 4.1 summarizes the MSA results with respect to the HADS' hierarchical structure and shows the levels of the hierarchical structure that are consistent with previous Rasch-model analysis results and factor analysis results. Since all inter-item correlations were positive, for $c = 0$ all 14 items were selected in one scale. At the next level, ten items constituted a single general psychological distress scale. Items $A4$, $A6$, $D5$, and $D7$ had $H_j < c$, and were not selected. At higher levels, five items constituted an anxiety scale and five items constituted a depression scale. Items $A4$ and $A6$ were excluded from the original 7-item Anxiety scale and items $D5$ and $D7$ were excluded from the original 7-item Depression scale. The reliability of the 5-item anxiety and depression scales was higher than of their 7-item versions. Hence, the reliability estimates confirmed the lower measurement quality of the four items.

Researchers using Rasch-model analysis (Gibbons et al., 2011; Pallant & Tennant, 2007) reported the 14-item general psychological-distress scale that MSA found for lower bound c-values close to 0. Moreover, fit assessment of the Rasch model showed evidence of the anxiety and depression scales, but did not reveal the low measurement quality of the items $A4$, $A6$, $D5$, and $D7$ (Gibbons et al., 2011; Pallant & Tennant, 2007). A problem of the Rasch model is that it assumes that all items relate to the attribute scale to the same degree, and not

Table 4.4: Exploratory and Confirmatory Mokken Scale Analysis Results for a Cardiac-Patients Sample (Adapted from Emons, Sijtsma, & Pedersen, 2012, Tables 1 and 5)

	Exploratory MSA					Confirmatory MSA		
	$c = 0$	$c = .4$		$c = .5$				
Item label		Scale 1	Scale 2	Scale 1	Scale 2	Anx	Depr	Restl
A1	.47	.51		.58		.58		
A2	.41	.45		.62		.62		
A3	.45	.50		.61		.61		
A4	.38		.47					.45
A5	.40	.45		.53		.53		
A6	.28							.38
A7	.43	.48		.62		.62		
D1	.41	.46			.59		.60	
D2	.46	.51			.61		.56	
D3	.44	.50			.57		.53	
D4	.43	.48			.51		.48	
D5	.32						.39	
D6	.39	.43			.58		.54	
D7	.31		.47					.39
H	.39	.48	.47	.59	.57	.59	.51	.40

all goodness-of-fit research may be able to pinpoint this cause of misfit (e.g., Molenaar, 1983; Glas & Verhelst, 1995). Alternatively, one may choose fitting the 2-parameter logistic item response model (Birnbaum, 1968), which assumes that different items relate to the attribute scale to different degrees, and thus may distinguish items relating relatively weakly to the scale from items relating stronger to the scale.

Exploratory factor analysis (e.g., Andrea et al., 2004; Mykletun et al., 2001) yielded a 2-factor solution in which all items with index A loaded on the Anxiety factor and all items with index D loaded on the Depression factor. Hence, exploratory factor analysis included the low-quality items in the factors, but MSA excluded the items from the two scales because for these items $H_j < c$. In confirmatory factor analysis (e.g., Hunt-Shanks et al., 2010; Martin et al., 2008), the fit indices were sensitive to the misfit of the low-quality items

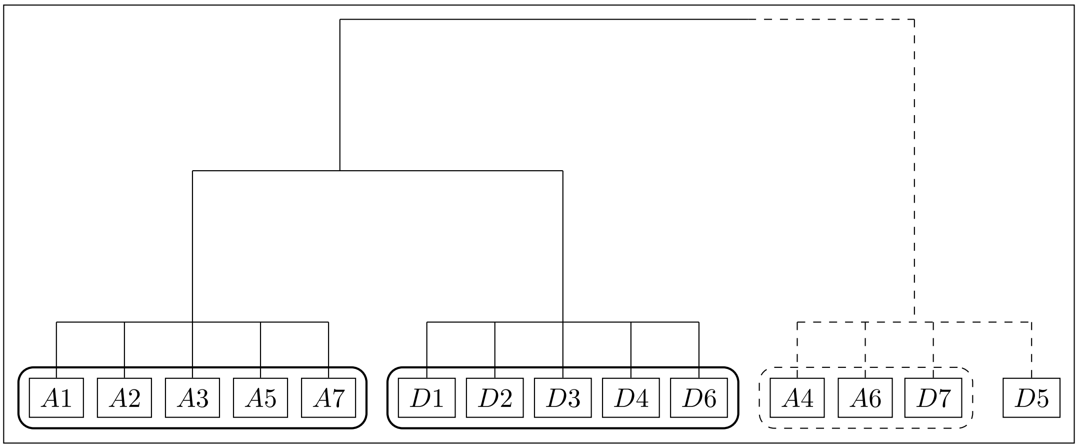


Figure 4.1: Graphical representation of the hierarchical structure of the HADS identified by MSA.

$A4$, $A6$, $D5$, and $D7$ and, as a result, a third factor had to be defined to obtain an acceptably fitting model. Emons et al. (2012) showed that the three item clusters from the Caci et al. (2003) 3-factor model were consistent with the three scales found in a cardiac-patients sample. The main difference between the exploratory and the confirmatory factor analysis results was that confirmatory factor analysis identified the low-quality items by a misfitting two-factor model.

In non-clinical samples, in which researchers (e.g., Mykletun et al., 2001; Andrea et al., 2004) found a 2-dimensional structure and in cardiac-patients samples, researchers (e.g., Hunt-Shanks et al., 2010; Martin et al., 2008) found a 3-dimensional structure. The dimensionality structure of the data was comparable in the non-clinical sample and the cardiac-patients sample. However, we found that the H_j and the H values were lower in the non-clinical sample than in the cardiac-patients sample. In the non-clinical sample, the H_j values of the items constituting the Restlessness scale were lower than 0.3 and, as a result, the items did not satisfy the item-selection criteria. In the cardiac-patients sample, confirmatory Mokken scale analysis identified the Restlessness scale of the Caci et al. (2003) 3-factor model. Hence, MSA confirmed a 2-dimensional structure in the non-clinical sample due to low item scalability, and a 3-dimensional structure in the cardiac-patients sample.

Zigmond and Snaith (1983) did not intend the HADS to measure restlessness in addition to anxiety and depression. Except the four Restlessness items, different dimensionality assessment methods used in different populations produce dimensionality results for the ten items that are consistent. An important question is whether the four Restlessness items cover important aspects of anxiety and depression. This question is difficult to answer. Like many questionnaires, the HADS is not the operationalization of a well-tested theory of anxiety and depression from which substantive arguments for the inclusion or exclusion of particular items were derived. Sijtsma (2012, in press) argued that in the absence of a well-tested theory about the attribute of interest, researchers can only rely on psychometric methods to decide about the dimensionality structure of their item sets. As a result, psychometric rather than theoretical arguments are highly dominant, perhaps too dominant, in instrument construction.

The HADS lacks a well-developed theoretical foundation, and as a result the items used predominantly define anxiety and depression instead of the other way around. A well-established theory about anxiety and depression should guide the operationalization into items that constitute the measurement instrument (Sijtsma, 2012, in press). The absence of a well-established theory and the resulting heavy reliance of researchers on psychometric methods for dimensionality assessment that each emphasize different levels of the hierarchical HADS structure, together explain the disagreement among different studies about the dimensionality structure of the HADS. MSA better reveals the hierarchy in a dimensionality structure than any of the other methods, and also provides a higher level of awareness with respect to the possibility that different dimensionality structures can be part of the same hierarchy. The HADS can have a future but needs to be based on better established and tested anxiety and depression theories. The two 5-item subscales can be an excellent basis for a novel HADS whereas the four Restlessness items may be discarded.

Chapter 5

Minimum Sample Size Requirements for Mokken Scale Analysis*

Abstract

An automated item selection procedure in Mokken scale analysis partitions a set of items into one or more Mokken scales, if possible. Two algorithms are available that pursue the same goal of selecting Mokken scales of maximum length: Mokken's original automated item selection procedure (AISP) and a genetic algorithm (GA). Minimum sample size requirements for Mokken scale analysis have not yet been established. In practical scale construction reported in the literature, we found that researchers used sample sizes ranging from 133 to 15,022 respondents. We investigated the effect of sample size on the assignment of items to the correct scales. Using a misclassification of 5% as a criterion, we found that Mokken scale analysis minimally required 250 to 500 respondents when item quality was high and 1250 to 1750 respondents when item quality was low.

*This chapter has been submitted for publication.

5.1 Introduction

For Mokken scale analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002; Van Schuur, 2011), minimum sample size requirements to obtain stable item selection results are unknown. Researchers use an automated item selection method to partition a set of items into one or more scales, if possible, of maximum length. A literature search of recent applications of MSA revealed that sample sizes for MSA ranged from 133 (Adler & Brodin, 2011) to 15,022 respondents (Prince et al., 2010). For $N = 15,022$, sample fluctuations are probably negligible, but for $N = 133$ sample fluctuations may be considerable. Researchers have a limited amount of time and finances to collect data (Hedeker, Gibbons, & Waternaux, 1999), but they also wish to replicate their findings in future studies (Jackson, 2003) and to have adequate statistical power for finding the effects they are interested in (Hedeker et al., 1999). In this study, we investigated minimum sample size requirements for two item selection methods in MSA that pursue the same goal using different algorithms.

Many studies investigated the minimally required sample size for other statistical methods such as regression analysis (e.g., Cohen, 1988; Green, 1991), factor analysis (e.g., Guadagnoli & Velicer, 1988; MacCallum, Widaman, Preacher, & Hong , 2002; Mundfrom, Shaw, & Ke, 2005; Velicer & Fava, 1998), multilevel analysis (e.g., Cohen, 2005; Hedeker et al., 1999; Snijders & Bosker, 1993), structural equation modeling (e.g., Bentler & Yuan, 1999; Jackson, 2003), and item response theory (e.g., Chuah, Drasgow, & Leucht, 2006; Hambleton & Jones, 1994; Hulin, Lissak, & Drasgow, 1982; Reise & Yu, 1990), but not for MSA. These studies investigated the sample size that is minimally required to have unbiased and precise parameter estimates. For MSA, parameter estimation is not of main interest, but researchers wish to know whether items are correctly partitioned into scales. To evaluate the similarity of two partitionings, previous studies used the indices Per Element Accuracy (PEA; Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005), and the minimum number of items to be moved to another scale for two partitionings to be equal (MIN; Van der Ark & Sijtsma, 2005). PEA and MIN suggest the extent to which items are assigned to the correct scales.

This paper is organized as follows. First, we discuss the monotone homogeneity model (Mokken, 1971; Sijtsma & Molenaar, 2002). Second, we discuss two automated item selection methods for MSA. Third, we study the minimally required sample size to find the correct partitioning of items. Fourth, we give recommendations to researchers about minimum sample sizes required for item selection in MSA.

5.2 Monotone Homogeneity Model

The monotone homogeneity model (MHM; Mokken, 1971, chap. 4; Sijtsma & Molenaar, 2002; Van Schuur, 2011) is defined by three assumptions: The latent variable θ is *unidimensional*, the J item score variables X_j ($j = 1, \ldots, J$) are *locally independent* given θ, and each expected item score is a *monotone nondecreasing* function of θ. These functions are called item response functions. Grayson (1988) proved that for a set of dichotomously scored items the MHM implies that the sum score on the J items, denoted $X_+ = \sum_j X_j$, stochastically orders people on θ, and thus can be used for ordinal person measurement. Van der Ark (2005) used a simulation study to demonstrate that a set of polytomously scored items consistent with the MHM can also be used for ordering persons.

Like the MHM, many parametric IRT models also assume unidimensionality and local independence but require parametric restrictions on the item response functions. Hemker, Van der Ark, and Sijtsma (2001) showed that the polytomous-item MHM encompasses well-known parametric IRT models such as the graded response model (Samejima, 1969) and the partial credit model (Masters, 1982). IRT models for dichotomous item scores such as the Rasch (1960) model and the 2-parameter logistic model (Birnbaum, 1968) are also special cases of the MHM. Next, we discuss two automated item selection methods in MSA that can be used to partition items into clusters that satisfy the definition of a Mokken scale and approximate the requirements of the MHM.

5.2.1 Mokken Scale Analysis

Let $Cov(X_j, X_k)$ denote the covariance between two items j and k, let $Cov_{max}(X_j, X_k)$ denote the maximum covariance between these items given their marginal item-score distributions, and let rest score $R_{(j)}$ denote the total score on $J-1$ items excluding item j; that is, $R_{(j)} = X_+ - X_j$. Then, the scalability coefficient for an item pair (j,k) is defined as

$$H_{jk} = \frac{Cov(X_j, X_k)}{Cov_{max}(X_j, X_k)};$$

the scalability coefficient for item j is defined as

$$H_j = \frac{Cov(X_j, R_{(j)})}{Cov_{max}(X_j, R_{(j)})};$$

and the scalability coefficient for the total scale is defined as

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

A set of items forms a Mokken scale (Sijtsma & Molenaar, 2002, pp. 67-69) if (1) all inter-item correlations are positive and (2) all coefficients H_j exceed a user-specified, positive lower bound c. Items that do not satisfy the criteria are defined to be unscalable. The requirements of the MHM and the definition of a Mokken scale do not coincide. The MHM implies the first criterion (Holland & Rosenbaum, 1986), but only implies the second criterion for $c = 0$ (Sijtsma & Molenaar, 2002, pp. 58-59). In practice, one requires a higher positive lower bound c (by default equal to .30) because higher values of coefficient H_j imply better item discrimination (Van der Ark, Croon, & Sijtsma, 2008). Automated item selection methods may be applied to partition J items into one or more Mokken scales, and possibly one or more items that may be unscalable (Mokken, 1971; Straat, Van der Ark, & Sijtsma, in press). Because the requirements of the MHM and the definition of a Mokken scale do not coincide, researchers are recommended to check afterwards whether the item response functions of selected items are monotone. Experience has shown that the discrepancy between Mokken scales and MHM requirements are often small in real-data analysis (e.g., Sijtsma, Emons, Bouwmeester, Nykliček, & Roorda, 2008; Straat, Van der Ark, & Sijtsma, 2012a; Wismeijer, Sijtsma, Van Assen, & Vingerhoets, 2008).

The *objective* of MSA's automated item selection methods is to select a first Mokken scale containing as many items as possible, then from the unselected items, if any, to select a second Mokken scale containing as many items as possible, and so on until there are no items left or until items remain that are unscalable (Mokken, 1971; Straat et al., in press). The R package `mokken` (Van der Ark, 2007) contains two item selection algorithms that pursue this objective. One algorithm is the automated item selection procedure (AISP; Sijtsma & Molenaar, 2002) and the other is the genetic algorithm (GA; Straat et al., in press). We briefly describe the two item selection procedures. For a more extensive description, see Straat et al. (in press).

Automated Item Selection Procedure

AISP is a bottom-up item selection procedure. AISP starts with selecting from all $\frac{1}{2}J(J-1)$ item pairs the item pair with the largest H_{jk} value that is significantly larger than 0 and exceeds lower bound c. Subsequently, AISP adds a third item to the scale that (a) correlates positively with the selected items j and k, (b) has an H_j coefficient with respect to the already selected items that is significantly larger than 0 and exceeds lower bound c, and (c) produces the largest H coefficient with the already selected items j and k among all unselected items that satisfy criteria (a) and (b). This step is repeated for a fourth item, a fifth item, and so on, until there are no items left that satisfy the criteria (a) and (b). If items remain unselected, AISP tries to construct a second scale from the unselected items, then a third scale, and so on, until there are no items left or the items left are unscalable.

Genetic Algorithm

GA has the same goal as AISP, but unlike the AISP bottom-up procedure GA mimics an evolutionary process to search among all possible partitionings the partitioning that satisfies MSA's scaling objective (Straat et al., in press). First, GA generates random partitionings and evaluates each partitioning with respect to the scaling objective. Second, the better a partitioning represents the scaling objective, the more likely it is that the partitioning is selected in a new, second population of partitionings that is drawn with replacement from the first population. Crossovers and mutations are applied to the partitionings in the second population, such that some of these partitionings become different from the original partitionings of the first population. Next, GA evaluates the partitionings in the second population and produces a third population following the same rules that were used to produce the second population. After the formation of each population, GA records which partitioning was the best partitioning until the most recent population. If the best partitioning remains the same after a pre-specified number of populations, this partitioning is reported as the final partitioning.

5.3 Method

Due to lack of an analytical method for deriving the minimally required sample size for MSA, we used a simulation study to investigate the minimally required sample size in two stages. In the first stage, we studied the effect of sample size (16 levels, ranging from 50 to 3,500) on the correct assignment of items to scales. In the second stage, we searched for the minimally required sample sizes to obtain at least 80%, 90%, 95%, and 99% correct item assignment.

We also included independent variables in our design that may interact with the effect of sample size on the correct assignment of items. In exploratory factor analysis, Hogarty et al. (2005) found that besides sample size, size of the factor loadings, test length, and correlation between factors may have an effect on correctly assigning items to scales based on the outcome of the exploratory factor analysis. Hence, we varied the same design characteristics, but we used size of the H_j values instead of size of the factor loadings.

5.3.1 Simulation Model

For the data simulation, we assumed a test consisting of J items each with five ordered answer categories scored $x = 0, \dots, 4$. A two-dimensional version of the graded response model (De Ayala, 1994) was used for data simulation. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ be the vector containing two latent variables. Let δ_{jx} ($j = 1, \dots, J$; $x = 1, \dots, 4$) be the difficulty parameter of item j and category x, and let $\boldsymbol{\alpha}_j = (\alpha_{j1}, \alpha_{j2})$ be the vector of discrimination parameters for item j. The two-dimensional graded response model describes the probability of obtaining a score of at least x on item j, given $\boldsymbol{\theta}$,

$$P(X_j \geq x|\boldsymbol{\theta}) = \frac{\exp[\alpha_{j1}(\theta_1 - \delta_{jx}) + \alpha_{j2}(\theta_2 - \delta_{jx})]}{1 + \exp[\alpha_{j1}(\theta_1 - \delta_{jx}) + \alpha_{j2}(\theta_2 - \delta_{jx})]}.$$

5.3.2 Design

Six design characteristics were fixed in the design of the simulation study: (1) the distribution of the latent variables was bivariate standard normal; (2) the number of latent variables equalled 2; (3) the number of answer categories equalled 5; (4) lower bound c equalled the default value of .3; (5) the number of replications in each design cell equalled 100; and (6) the location parameters of the J items were spaced equidistantly, such that location parameters of item j equalled $(-1.5 + \frac{j-1}{J-1}, -1.0 + \frac{j-1}{J-1}, -0.5 + \frac{j-1}{J-1}, 0.0 + \frac{j-1}{J-1})$.

Sample size. We investigated 16 different sample sizes (50, 100, 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, and 3500). A pilot study showed that sample sizes larger than 3,500 AISP and GA produced stable partitionings. Thus, studying larger sample sizes did not seem necessary.

H_j value. A higher item discrimination causes a higher H_j value (De Koning, Sijtsma, & Hamers, 2002). We chose the discrimination parameters such that we obtained conditions with H_j values exceeding .20, .30, or .40 (Table 5.1). One condition had all H_js approximately equal to .22 ($\alpha = 1$), one condition had all H_js approximately equal to .32 ($\alpha = 1.3$), and one condition had all H_j approximately equal to .42 ($\alpha = 1.6$).

Table 5.1: Range of H_j Values.

	Test length		
α	5	10	20
1	.219-.230	.218-.229	.216-.227
1.3	.320-.335	.319-.333	.316-.330
1.6	.414-.430	.412-.430	.411-.428

Test length. We investigated short tests containing 10 items and long tests containing 20 items.

Correlation between latent variables. We chose three values for the correlation between the latent variables: weak ($r(\theta_1, \theta_2) = .3$), strong $r(\theta_1, \theta_2) = .6$), and perfect ($r(\theta_1, \theta_2) = 1.0$) resulting in one effective latent variable. For weakly and strongly correlated latent variables, we simulated data assuming a simple structure with $\alpha_{j1} > 0$ and $\alpha_{j2} = 0$ for the odd-numbered items, and $\alpha_{j1} = 0$ and $\alpha_{j2} > 0$ for the even-numbered items.

Item selection procedure. We used AISP and GA to analyze each data set. We used the R package `mokken` (Van der Ark, 2007) to run AISP and GA.

5.3.3 Dependent variable

PEA and MIN are indices for evaluating a partitioning of a set of items in one or more scales by comparing the obtained partitioning with a baseline partitioning. PEA is defined as the proportion of items that is classified in agreement with the baseline partitioning. Thus, PEA is the proportion of correctly classified items. MIN is defined as the number of items to be moved from one scale to another scale to retain the baseline partitioning. Thus, MIN counts the number of misclassified items. Dividing MIN by the test length yields the proportion of misclassified items, which obviously is the complement of the proportion of correctly classified items; hence, $PEA = 1 - \frac{MIN}{J}$. Proportions, such as PEA, are easier to compare between different test length than counts, such as MIN, because proportions are independent of test length. Hence, we used PEA as the dependent variable in the first stage of the study.

To obtain baseline partitionings, we determined for each condition the ”true” partitioning by simulating one data set containing 1,000,000 observations. The baseline partitionings were the following. For $H_j \approx .22$, all items were unscalable; for $H_j \approx .32$ and $H_j \approx .42$, the “true” partitioning depended on the correlation between the latent variables. If the latent variables had perfect correlation (i.e., $r(\theta_1, \theta_2) = 1$), the “true” partitioning was that the item selection algorithms assigned all items to one scale. If the latent variables had a correlation of .3 or .6, the item selection algorithms assigned the odd-numbered items to one scale and the even-numbered items to a another scale.

In the first stage, we computed in each condition PEA as the proportion of items in agreement with the “true” partitioning. The proportions were based on the product of the number of items and the number of replications. Hence, for $J = 10$ we used 1,000 item classifications and for $J = 20$ we used 2,000 item classifications to obtain stable estimates of the proportions. In the Results section, we report the results from the second stage consisting of the minimally required sample size for different PEA values. Because the literature does not provide guidelines for the interpretation of PEA, we based values of PEA on typical values for Type-I error rates in hypothesis testing (.20, .10, .05, and .01). Type-I error rates refer to wrong decisions, whereas PEA refers to correct decisions. Hence, we used values equal to 1 - Type-I error rates (i.e., .80, .90, .95, and .99). We evaluated the minimally required sample size for mediocre PEA (at least 80% of the items correctly classified), adequate PEA (at least 90% of the items correctly classified), good PEA (at least 95% of the items correctly classified), and excellent PEA (at least 99% of the items correctly classified). We realize that the labels are arbitrary, but we believe that the labels facilitate interpretation of the results.

5.4 Results

The Appendix shows the PEA for each combination of sample size, H_j value, test length and correlation between the latent variables. From the results in the Appendix, we derived the minimally required sample sizes for the four pre-specified levels of PEA (Table 5.2). The results for AISP and GA were almost equal except for the condition with strongly correlated latent variables (i.e., $r(\theta_1, \theta_2) = .6$) and highly discriminating items (i.e., $H_j \approx .42$). In these conditions, GA obtained one scale containing all items instead of two scales, each containing $\frac{J}{2}$ items. Hence, PEA for GA was approximately .5 for all sample sizes because $\frac{J}{2}$ items were correctly assigned to the scale and the other $\frac{J}{2}$ items were incorrectly assigned to the same scale. In Table 5.2 we only report the minimally required sample sizes for AISP.

The minimally required sample size for different levels of PEA mainly depended on the H_j values. If the H_j values were approximately .22 or .32, larger sample sizes were needed for at least an adequate PEA than if H_j values were .42. For $H \approx .22$, the sample size should be at least 750 to 1000 to produce at least mediocre PEA, at least 1000 to 1250 to produce at least adequate PEA, at least 1250 to 2000 to produce at least good PEA, and at least 2750 to 3500 for excellent PEA. For $H_j \approx .32$, the sample size should be

Table 5.2: Minimum Sample Size Requirements for MSA (AISP and GA) for Four Different Levels of PEA.

			Per element accuracy			
H_j	$r(\theta_1, \theta_2)$	J	Mediocre	Adequate	Good	Excellent
.22	.3	10	750	1000	1250	2500
		20	750	1250	1750	3000
	.6	10	500	1000	1250	2750
		20	750	1250	1500	2750
	1.0	10	1000	1250	2000	3250
		20	1000	1750	2000	3500
.32	.3	10	250	750	1500	3500
		20	250	750	1500	3500
	.6	10	500	750	1500	3000
		20	250	750	1250	3000
	1.0	10	250	750	1250	3000
		20	250	750	1500	2750
.42	.3	10	50	50	250	250
		20	50	50	250	250
	.6	10	250	250	500	750
		20	250	250	500	750
	1.0	10	50	50	250	250
		20	50	50	250	250

Note: Per element accuracy is called mediocre if higher than .80, adequate if higher than .90, good if higher than .95, and excellent if higher than .99.

at least 250 to produce at least mediocre PEA, at least 750 to produce at least adequate PEA, at least 1250 to 1500 to produce at least good PEA, and at least 3000 to 3500 for excellent PEA. For $H \approx .42$, the sample size should be at least 50 to 250 to produce at least mediocre to adequate PEA, at least 250 to 500 to produce at least good PEA, and at least 250 to 750 for excellent PEA.

The results showed that the minimally required sample size was larger if the H_j values were close to lower bound .3 (i.e., $H_j \approx .32$). We investigated whether this result could be generalized to the other H_j values (i.e., $H_j \approx .22$ and $H_j \approx .42$) close to lower bounds .2 and .4, respectively. We found that the results for $H_j \approx .22$ with $c = .2$ and $H_j \approx .42$ with $c = .4$ were similar to the results for $H_j \approx .32$ with $c = .3$ (reported in Table 5.2).

5.5 Discussion

We found that the minimally required sample size for item selection in MSA depends on the H_j values relative to lower bound c. For high-quality items with H_j values exceeding lower bound c, MSA rarely misclassifies items if the sample size is at least 250. Test constructors knowing their craft well develop tests based on a well-founded theory and thoroughly think about items that are qualitatively good indicators of the construct of interest. Using the item selection procedures in MSA to partition the set of items into one or more Mokken scales, they easily find the correct partitionings. Hence, test constructors should put additional effort in the construction of good items.

The effect of the difference between H_j values and lower bound c on the minimally required sample size is comparable to the effect of the effect size on the minimally required sample size for t-tests, F-tests in analyses of variance and F-tests and t-tests in regression analyses. The larger the difference between the "true" effect and the effect expressed in the null-hypothesis, the smaller the minimally required sample size to find a significant effect. In MSA, the lower bound c is the effect under the null-hypothesis. The larger the difference between the "true" H_j and the "null" c, the smaller the required sample size to assign the item to the correct scale.

In practice, researchers like to know whether the sample size was sufficiently large to find the correct partitionings given the differences between H_j values and lower bound c. Researchers may consult the standard errors of the H_j values (Kuijpers, Van der Ark, & Croon, 2012), which are available in the R package `mokken`. The standard errors can be used for constructing confidence intervals for the H_j values so as to check whether lower bound c lies within the confidence intervals.

Minimally required sample sizes have already been established for exploratory factor analysis. For exploratory factor analysis, the determination of the minimally required sample size is a complex interplay between the size of the factor loadings and the number of items per factor. The effects of test length and number of latent variables on the minimally required sample size were negligible for MSA. The decision rule for the minimally required sample size for item selection in MSA is easier: For high-quality items, MSA performs well with small sample sizes.

Appendix: Results from Investigation of PEA

Table 5.3: Per Element Accuracy Results for AISP.

		$H_j \approx .22$			$H_j \approx .32$			$H_j \approx .42$		
		$r(\theta_1, \theta_2)$			$r(\theta_1, \theta_2)$			$r(\theta_1, \theta_2)$		
J	N	.3	.6	1.0	.3	.6	1.0	.3	.6	1.0
10	50	0.22	0.23	0.37	0.62	0.48	0.72	0.90	0.65	0.93
	100	0.49	0.45	0.34	0.76	0.71	0.77	0.94	0.78	0.94
	250	0.66	0.67	0.50	0.82	0.79	0.81	0.99	0.92	1.00
	500	0.79	0.81	0.70	0.85	0.87	0.88	1.00	0.97	1.00
	750	0.87	0.87	0.79	0.91	0.90	0.91	1.00	0.99	1.00
	1000	0.93	0.92	0.87	0.92	0.90	0.93	1.00	1.00	1.00
	1250	0.96	0.95	0.90	0.94	0.94	0.95	1.00	1.00	1.00
	1500	0.96	0.97	0.93	0.95	0.95	0.96	1.00	1.00	1.00
	1750	0.97	0.98	0.94	0.96	0.95	0.96	1.00	1.00	1.00
	2000	0.98	0.98	0.96	0.96	0.96	0.95	1.00	1.00	1.00
	2250	0.98	0.98	0.96	0.97	0.96	0.98	1.00	1.00	1.00
	2500	0.99	0.98	0.97	0.97	0.97	0.98	1.00	1.00	1.00
	2750	1.00	0.99	0.98	0.97	0.98	0.98	1.00	1.00	1.00
	3000	1.00	1.00	0.98	0.97	0.98	0.99	1.00	1.00	1.00
	3250	1.00	1.00	0.99	0.98	0.98	0.99	1.00	1.00	1.00
	3500	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00	1.00
20	50	0.24	0.22	0.36	0.69	0.57	0.77	0.91	0.66	0.94
	100	0.49	0.45	0.34	0.76	0.71	0.77	0.94	0.78	0.94
	250	0.53	0.49	0.40	0.81	0.80	0.84	0.99	0.93	1.00
	500	0.70	0.69	0.60	0.85	0.86	0.88	1.00	0.98	1.00
	750	0.80	0.80	0.72	0.91	0.90	0.93	1.00	0.99	1.00
	1000	0.88	0.87	0.81	0.93	0.92	0.93	1.00	1.00	1.00
	1250	0.91	0.93	0.86	0.93	0.95	0.94	1.00	1.00	1.00
	1500	0.94	0.95	0.89	0.95	0.94	0.96	1.00	1.00	1.00
	1750	0.96	0.96	0.92	0.97	0.95	0.96	1.00	1.00	1.00
	2000	0.97	0.97	0.95	0.97	0.97	0.97	1.00	1.00	1.00
	2250	0.97	0.97	0.96	0.98	0.97	0.98	1.00	1.00	1.00
	2500	0.98	0.98	0.96	0.97	0.97	0.98	1.00	1.00	1.00
	2750	0.98	0.99	0.97	0.98	0.98	0.99	1.00	1.00	1.00
	3000	0.99	0.99	0.98	0.98	0.99	0.99	1.00	1.00	1.00
	3250	0.99	0.99	0.98	0.98	0.99	0.99	1.00	1.00	1.00
	3500	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00

Note: N is sample size. J is number of items. $r(\theta_1, \theta_2)$ is the correlation between the latent variables.

Chapter 6

Using Conditional Association to Identify Locally Independent Item Sets[*]

Abstract

The ordinal, unidimensional monotone latent variable model assumes local independence, unidimensionality, and monotonicity, and implies the observable property of conditional association. We specialized conditional association into three useful observable consequences and implemented them in a new procedure. The new procedure aims at identifying items that are locally dependent, removing those items from the initial item set, and producing a subset of items that is consistent with the assumption of local independence. We compared the new procedure with the scaling procedures DETECT and Mokken scale analysis, and found that the new procedure produced larger item sets consistent with the unidimensional monotone latent variable model.

[*] This chapter has been submitted for publication.

6.1 Introduction

The unidimensional monotone latent variable model (UMLVM; Holland & Rosenbaum, 1986; also, see Hemker, Sijtsma, Molenaar, & Junker, 1997; Molenaar, 1997) is a general IRT model that is based on three assumptions. Before we discuss the three assumptions, we introduce notation and definitions. Let j be an item subscript, X_j a polytomous item-score variable adopting discrete, ordered scores $x = 0, \ldots\ m$, J the number of items in the test, and θ the latent variable the items measure. The total score on the J items equals $X_+ = \sum_{j=1}^{J} X_j$ and has realization $x_+ = 0, \ldots, mJ$. For dichotomous items, which may be considered a special case of polytomous items scored $x = 0, 1$, the total score runs from 0 to J. Using this notation, the three assumptions of the UMLVM are:

1. Unidimensionality: latent variable θ is unidimensional;
2. Local independence: item scores are independent conditional on θ;
$$P(X_1 = x_1, \ldots, X_J = x_J|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta); \tag{6.1}$$
3. Monotonicity: the IRFs are monotone nondecreasing in θ; that is,
$$E(X_j|\theta) \text{ is nondecreasing in } \theta. \tag{6.2}$$

For dichotomous items, the UMLVM implies stochastic ordering of the latent variable θ by the total score X_+, abbreviated SOL (Hemker et al. 1997). Let t be an arbitrary value of θ. Then, for two values of the total score denoted C and K such that $0 \leq C < K \leq J$ and any value t, SOL is defined as

$$P(\theta > t|X_+ = C) \leq P(\theta > t|X_+ = K). \tag{6.3}$$

SOL implies that respondents with higher total scores on average have higher θ-values.

For polytomous items, Hemker et al. (1997) showed that SOL does not hold. Hence, the use of the total score as an ordinal estimator of the latent variable is not justified for polytomous items. However, using simulated data Van der Ark

(2005) demonstrated that in real data SOL holds by approximation and may be assumed for all practical purposes. Moreover, Van der Ark and Bergsma (2010) showed that for polytomous items the UMLVM implies weak SOL. To define weak SOL, we use total score x_+ rather than concrete values C and K as in Equation 6.3, so that for each value $x_+ = 1, \ldots, mJ$ weak SOL is defined as

$$P(\theta > t | X_+ < x_+) \leq P(\theta > t | X_+ \geq x_+). \tag{6.4}$$

SOL implies weak SOL but weak SOL does not imply SOL (Van der Ark & Bergsma, 2010). Weak SOL implies a less fine-grained stochastic order of the distribution of θ, which holds for dichotomizations of the total-score scale into disjoint person subsets but not for each value of X_+ (Equation 6.3). For example, for all cut scores x_{+0} based on the X_+ scale, Equation 6.4 implies that the rejected group ($x_+ < x_{+0}$) has a lower mean θ than the selected group ($x_+ \geq x_{+0}$). An application of rejection/selection using an a priori determined cut score is to decide who does not and who does receive a treatment. Also, when one selects a fixed percentage of, say, $P\%$, of the highest-scoring applicants for a course one implicitly cuts the total-score scale into a lower and a higher part.

The UMLVM does not impose a parametric structure on the response probabilities $P(X_j = x_j | \theta)$. As a result, the UMLVM does not enable the numerical estimation of the latent variable. The importance of equations 6.3 and 6.4 is that we have ordinal scales for θ even if θ cannot be numerically estimated. To have an ordinal scale for a real test that satisfies SOL (Equation 6.3) or weak SOL (Equation 6.4), one first has to assess the fit of the UMLVM to the data and conclude that the model indeed fits the data well. Then, by implication the test measures the attribute on an ordinal scale. For unidimensional θ, several goodness-of-fit methods exist that assess observable consequences of the UMLVM; Sijtsma and Molenaar (2002) provide an overview. One observable consequence that could be used for fit assessment is conditional association (CA; Holland & Rosenbaum, 1986; Rosenbaum, 1984, 1988). CA opens a wide array of potentially powerful tools to assess the goodness-of-fit of the UMLVM but has been largely ignored by IRT theorists. The purpose of this article was to study this potential by linking CA to the assumptions of the UMLVM, finding out which assumptions CA can assess best, and combining the results in a new procedure for identifying locally independent item sets.

This article is organized as follows. First, CA is introduced, three special cases of CA are proposed as candidates for investigating the fit of the UMLVM, and the results of a computational study that investigated the three special cases when assumptions of the UMLVM do not hold are presented. Second, based on the results of the computational study, a procedure for identifying locally independent sets of items is discussed. Third, a simulation study is presented in which the use of the new procedure is compared with the methods DETECT and Mokken scale analysis. Fourth, the new procedure is applied to real data. Fifth, merits and drawbacks of the new procedure are discussed.

6.2 Conditional Association

6.2.1 Definition of Conditional Association

Let $\mathbf{X}$ contain the J item-score random variables and let these variables be divided in two mutually exclusive sets of item scores, denoted by $\mathbf{Y}$ and $\mathbf{Z}$, so that $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. Further, let f_1 and f_2 be nondecreasing functions and let h be any function. Let $\sigma(.,.)$ denote the population covariance and $s(.,.)$ the sample covariance. Holland and Rosenbaum (1986, Theorem 6) proved that the UMLVM implies CA, which is defined as

$$\sigma\left[f_1(\mathbf{Y}), f_2(\mathbf{Y}) | h(\mathbf{Z}) = \mathbf{z}\right] \geq 0. \tag{6.5}$$

Specific choices of $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ and f_1, f_2, and h enable a large number of special cases that each impose restrictions on the data. Checking whether the data are consistent with these restrictions provides a powerful method to assess the goodness-of-fit of the UMLVM but also confronts the data analyst with the question of how to limit the number of possibilities and how to combine the numerous results that each of the possibilities produces. The multitude of special cases of Equation 6.5 and concrete data results might explain why CA has not become a standard tool in IRT goodness-of-fit research.

6.2.2 Three Special Cases

Let items be identified by subscripts j, k and l. In three special cases of CA, $\mathbf{Y}$ contains the scores on a pair of items j and k. The cases are the following.

1. Let $f_1(\mathbf{Y}) = X_j$ and $f_2(\mathbf{Y}) = X_k$, and ignore set $\mathbf{Z}$. Then CA reduces to the well-known inter-item covariance (Holland & Rosenbaum, 1986, p. 1537),

   $$\sigma(X_j, X_k) \geq 0. \tag{6.6}$$

   In practical item analysis, researchers intuitively adopt the idea that items measuring the same attribute must correlate positively, and IRT models such as the UMLVM that assume local independence and monotonicity imply Equation 6.6 (Mokken, 1971, p. 120).

2. Let $h(\mathbf{Z}) = X_l$, so that CA reduces to

   $$\sigma(X_j, X_k | X_l = x_l) \geq 0. \tag{6.7}$$

   Equation 6.7 shows that in the subgroup scoring $X_l = x_l$, the inter-item covariance is always non-negative provided the UMLVM is the correct model.

3. Holland and Rosenbaum (1986, Equation 6.1) suggested the third case. Let $R_{(jk)}$ be the total score on the items except items j and k, also known as the rest score. Then if $h(\mathbf{Z}) = R_{(jk)} = \sum_{i \neq j,k} X_i$, CA reduces to

   $$\sigma(X_j, X_k | R_{(jk)} = r) \geq 0. \tag{6.8}$$

   Equation 6.8 shows that for any subgroup of individuals that have the same rest score the inter-item covariance is non-negative.

The UMLVM implies CA; hence, a negative sign of any of the sample estimates of the covariances in equations 6.6, 6.7, and 6.8 found in a data set is inconsistent with the UMLVM and leads to the conclusion that, strictly speaking (i.e., ignoring sampling fluctuation), the UMLVM is not the model that generated the data. Reversely, if one finds only positive signs in the data for equations 6.6, 6.7, and 6.8, this result logically does not imply that the UMLVM is the data-generating model but many positive signs do provide increasing support for the UMLVM. This is where the strength of CA resides: The many covariances one may check in the data together build a strong case for the UMLVM. As noticed, CA's drawback is the abundance of covariances that have to be checked. For example, for $J = 20$ Likert items with $m = 5$ ordered answer categories, there are $\binom{J}{2} = 190$ covariances, $\sigma(X_j, X_k)$ (Equation 6.6); $m\binom{J}{3} = 5,700$ covariances conditional on an item score, $\sigma(X_j, X_k|X_l)$ (Equation 6.7); and $(m-1)(J-2)\binom{J}{2} = 13,680$ covariances conditional on the rest score, $\sigma(X_j, X_k|R_{(jk)})$ (Equation 6.8).

If the UMLVM does not fit the data, the interesting question that forces itself upon the researcher is whether the misfit is due to assumptions of the model that are inconsistent with the data. One inconsistency is that a multidimensional $\boldsymbol{\theta}$ is needed to explain the associations between the items, and an incorrectly assumed unidimensional θ (Equation 6.1) produces dependencies among particular item pairs thus suggesting UMLVM misfit. Another inconsistency stems from non-monotone relationships between item scores and the latent variable, so that monotonicity (Equation 6.2) cannot capture the true relationship well. For practical data analysis, two questions need to be answered.

First, how are equations 6.6, 6.7, and 6.8 related to model violations of local independence (Equation 6.1) and monotonicity (Equation 6.2)? A distinction is made between two violations of local independence. One violation is positive local dependence (PLD; $\sigma(X_j, X_k|\theta) > 0$), and the other violation is negative local dependence (NLD; $\sigma(X_j, X_k|\theta) < 0$); also, see Chen and Thissen (1997), Rosenbaum (1988), and Yen (1984). It has to be decided whether, for example, $\sigma(X_j, X_k|X_l) < 0$ means that items j and k are PLD or NLD. Next, one also needs to know whether the negative sign may be due to a violation of monotonicity in one or more of the IRFs. Once this question has been answered, specific conditional covariances may be selected to assess a specific model violation.

Second, how can we combine the information from, say, 5,700 covariances of the type $\sigma(X_j, X_k|X_l)$ into meaningful conclusions, meanwhile taking into account that several negative signs may be due to sampling fluctuation? A systematic procedure is required that navigates through the abundance of information.

6.2.3 Detecting Violations of UMLVM Assumptions

Suppose we investigate covariances pertaining to items a, b ($\{X_a, X_b\} \in \mathbf{Y}$), and c ($X_c \in \mathbf{Z}$). We use shorthand notation $\sigma_{ab} = \sigma(X_a, X_b)$, $\sigma_{ab|c} = \sigma(X_a, X_b|X_c)$, and $\sigma_{ab|R} = \sigma(X_a, X_b|R_{(ab)})$; and s replaces σ when sample covariances are considered. Subscripts j, k, and l refer to any other item in the test. Notation PLD(j, k) means that items j and k are PLD, and NLD(j, k) that the items are NLD. Notation NM(j) means that the IRF of item j is non-monotone.

Holland and Rosenbaum (1986) and Rosenbaum (1988) provided analytical proof that several covariances based on equations 6.6, 6.7, or 6.8 are also positive when the UMLVM does not hold. Thus, these covariances are useless to detect misfit of the UMLVM. In Table 6.1, 0 values refer to these covariances. Proof of their positivity was given by Rosenbaum (1988, Theorem 4; superscript 1 in Table 6.1), Rosenbaum (1988, Theorem 1; superscript 2), and Holland and Rosenbaum (1986, Equation 5; superscript 3).

For the covariances in equations 6.6, 6.7, or 6.8 that can be negative when the UMLVM does not hold, we did a computational study to identify which of these covariances were negative with high probability when particular assumptions did not hold. Two violations of local independence that we studied were PLD, NLD, and we studied non-monotone IRFs. Covariances that are negative with high probability when PLD, NLD or NM holds are well suited to detect the violation.

Table 6.1 shows the proportions of negative values covariances σ_{ab}, $\sigma_{ab|c}$, and $\sigma_{ab|R}$ (columns) attained under violations of the UMLVM (rows). Note that the violation given in a particular row is the only violation of the UMLVM. The proportion of negative values may be interpreted as the power of a covariance to identify PLD, NLD or NM. The first column refers to covariance σ_{ab}; hence, all cells that refer to item c are empty. We notice that proportions depend on particular design choices made in the computational study, and that different choices might have resulted in somewhat different proportions; see the Appendix. For our purpose, these small differences are unimportant as we meant to identify covariances that have enough power to be useful for assessing fit of the UMLVM to data.

We summarize the results of each column in Table 6.1. The first column shows that σ_{ab} may be negative if items a and b are NLD. However, $\sigma_{ab|R}$ (third

Table 6.1: Power of Covariances (Columns) to Detect Model Violations (Rows).

		Covariances		
Type of violation		σ_{ab}	$\sigma_{ab\|c}$	$\sigma_{ab\|R}$
Both items a, b in PLD item-pair	PLD(a, b)	0^1	0^1	0^1
One item a/b and conditioning item c in PLD item-pair	PLD(a, c)		.314	.318
	PLD(b, c)		.314	.318
One item a/b in PLD item-pair	PLD(a, j)	$0^{1,2}$	$0^{1,2}$	.318
	PLD(b, j)	$0^{1,2}$	$0^{1,2}$	.318
Conditioning item c in PLD item-pair	PLD(c, j)		$0^{1,2}$	0^1
Both items a, b in NLD item-pair	NLD(a, b)	.497	.652	.774
One item a/b and conditioning item c in NLD item-pair	NLD(a, c)		.000	.000
	NLD(b, c)		.000	.000
One item a/b in NLD item-pair	NLD(a, j)	0^2	0^2	.000
	NLD(b, j)	0^2	0^2	.000
Conditioning item c in NLD item-pair	NLD(c, j)		0^2	0^2
One item a/b violates M	NM(a)	.000	.000	.000
	NM(b)	.000	.000	.000
Conditioning item c violates M	NM(c)		0^3	0^3

^{1,2,3} Superscripts are explained in the text.

column) is more powerful than σ_{ab} for detecting NLD(a, b). The result for σ_{ab} is not pursued further. The second column shows that a negative value of $\sigma_{ab|c}$ may occur for PLD(a, c), PLD(b, c), or NLD(a, b). Hence, sample covariances $s_{ab|c} < 0$ may be used to detect PLD(a, c) and PLD(b, c); this result we call Result 1. For detecting NLD(a, b), $\sigma_{ab|R}$ is more powerful than $\sigma_{ab|c}$. Therefore, detecting NLD by means of $s_{ab|c}$ is not pursued further. The third column shows that a negative value of $\sigma_{ab|R}$ may occur when either a or b is in a PLD item pair. Sample covariances $s_{aj|R} < 0$ and $s_{bj|R} < 0$ (for $j \neq a, b$) may be used to detect these PLD pairs; this result we call Result 2. A negative value of $\sigma_{ab|R}$ may also occur when both a and b are in the same NLD item pair. Because $\sigma_{ab|R}$ is the most powerful covariance for detecting NLD(a, b), sample covariances $s_{ab|R} < 0$ may be used to detect NLD(a, b); this is Result 3.

Finally, the conditional covariances did not have sufficient power to detect NM. Hence, CA as operationalized here cannot be used to assess monotonicity; see Sijtsma and Molenaar (2002, chap. 3) for an alternative method to assess monotonicity.

6.3 A Procedure to Identify a Locally Independent Item Set

We propose a procedure, called CA procedure, that uses Results 1, 2, and 3 to flag items that are suspected to be locally dependent, and that enables us to delete some or all of the flagged items to obtain a locally independent item subset from the original J-item set.

6.3.1 Flagging Suspected Items

Three indices, denoted $W^{(1)}$, $W^{(2)}$ and $W^{(3)}$, were used to flag suspected items. The three indices are counts of negative conditional sample covariances. A problem of simply counting sample covariances is that small samples produce many negative covariances simply due to sampling fluctuation. Thus, it makes sense to give more weight to the count of a negative covariance if the covariance was estimated in a larger sample. It is well known that the gain in precision diminishes as sample size is larger, and for the standard error of the sample mean, $SE(\bar{X}) = \frac{s_X}{\sqrt{n}}$, $\sqrt{n}$ expresses this phenomenon. For covariances, a simple equation for the standard error is unavailable and for simplicity we thus weigh negative covariance by $\sqrt{n}$, where n is the size of the sample in which the covariance was estimated.

Let $I(A)$ be an indicator function attaining values $I(A) = 1$ if A is true, and $I(A) = 0$ otherwise. The three indices are defined as follows.

Index $W^{(1)}$ is determined for each item pair, and for item-pair (a, c),

$$W_{ac}^{(1)} = \sum_{j \neq a,c} \sum_{x} I[s(X_a, X_j | X_c = x) < 0] \times \sqrt{n_x}. \tag{6.9}$$

In Equation 6.9, $I[s(X_a, X_j|X_c = x) < 0]$ indicates whether the covariance is negative (value 1) and n_x is the size of the group scoring $X_c = x$. Note that

$$W_{ac}^{(1)} = \sum \sum_{x} I[s(X_a, X_j | X_c = x) < 0] \times \sqrt{n_x}$$

and

$$W_{ca}^{(1)} = \sum \sum_{x} I[s(X_c, X_j | X_a = x) < 0] \times \sqrt{n_x}$$

are not the same, so we have a total of $J(J-1)$ different $W^{(1)}$ values. If $W_{ac}^{(1)}$ is large, then item-pair (a, c) likely is PLD.

Index $W^{(2)}$ is determined for each item. Index $W^{(2)}$ is based on Result 2, and it is a weighted count of all negative covariances $s_{aj|R}$ in which item a is involved. For item a,

$$W_a^{(2)} = \sum_{j \neq a} \sum_{r} I[s(X_a, X_j | R_{(aj)} = r) < 0] \times \sqrt{n_r}.$$

If $W_a^{(2)}$ is large, then item a likely is in a PLD item pair.

Index $W^{(3)}$ is determined for each item pair. Index $W^{(3)}$ is based on Result 3, and it is a weighted count of all negative covariances $s_{ab|R}$ in which item-pair (a, b) is involved. For item-pair (a, b),

$$W_{ab}^{(3)} = \sum_{r} I[s(X_a, X_b | R_{(ab)} = r) < 0] \times \sqrt{n_r}.$$

If $W_{ab}^{(3)}$ is large, then item-pair (a, b) likely is NLD.

The next step is to determine which index values are large enough to flag item(s) as suspect. If most items are locally independent, most W values are low and the distribution of each index is positively skewed. We used Tukey's fences (Tukey, 1977, a.k.a. the box plot) to determine whether a score is extremely high, but we adjusted the box plot for skewness (Hubert & Vandervieren, 2008; Kimber, 1990). Let M and Q_3 be the median and the third quartile of the distribution, then a W index is discordant if it exceeds the upper fence chosen at $Q_3 + 3 \times (Q_3 - M)$. Each item for which one or more W values are discordant is flagged.

6.3.2 Removing Flagged Items

Each item relates to $2(J-1)$ indices $W^{(1)}$, 1 index $W^{(2)}$, and $J-1$ indices $W^{(3)}$, so that each item may be flagged any number of times between 0 and $3J-2$.

Removing an item from the test may affect the number of flags for the other items. For example, if item pair (a, b) is flagged by index $W^{(1)}$, and items a and b have not been flagged elsewhere, then removing b from the item set clears the flags for item a. After removal of item b, item a is consistent with the UMLVM. This suggests removing items one by one rather than removing all flagged items at once. We advocate the following procedure.

We want the procedure to remove the smallest number of items possible so as to obtain the longest locally independent item set. We based the removal of items on an algorithm that Ligtvoet, Van der Ark, Te Marvelde, and Sijtsma (2010) proposed for a problem that is different from ours but has some formal similarities. First, for each item we counted the number of flags across the three W indices. Then, we removed the item with the largest number of flags, for each item counted the number of flags again and removed the item with the largest number of flags that appeared at this stage. This procedure was repeated until there were no flags left. At any stage of the removal algorithm, two or more items may have the same total number of flags. Thus, one has to consider an additional criterion to remove items. For his purpose, we chose to remove the item with the weakest discrimination power using Mokken's (1971, pp. 151-152) item-scalability coefficient H_j (Van Abswoude, Van der Ark, & Sijtsma, 2004).

6.4 Comparison of Methods Assessing Fit to UMLVM

Next, we compared the procedure with two methods that also aim at selecting items that are consistent with the UMLVM.

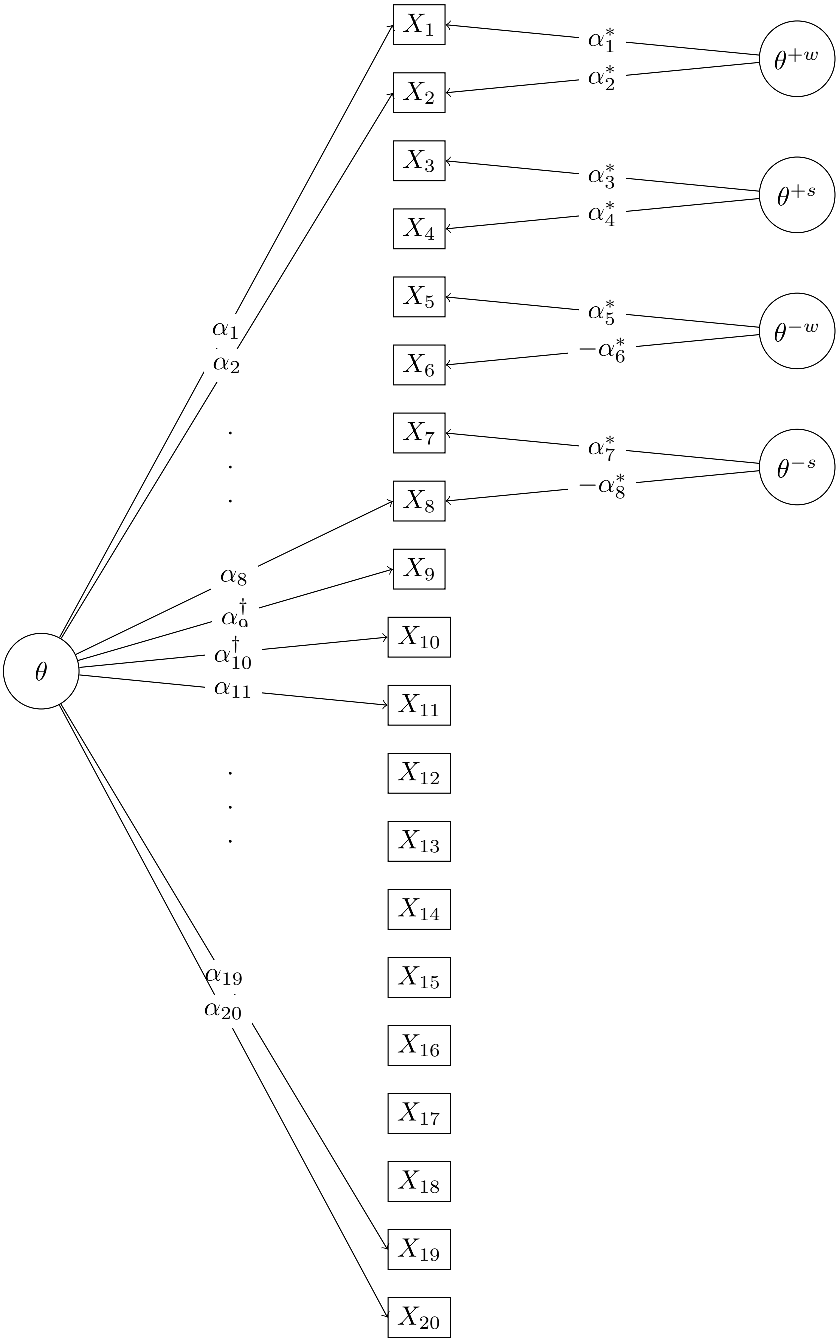
6.4.1 Method

We investigated the sensitivity and the specificity of the CA procedure, DETECT (Zhang, 2007; Zhang & Stout, 1999a, 1999b), and Mokken scale analysis (Mokken 1971; Sijtsma & Molenaar, 2002) for identifying items that are inconsistent with local independence and monotonicity. The CA procedure identifies and removes PLD and NLD item-pairs to obtain a locally independent item set. DETECT uses the *DETECT*-index, which is based on the mean of $\sigma_{jk|R}$ and $\sigma_{jk|X_+}$, to identify one or several subsets of items, if present in the data, that are locally independent. As items j and k are both included in total score X_+, covariance $\sigma_{jk|X_+}$ is not a special case of CA, but DETECT averages the covariances $\sigma_{jk|R}$ and $\sigma_{jk|X_+}$ in an effort to reduce bias in the estimation of $\sigma_{jk|\theta}$ (Zhang, 2007; Zhang & Stout, 1999a, 1999b). Mokken scale analysis uses scalability coefficient H_j, which is based on σ_{jR}, to identify within a set of items one or several *Mokken scales* that satisfy particular scaling criteria (Mokken 1971, pp. 184-185; Sijtsma & Molenaar, 2002, pp. 67-68). As the H_j coefficient is positively related to the item discrimination, Mokken scale analysis relies on the slope of the IRFs to identify item subsets, and unlike the CA procedure and DETECT, Mokken scale analysis concentrates on the monotonicity assumption.

Design

Figure 6.1 illustrates the simulation model. We used the multidimensional graded response model with $\boldsymbol{\theta} = (\theta, \theta^*)$ to simulate 100 data sets. Each data set contained the scores of 1000 persons on 20 polytomous items with five ordered item scores ($m = 4$). The θs were standard normal. The 20 items measured the dominant latent variable θ. The item discrimination parameters for this latent variable were drawn from $\ln[N(0.2, 0.05)]$. For each item, $m = 4$ location parameters were drawn from $N(0, 1)$. Graded response models require that the location parameters of the same item have an increasing order, hence the four sampled location parameters were ordered from smallest to largest.

Ten items were inconsistent with UMLVM assumptions. Violations were either weak or strong. Discrimination parameters for the nuisance latent variable θ^* were drawn from $\ln[N(0.2, 0.05)]$ and $\ln[N(0.9, 0.01)]$, respectively. Thus, a weak violation involved a smaller discrimination parameter (mean: $\alpha^* = e^{0.2} = 1.22$) than a strong violation (mean: $\alpha^* = e^{0.9} = 2.46$). For θ^*, we used superscripts w for weak local dependence and s for strong local dependence, and combined each of them with either "+" for PLD or "−" for NLD. Thus, local dependence was induced by $\theta^{w+}(X_1$ and $X_2), \theta^{s+}(X_3$ and $X_4), \theta^{w-}(X_5$ and $X_6)$, and $\theta^{s-}(X_7$ and $X_8)$. A weak violation of monotonicity (X_9) involved a decreasing IRF with discrimination parameter $-\alpha_9^\dagger$ on $\theta \in (-0.5, 0.5)$, thus affecting 38% of the sample, and a strong violation of monotonicity (X_{10}) involved a decreasing IRF with discrimination parameter



Note: Discrimination parameters with an asterisk induce local dependence and discrimination parameters with a dagger induce violations of monotonicity.

Figure 6.1: Graphical representation of the multidimensional graded response model used in Study II.

$-\alpha_{10}^{\dagger}$ on $\theta \in (-1, 1)$, thus affecting 68% of the sample.

Dependent variables

Specificity was defined as the proportion of analyses within a design cell in which a method correctly identified an item or an item pair to be *consistent* with the UMLVM. For PLD and NLD, sensitivity was defined twofold: (Type 1) the proportion of analyses within a design cell in which a method correctly removed *one item* from a locally dependent item pair; and (Type 2) the proportion of analyses within a design cell in which a method correctly removed *one or both items* from a locally dependent item pair. Data analysis based on the first definition aims to retain as many items as possible in the scale; that is, it does not delete more items than necessary. Sensitivity according to the second definition is always higher than sensitivity according to the first definition. For violations of monotonicity, sensitivity was defined as the proportion of analyses in which a method correctly removed an item violating monotonicity. We used the R package `CAprocedure` (Straat, 2012), the program DETECT for polytomous items (Zhang, 2007), and the R package `mokken` (Van der Ark, 2007) for the CA procedure, DETECT, and Mokken scale analysis, respectively.

6.4.2 Results

The CA procedure had the best specificity: The CA procedure correctly identified 97.6 percent of the items that were consistent with the UMLVM, DETECT identified 68.8 percent, and Mokken scale analysis 67.5 percent. Table 6.2 shows that for weak PLD, the CA procedure more often correctly removed one item from the item set (higher Type-1 sensitivity) than DETECT, but the CA procedure also more often failed to remove items (lower Type-2 sensitivity). For strong PLD, the CA procedure removed one item in 71 percent of the analyses and it produced longer item sets than DETECT and Mokken scale analysis. For NLD, DETECT was the most sensitive method followed by the CA procedure and Mokken scale analysis, respectively. For weak violations of monotonicity, the CA procedure was less sensitive than DETECT and Mokken scale analysis, but for strong violations the three methods were equally sensitive.

Table 6.2: Sensitivity of the New Procedure, DETECT, and Mokken Scale Analysis.

Scaling procedure	Violation					
	PLD		NLD		Violation M	
Type	Weak	Strong	Weak	Strong	Weak	Strong
New procedure						
1 item	16%	71%	85%	78%	1%	84%
1 or 2 items	16%	77%	86%	100%		
DETECT						
1 item	3%	0%	94%	96%	50%	84%
1 or 2 items	62%	99%	97%	100%		
Mokken scale analysis						
1 item	24%	6%	49%	25%	35%	84%
1 or 2 items	60%	96%	82%	85%		

6.5 Empirical Example: The Type D Scale-14

To study the performance of the CA procedure in real-data analysis, we used data from 3,181 persons who responded to the Type D Scale-14 (DS14) questionnaire (Table 6.3). The DS14 is a standard measurement instrument for the distressed personality trait – Type D, for short – and contains two seven-item scales measuring the traits negative affectivity (NA) and social inhibition (SI). Three substraits called feelings of dysphoria, anxious apprehension, and irritability drive NA, and three substraits called discomfort in social situations, reticence, and lack of social poise drive SI (Denollet, 2005; Svansdottir et al. 2011). Different subsets of items from the seven-item sets for the NA-scale and the SI-scale represent the two subtrait triplets: Items NA1, NA2, and NA3 represent feelings of dysphoria; items NA4 and NA5 represent anxious apprehension; items NA6 and NA7 represent irritability; items SI1, SI2, and SI3 represent discomfort in social situations; items SI4 and SI5 represent reticence; and items SI6 and SI7 represent lack of social poise. Given the item structure (Table 6.3), we expect that a set of items measuring the same substrait is PLD. Next, we discuss the results of each step of the CA procedure of the DS14 data.

Table 6.3: H_j Coefficients for the Negative Affectivity Scale and the Social Inhibition Scale.

Item	Content	H_j
Negative affectivity scale		
NA1	Often feels unhappy	0.487
NA2	Takes gloomy view of things	0.555
NA3	Is often down in the dumps	0.589
NA4	Worries about unimportant things	0.430
NA5	Often worries about something	0.527
NA6	Is easily irritated	0.470
NA7	Is often in a bad mood	0.464
Social inhibition scale		
SI1	Inhibited in social interactions	0.491
SI2	Difficulties starting a conversation	0.547
SI3	Does not find things to talk about	0.527
SI4	Closed kind of person	0.515
SI5	Keeps others at a distance	0.493
SI6	Makes contact easily	0.551
SI7	Often talks to strangers	0.457

Table 6.3 shows the H_j coefficients that the CA procedure uses in case of ties with respect to the number of flags per item. Table 6.4 shows the three W indices for the NA scale. The upper fences for the box plots of the W indices were 0, 503, and 186.43, respectively. The CA procedure flagged all items that had at least one $W^{(1)}$ value larger than upper fence 0. For indices $W^{(2)}$ and $W^{(3)}$, none of the items had W values exceeding the upper fence. Item NA3 had six flags and was removed first. Removal of Item NA3 resulted in the vanishing of the flag initially assigned to item pair (NA3, NA7), leaving four flags for Item NA7. Consequently, Item NA7 was the second item that was removed. Without items NA3 and NA7, no flags were left.

Table 6.5 shows the $W^{(2)}$ and $W^{(3)}$ values for the SI scale. All $W^{(1)}$ values equalled 0. For $W^{(2)}$ and $W^{(3)}$, the upper fences were 445.02 and 203.01, respectively. Hence, the CA procedure flagged no items. However, Item SI5 had a $W^{(2)}$ value (i.e., 431.04) close to the upper fence (i.e., 445.02) and item pairs (SI1,SI5) and (SI4,SI7) had $W^{(3)}$ values (i.e., 187.14 and 164.28, respectively)

Table 6.4: W Indices for the Negative Affectivity Scale.

	$W^{(1)}$							$W^{(2)}$	$W^{(3)}$						
Item	NA1	NA2	NA3	NA4	NA5	NA6	NA7		NA1	NA2	NA3	NA4	NA5	NA6	NA7
NA1		0	22.58	0	0	0	0	220.74							
NA2	0		16.70	0	0	0	4.47	171.53	5.39						
NA3	0	0		0	0	0	0	255.87	18.41	28.79					
NA4	0	0	16.70		0	0	4.47	377.09	104.81	54.12	72.86				
NA5	0	0	40.50	0		0	4.47	193.57	23.28	23.42	0	3.46			
NA6	0	0	5.57	0	0		4.47	224.52	4.69	37.06	112.81	22.86	43.55		
NA7	0	0	57.51	0	0	0		332.72	64.17	22.76	22.99	118.99	100.07	3.74	

close to the upper fence (i.e., 203.01).

Table 6.5: *W* Indices for the Social Inhibition Scale.

	$W^{(2)}$	$W^{(3)}$						
Item		SI1	SI2	SI3	SI4	SI5	SI6	SI7
SI1	328.03							
SI2	191.46	2.65						
SI3	161.53	13.28	2.64					
SI4	300.90	32.66	0	42.24				
SI5	431.04	187.14	10.35	79.31	0			
SI6	316.71	61.11	75.24	4.24	61.72	107.47		
SI7	369.55	31.20	100.57	19.81	164.28	46.77	6.92	

The CA procedure suggested to remove two items from the NA-scale (NA3 and NA7) and to keep all items in the SI-scale. From a theoretical viewpoint (Denollet, 2005), because these items measured different subtraits it was to be expected that these items had to be removed from the NA scale. Hence, the CA procedure identified two of the three NA subscales, but none of the three SI subscales.

6.6 Discussion

The CA procedure has higher specificity than DETECT and Mokken scale analysis. Thus, the latter two methods show a tendency to remove items that are consistent with the UMLVM, whereas the CA procedure procedure tends to keep such items in the item set. Not rejecting fitting items is desirable, in particular when the number of available items is small as with narrowly defined attributes, and losing items that only deviate little from the majority might unnecessarily reduce reliability and trait coverage.

For detecting PLD and NLD, the CA procedure has sensitivity similar to DETECT and Mokken scale analysis. Thus, the methods are equally good at identifying locally dependent items, which next are candidates for removal from the scale. An advantage of the CA procedure is that it suggests removing only one item in a locally dependent item pair, whereas DETECT and Mokken scale analysis usually remove both items. As a result, the CA procedure retains more items in the item set than DETECT and Mokken scale analysis and again avoids the unnecessary removal of items.

The CA procedure did not identify violations of monotonicity well. As local independence and monotonicity together imply CA, the lack of power of the CA procedure to identify items with decreasing IRFs came rather unexpectedly. To identify decreasingness by means of conditional covariances, it appears that along a large interval of the latent-variable scale where the population is located the IRF of one item should be increasing and the IRF of the other item should be decreasing. In model fit research, nonparametric regression methods of item score on rest score should be used to estimate IRFs and to detect violations of monotonicity (Junker & Sijtsma, 2000; Ramsay, 1991).

In the real-data example, we found an upper fence equal to 0, which implies that even for small sample fluctuations resulting in negative covariances items are flagged. This may happen when all items are consistent with the UMLVM. Hence, all items with low but nonzero W values are flagged and a large number of items is incorrectly removed. To correct the unfortunate removal of items, in real-data analysis a minimum value for the upper fence equal to, say, 20, may be used.

Given the results for the CA procedure, we suggest using the method as follows. If the scale construction is exploratory, we suggest to use the automated item selection procedure from Mokken scale analysis (Sijtsma & Molenaar, 2002, chap. 5; Straat, Van der Ark, & Sijtsma, in press). Then, for each identified scale the CA procedure might be used to exclude PLD and NLD items, and nonparametric regression might be used to assess IRF monotonicity. The program `mokken` (Van der Ark, 2007) may be used to run the automated item selection procedure and next to investigate monotonicity. If the scale construction is confirmatory, one may start right away using the CA procedure followed by nonparametric-regression IRF assessment.

Appendix

The aim of the computational study was to estimate the probability of a negative population covariance of the type σ_{jk}, $\sigma_{jk|l}$, and $\sigma_{jk|R}$, given either

PLD, NLD, or a violation of monotonicity. Only if the probability is high, a covariance is a likely candidate for detecting violations of UMLVM assumptions. The study only considered the cases for which a nonnegative covariance could not be proven analytically; these cases correspond to the cells in Table 6.1 containing proportions.

Method

Computational model

We assumed a 5-item test with items scored $x = 0, \ldots, 4$. Population covariances were derived from a two-dimensional graded response model (De Ayala, 1994). The computational details for the population covariances can be obtained from the first author. Vector $\boldsymbol{\theta} = (\theta, \theta^*)$ contained a dominant latent variable θ and a nuisance latent variable θ^*. Let δ_{jx} ($j = 1, \ldots, 5$; $x = 1, \ldots, 4$) be the difficulty parameter of item j and category x, and $\boldsymbol{\alpha}_j = (\alpha_j, \alpha_j^*)$ the vector of discrimination parameters for item j. Parameter α_j^* is the discrimination of item j on θ^*. The two-dimensional graded response model is defined as

$$P(X_j \geq x|\boldsymbol{\theta}) = \frac{\exp[\alpha_j(\theta - \delta_{jx}) + \alpha_j^*(\theta^* - \delta_{jx})]}{1 + \exp[\alpha_j(\theta - \delta_{jx}) + \alpha_j^*(\theta^* - \delta_{jx})]}.$$

Latent variables θ and θ^* were standard normally distributed and correlated zero. A histogram of 51 equidistant intervals ranging from -2.5 to 2.5 approximated the distributions of the latent variables. A pilot study showed that item difficulty did not affect the sign of the covariances under investigation; hence, difficulties were fixed. Let $\boldsymbol{\delta}_j = (\delta_{j1}, \delta_{j2}, \delta_{j3}, \delta_{j4})$. The values of the difficulty parameters were $\boldsymbol{\delta}_1 = (-1.5, -0.75, 0.25, 1)$, $\boldsymbol{\delta}_2 = (-1.25, -0.5, 0.5, 1.25)$, $\boldsymbol{\delta}_3 = (-1, -0.25, 0.75, 1.5)$, $\boldsymbol{\delta}_4 = (-0.75, 0, 1, 1.75)$, and $\boldsymbol{\delta}_5 = (-0.5, 0.25, 1.25, 2)$.

PLD (Table 6.1, upper panel)

Items 1 and 2 were PLD, and items 3, 4, and 5 were consistent with the UMLVM. Discrimination parameters relative to θ were positive ($\alpha_j > 0$, $j = 1, \ldots, 5$). Items 1 and 2 had positive discrimination relative to θ^* (α_1^*, $\alpha_2^* > 0$), and α_3^*, α_4^*, $\alpha_5^* = 0$. As a result, $\sigma_{12|\theta} > 0$ (the covariance depends on θ^*) and all other $\sigma_{jk|\theta} = 0$.

Discrimination parameters α_3, α_4, and α_5 were fixed to 1.5. A pilot study showed that discrimination parameters not related to model violations had negligible effects on the sign of the covariances. The independent variables were discrimination parameters α_1, α_2, α_1^*, and α_2^*. Each discrimination parameter had 13 levels equally spaced between 0.25 to 3.25 (Table 6.6, first column), yielding $13^4 = 28,651$ combinations of discrimination parameters.

The two dependent variables were: (1) the proportion of negative values for $\sigma_{13|2}$ and $\sigma_{23|1}$ (in 28,651 models), which estimated the probability of finding a negative value of $\sigma_{jk|l}$ under PLD(j,l) (denoted $P[\sigma_{jk|l} < 0|\text{PLD}(j,l)]$; see Table 6.1, second row, second column); and (2) the proportion of negative values for $\sigma_{13|R}$ and $\sigma_{23|R}$, which estimated both $P[\sigma_{jk|R} < 0|\text{PLD}(j,l)]$ and $P[\sigma_{jk|R} < 0|\text{PLD}(j,g)]$ (Table 6.1, second and third row, third column).

Table 6.6: Discrimination Parameters for Local Dependence and Violation of Monotonicity Conditions.

Discrimination	Type of violation	
Parameter	Local dependence	Violation of monotonicity
α_1	0.25, 0.5,..., 3.25	0.25, 0.5,..., 3.25
α_2	0.25, 0.5,..., 3.25	0.25, 0.5,..., 3.25
α_3	1.5	0.25, 0.5,..., 3.25
α_4	1.5	1.5
α_5	1.5	1.5
α_1^*	0.25, 0.5,..., 3.25	0
α_2^*	0.25, 0.5,..., 3.25	0

NLD (Table 6.1, middle panel)

We chose $\alpha_2^* < 0$ to induce NLD$(1,2)$. The five dependent variables were proportions of negative values of (1) σ_{12} estimating $P[\sigma_{jk} < 0|\text{NLD}(j,k)]$ (Table 6.1, fifth row, first column), (2) $\sigma_{12|3}$ estimating $P[\sigma_{jk|l} < 0|\text{NLD}(j,k)]$ (Table 6.1, fifth row, second column), (3) $\sigma_{12|R}$ estimating $P[\sigma_{jk|R} < 0|\text{NLD}(j,k)]$ (Table 6.1, fifth row, third column), (4) $\sigma_{13|2}$ and $\sigma_{23|1}$

estimating $P[\sigma_{jk|l} < 0|\text{NLD}(j,l)]$ (Table 6.1, sixth row, second column), and (5) $\sigma_{13|R}$ and $\sigma_{23|R}$ estimating $P[\sigma_{jk|R} < 0|\text{NLD}(j,l)]$ (Table 6.1, sixth row, third column).

Investigating violations of monotonicity (Table 6.1, lower panel)

Only item 3 violated monotonicity. IRF $E(X_3|\theta)$ decreased either between $(-0.5; 0.5)$ or between $(0.5; 1.5)$. Discrimination parameters α_1, α_2, and α_3 had 13 levels (Table 6.6, second column), and $\alpha_4 = \alpha_5 = 1.5$. A pilot study showed that the effects of $\alpha_4 = \alpha_5 = 1.5$ on the sign of the covariances were negligible. The two θ-intervals and the 13^3 combinations of discrimination parameters produced $4,394$ combinations in total.

The three dependent variables were the proportion of negative values in (1) σ_{13} and σ_{23}, which estimated $P[\sigma_{jk} < 0|\text{VM}(j)]$ (Table 6.1, seventh row, first column); (2) $\sigma_{13|2}$ and $\sigma_{23|1}$, which estimated $P[\sigma_{jk|l} < 0|\text{VM}(j)]$ (Table 6.1, seventh row, second column); and (3) $\sigma_{13|R}$ and $\sigma_{23|R}$, which estimated $P[\sigma_{jk|R} < 0|\text{VM}(j)]$ (Table 6.1, seventh row, third column).

Results and Conclusions

Table 6.1 shows the results. Different parameter choices might have resulted in somewhat different results. We emphasize that the exact proportions of negative covariances were not of main interest but rather the knowledge that the proportions were considerable. Given a violation of a particular assumption, a considerable proportion of negative values suggests that the covariance can be used to identify the violation. Additional computations showed that for violations of monotonicity (i.e., decreases of the IRF) across at least two standard deviations of the latent variable, the proportion of negative values was considerable but we considered such large decreases unrealistic and ignored the results.

Chapter 7

Epilogue

The measurement model that is central in this thesis is the nonparametric item response model of monotone homogeneity. The model is based on three assumptions. Unidimensionality means that one latent variable drives the responses to the items. Local independence means that the item scores are independent conditional on the unidimensional latent variable. Finally, monotonicity means that the item response function (dichotomous item) and the item step response functions (polytomous items) are monotone nondecreasing functions of the latent variable. The monotone homogeneity model is important because it implies an ordinal scale for person measurement, and because many frequently used, parametric item response models are special cases of the more general, nonparametric monotone homogeneity model. The generality of the monotone homogeneity model implies that if the model does not fit the data collected by means of a test or a questionnaire, a wide array of more specific parametric item response models also show misfit to these data and need not be investigated anymore.

The psychometric part of the thesis deals with item selection and investigating local independence, and contributes to the topic of fit assessment of measurement models, in particular, the monotone homogeneity model. We discuss three topics. In Chapter 2, we discuss the bottom-up automated item selection procedure that Mokken (1971, pp. 190-193) proposed, and notice that the procedure suffers from two problems. The first problem is that, after completion of the procedure, Mokken's automated item selection procedure may have produced a scale that is inconsistent with the definition of a Mokken scale. The second problem is that, given the definition of a scale, Mokken's objective was to partition a set of items into one or more scales that contain as many items as possible but the procedure may not attain this goal. We proposed a genetic-algorithm version of the automated item selection procedure that solves the first problem and almost always produces the longest scale(s) given the definition of a Mokken scale.

We recommend using the genetic-algorithm version for item selection. The genetic-algorithm version can also be used in combination with the methodology Hemker et al. (1995) proposed to investigate the dimensionality structure of an item set. The methodology consists of running the automatic item selection procedure for varying scalability lower-bound values c. In doing this, several item-scalability values H_j are unavoidably close to at least one of the c values, thus inducing the first problem that sometimes a scale is inconsistent with the definition of a Mokken scale. This problem disappears when the genetic algorithm is used, and the resulting dimensionality structure is more trustworthy. A drawback of the genetic algorithm is that it is slow for tests containing more than 20 items and that it does not always find the largest possible item clusters. Brusco, Koehn, and Steinley (2011) proposed a branch-and-bound max-cardinality algorithm version of Mokken's automated item selection that more easily finds the global maximum.

In Chapter 5, we discussed the stability of Mokken's automated item selection algorithm and the genetic-algorithm version for varying sample sizes, and found the same results for both procedures. The results show that the difference between values of item-scalability coefficients H_j and scalability lower-bound c has the largest effect on the minimally required sample size. For H_j values close to c, one needs a larger sample size (N ranging from 1250 to 1750) than for H_j values that well exceed c (N ranging from 250 to 750). Thus, we recommend researchers to construct and retain items having absolutely high H_j values. When the Hemker et al. (1995) methodology for investigating the dimensionality structure of the item set is run using the automated item selection procedure, we recommend a minimum sample size of at least 1250 respondents. This methodology consists of different analysis steps. In each step, the item partitioning is evaluated for a particular value of lower bound c. Thus the methodology evaluates lower-bound values increasing in steps of .05 from 0 to, say, .65 (Hemker et al., 1995).

Therefore, it is unavoidable that for some of these analysis steps, lower bound c is close to one or more H_j values. This condition requires a large sample size to obtain stable results.

The assumption of local independence has received little attention in the context of investigating the goodness of fit of the monotone homogeneity model. One could argue that the use of the automated item selection procedure and its genetic-algorithm version already lead to unidimensional scales for which local independence probably holds, but it remains worthwhile to investigate local independence separately to find out whether this investigation produces additional and possibly interesting information about item structure that item selection does not pick up. In Chapter 6, we discussed how the property of conditional association can be used to investigate local independence. We proposed a new procedure, called CA procedure, that uses three special cases of conditional association to identify a set of locally independent items. The results show that CA procedure produces larger sets of items than Mokken's automated item selection procedure and method DETECT. We concluded that using Mokken's automated item selection procedure to produce scales in combination with CA procedure to analyze each separate scale to identify locally dependent items that might be removed from the scale, was not a fruitful strategy. More research is needed to show how CA procedure may be used to assess the fit of the monotone homogeneity model.

In chapters 3 and 4, we discussed real-data applications of the automated item selection procedure and its genetic-algorithm version. In Chapter 3, we investigated the dimensionality structure of the DS14 (Denollet, 2000, 2005), which is a questionnaire that measures distressed (Type D) personality. In Chapter 4, we investigated the HADS (Zigmond & Snaith, 1983), which is a questionnaire that measures anxiety and depression in physically ill hospital patients. Both questionnaires have become the topic of debate in the medical, clinical and health psychological literature. The debate about the DS14 concentrates on which of three different factor models that each describe the hierarchical structure of the DS14 best represents the true dimensionality structure. The Hemker et al. (1995) methodology that used the genetic-algorithm version of Mokken's automated item selection procedure supported the theoretical three-level hierarchical structure of the DS14. The debate about the HADS concentrates on the disagreement among different studies with respect to the dimensionality structure of the HADS. The Hemker et al. (1995) methodology revealed that the different dimensionality structures that different studies found represent different levels of a hierarchical structure.

Our analyses with respect to the DS14 and the HADS illuminated two problems that were both ignored or missed in the debates in the relevant literature. The first problem is that the outcomes of dimensionality research strongly depend on the method used and the population in which the research was done. By using the Hemker et al. (1995) methodology, we revealed the hierarchical structure of the data and showed that methods frequently reported in the relevant literature, which are Rasch analysis, exploratory factor analysis, and confirmatory factor analysis, each assessed only one particular level of the hierarchical HADS structure but never revealed all levels. The second problem seems to be that questionnaires such as the DS14 and the HADS lack a well-tested theoretical foundation, and are in strong need of such a foundation, which then is leading in instrument construction. In the absence of a theoretical foundation, the items in a questionnaire define the attribute that the questionnaire purports to measure. However, a well-tested theory should define the attribute and guide the operationalization into items (Sijtsma, 2012, in press). This strategy in which theory guides instrument construction avoids debates about the "true" dimensionality structure that in fact discuss methodological artifacts.

References

Adler, M., & Brodin, U. (2011). An IRT validation of the Affective Self Rating Scale. *Nordic Journal of Psychiatry, 65*, 396-402.

Alterman, A. I., Cacciola, J. S., Habing, B., & Lynch, K.G. (2007). ASI recent and lifetime summary indices based on nonparametric IRT models. *Psychological Assessment, 19*, 119-132.

Andrea, H., Bültmann, U., Beurskens, A. J., Swaen, G. M., Van Schayck, C. P., & Kant, I. J. (2004) Anxiety and depression in the working population using the HAD Scale–psychometrics, prevalence and relationships with psychosocial work characteristics. *Social Psychiatry and Psychiatric Epidemiology, 39*, 637-646.

Aquarius, A. E. A. M., Smolderen, K. G. E., Hamming, J. F., De Vries, J., Vriens, P. W., & Denollet, J. (2009). Type D personality and mortality in peripheral arterial disease: A pilot study. *Archives of Surgery, 144*, 728-733.

Balàzs, K., Hidegkuti, I., & De Boeck, P. (2006). Detecting heterogeneity in logistic regression models. *Applied Psychological Measurement, 30*, 322-344.

Bentler P. M. (1989). *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34*, 181-197.

Bergvik, S., Sørlie, T., Wynn, R., & Sexton, H. (2010). Psychometric properties of the Type D scale (DS14) in Norwegian cardiac patients. *Scandinavian*

Journal of Psychology, 51, 334-340.

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

Bernstein, I. H., & Teng, G. (1989). Factoring items and factor scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477.

Birnbaum, A. (1968). Some latent variable models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Bjelland, I., Dahl, A. A., Tangen Haug, T., & Neckelmann, D. (2002). The validity of the Hospital Anxiety and Depression Scale: An updated literature review. *Journal of Psychosomatic Research, 52*, 69-77.

Bollen K. A. (1989). *Structural equations with latent variables.* New York: John Wiley & Sons, Inc.

Brennan, C., Worrall-Davies, A., McMillan, D., Gilbody, S., & House, A. (2010). The Hospital Anxiety and Depression Scale: A diagnostic meta-analysis of case-finding ability. *Journal of Psychosomatic Research, 69*, 371-378.

Browne, M. W, & Cudeck R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models*, pp. 136-162. Newbury Park, CA: Sage.

Brusco, M. J., Koehn, H. F., & Steinley, D. (2011). *A branch-and-bound max-cardinality algorithm for exploratory Mokken scale analysis.* Paper presented at the International Meeting of the Psychometric Society, Hong Kong, July, 2011.

Caci, H., Bayle, F. J., Dossios, C., Robert, P., & Boyer, P. (2003). How does the Hosporal Anxiety and Depression Scale measure anxiety and depression in healthy subjects? *Psychiatric Research, 118*, 89-99.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Chuah, S. C., Drasgow, F., & Leucht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education, 19*, 241-255.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Earlbaum.

Cohen, M. P. (2005). Sample size considerations for multilevel surveys. *International Statistical Review, 73*, 279-287.

Cosco, T. D., Doyle, F., Ward, M., & McGee, H. (2012). Latent structure of the Hospital Anxiety and Depression Scale: A 10-year systematic review. *Journal of Psychosomatic Research, 72*, 180-184.

Cosco, T. D., Doyle, F., Watson, R., Ward, M., & McGee, H. (2012). Mokken scaling analysis of the Hospital Anxiety and Depression Scale in individuals with cardiovascular disease. *General Hospital Psychiatry, 34*, 167-172.

Coyne, J. C., Jaarsma, T., Luttik, M. L., Van Sonderen, E., Van Veldhuisen, D. J., & Sanderman, R. (2011). Lack of prognostic value of Type D personality for mortality in a large sample of heart failure patients. *Psychosomatic Medicine, 73*, 557-562.

Coyne, J. C., & Van Sonderen, E. (2012). No further research needed: Abandoning the Hospital and Anxiety Depression Scale (HADS). *Journal of Psychosomatic Research, 72*, 173-174.

Dannemann, S., Matschke, K., Einsle, F., Smucker, M. R., Zimmermann, K., Joraschky, P., et al. (2010). Is Type D a stable construct? An examination of Type D personality in patients before and after cardiac surgery. *Journal of Psychosomatic Research, 69*, 101-109.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*, 155-170.

De Jonge, P., Denollet, J., Van Melle, J. P., Kuyper, A., Honig, A., Schene, A. H., & Ormel, J. (2007). Associations of Type D personality and depression with somatic health in myocardial infarction patients. *Journal of Psychosomatic Research, 63*, 477-482.

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

Denollet, J. (2000). Type D personality: a potential risk factor refined. *Journal of Psychosomatic Research, 49*, 255-266.

Denollet, J. (2005). DS14: standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine, 67*, 89-97.

Denollet, J., Schiffer, A. A., & Spek, V. (2010). A general propensity to psychological distress affects cardiovascular outcomes: evidence from research on the Type D (distressed) personality profile. *Circulation: Cardiovascular Quality and Outcomes, 3*, 546-557.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309-326.

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). A confirmatory factor analysis of the Hospital Anxiety and Depression Scale: Comparing empirically and theoretically derived structures. *British Journal of Clinical Psychology, 39*, 79-94.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Emons, W. H. M., Meijer, R. R., & Denollet, J. (2007). Negative affectivity and social inhibition in cardiovascular disease: evaluating Type D personality and its assessment using item response theory. *Journal of Psychosomatic Research, 63*, 27-39.

Emons, W. H. M., Sijtsma, K., & Pedersen, S. S. (2012). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in cardiac patients: comparison of Mokken scale analysis and factor analysis. *Assessment* doi:10.1177/1073191110384951.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12*, 105-120.

Ferguson, E., Williams, L., O'Connor, R. C., Howard, S., Hughes, B. M., Johnston, D. W., Allan, J. L., O'Connor, D. B., Lewis, C. A., Grealy, M. A., & O'Carroll, R. E. (2009). A taxometric analysis of Type D personality. *Psychosomatic Medicine, 71*, 981-986.

Friedman, S., Samuelian, J. C., Lancrenon, S., Even, C., & Chiarelli, P. (2001). Three-dimensional structure of the Hospital Anxiety and Depression Scale in a large French primary care population suffering from major depression. *Psychiatric Research, 104*, 247-257.

Gibbons, C. J., Mills, R. J., Thornton, E. W., Ealing, J., Mitchell, J. D., Shaw, P. J., et al. (2011). Rasch analysis of the Hospital Anxiety and Depression Scale (HADS) for use in motor neurone disease. *Health and Quality of Life Outcomes, 9*, 82-89.

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: Springer.

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

Grande, G., Glaesmer, H., & Roth, M. (2010). The construct validity of social inhibition and the Type D taxomony. *Journal of Health Psychology, 15*, 1103-1112.

Grande, G., Romppel, M., Glaesmer, H., Petrowski, K., & Herrmann-Lingen, C. (2010). The Type D scale (DS14): Norms and prevalence of Type D personality in a population-based representative sample in Germany. *Personality and Individual Differences, 48*, 935-939.

Grande, G., Romppel, M., Versper, J. M., Schubman, R., Glaesmer, H., & Herrmann-Lingen, C. (2011). Type D personality and all-cause mortality in cardiac patients – Data from a German cohort study. *Psychosomatic Medicine, 73*, 548-556.

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

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499-510.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*, 171-186.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Hausteiner, C., Klupsch, D., Emeny, R., Baumert, J., & Ladwig, K.H. (2010). Clustering of negative affectivity and social inhibition in the communitiy: prevalence of Type D personality as a cardiovascular risk marker. *Psychosomatic Medicine, 72*, 163-171.

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between groups. *Journal of Educational and Behavioral Statistics, 24*, 70-93.

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.

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.

Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66*, 487-506.

Herrmann C. (1997). International experiences with the Hospital Anxiety and Depression Scale – A review of validation data and clinical results. *Journal of Psychosomatic Research, 42*, 17-41.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement, 65*, 202-226.

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hubert, M. & Vandervieren, E. (2008). An adjusted box plot for skewed distributions. *Computational Statistics and Data Analysis, 52*, 5186-5201.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Hunt-Shanks, T., Blanchard, C., Reid, R., Fortier, M., & Cappelli M. (2010). A psychometric evaluation of the Hospital Anxiety and Depression Scale in cardiac patients: Addressing factor structure and gender invariance. *British Journal of Health Psychology, 15*, 97-114.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59*, 77-79.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10*, 128-141.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56*, 255-278.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65-81.

Kimber, A. C. (1990). Exploratory data analysis for possibly censored data from skewed distributions. *Applied Statistics, 39*, 21-30.

Krosnik, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2012). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. Manuscript submitted for publication.

Kupper, N., & Denollet, J. (2007). Type D personality as a prognostic factor in heart disease: assessment and mediating mechanisms. *Journal of Personality Assessment, 89*, 65-276.

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.

Lim, H.E., Lee, M., Ko, Y., Park, Y., Joe, S., Kim, Y., Han, S., Lee, H., Pedersen, S. S., & Denollet, J. (2011). Assessment of the Type D personality construct in the Korean population: a validation study of the Korean DS14. *Journal of Korean Medical Science, 26*, 116-123.

Lubke G., & Muthén B. (2004). Factor-analyzing Likert scale data under the assumption of multivariate normality complicates a meaningful comparison of observed groups or latent classes. *Structural Equation Modeling, 11*, 514-534.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong S. (2002). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 11*, 611-637.

Martin, C. R., Thompson, D. R., & Barth J. (2008). Factor structure of the Hospital Anxiety and Depression Scale in coronary heart disease patients in three countries. *Journal of Evaluation in Clinical Practice, 14*, 281-287.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354-368.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283-298.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*, 293-299.

Michalewicz, Z. (1996). *Genetic algorithms + data structures = evolution programs.* New York: Springer.

Miller, A. (2002). *Subset selection in regression.* New York: Chapman and Hall.

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

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

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

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48*, 49-72.

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

Moorey, S., Greer, S., Watson, M., Gorman, C., Rowden, L., Tunmore, R., et al. (1991). The factor structure and factor stability of the Hospital Anxiety and Depression Scale in patients with cancer. *British Journal of Psychiatry, 158*, 255-259.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*, 159-168.

Mykletun, A., Stordal, E., & Dahl, A. A. (2001). Hospital anxiety and depression (HAD) scale: factor structure, item analyses and internal consistency in a large population. *British Journal of Psychiatry, 179*, 540-544.

Nandakumar, R., & Stout W. F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18*, 41-68.

Norton, S., Sacker, A., & Done, J. (2012). Further research needed: A comment on Coyne and van Sonderen's call to abandon the Hospital Anxiety and Depression Scale. *Journal of Psychosomatic Research, 73*, 75-76.

O'Dell, K. R., Masters, K. S., Spielmans, G. I., & Maisto, S. A. (2011). Does Type D personality predict outcomes among patients with cardiovascular disease? A meta-analytic review. *Journal of Psychosomatic Research, 71*, 199-206.

Olsson U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 481-500.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*, 1-18.

Prince, M., Acosta, D., Ferri, C. P., Guerra, M., Huang, Y., Jacob, K. S., et al. (2010). A brief dementia screener suitable for use by non-specialists in resource poor settings — the cross-cultural derivation and validation of the brief Community Screening Instrument for Dementia. *International Journal of Geriatric Psychiatry, 26*, 899-907.

Ramsay, J. C. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287-297.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Revelle, W. (2012). psych: Procedures for Personality and Psychological Research (Version 1.1.12). Northwestern University: Evanston. Retrieved from http://cran.r-project.org/web/packages/psych/index.html

Rosenbaum, P. R. (1984). Testing conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

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

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1-30.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Schiffer, A. A. J., Smith, O. R. F., Pedersen, S. S., Widdershoven, J. W., & Denollet, J. (2010). Type D personality and cardiac mortality in patients with chronic heart failure. *International Journal of Cardiology, 142*, 230-235.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 77*, 4-20.

Sijtsma, K. (in press). Psychological measurement between physics and statistics. *Theory & Psychology.*

Sijtsma, K., Emons, W. H. M., Bouwmeester, S., Nyklicek, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality of life scales and its application to the World Health Organization Quality of Life scale (WHOQOL-Bref). *Quality of Life Research, 17*, 275-290.

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: Elsevier, North Holland.

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

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Testing, 18*, 237-259.

Spindler, H., Kruse, C., Zwisler, A., & Pedersen, S. S. (2009). Increased anxiety and depression in Danish cardiac patients with a Type D personality: cross-validation of the Type D scale (DS14). *International Journal of Behavioral Medicine, 16*, 98-107.

Straat, J. H. (2012). CAprocedure (Version 1.0) [Computer software]. Tilburg, The Netherlands: Tilburg University. Retrieved from http://cran.r-project.org/web/packages/CAprocedure/index.html

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2012a). Using conditional association to identify locally independent items sets. Manuscript submitted for publication.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2012b). Multi-method analysis of the internal structure of the Type D Scale-14 (DS14). *Journal of Psychosomatic Research, 72*, 258-265.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (in press). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification.*

Svansdottir, E., Karlsson, H. D., Gudnason, T., Olason, D. T., Thorgilsson, H., Sigtryggsdottir, U., Sijbrands, E. J., Pedersen, S. S., & Denollet J. (2011). Validity of Type D personality in Iceland: association with disease severity and risk markers in cardiac patients. *Journal of Behavioral Medicine, 35*, 155-166.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics.* Needham Heights, MA: Allyn & Bacon, Inc.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159-203.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3-24.

Van Abswoude, A. A. H., Vermunt, J. K., & Hemker, B. T. (2007). Assessing dimensionality by maximizing H coefficient-based objective functions. *Applied Psychological Measurement, 31*, 308-330.

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), 1-19.

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 Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika, 73*, 183-208.

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 Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*, 380-392.

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

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.

Van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis.* Thousand Oaks, CA: Sage.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231-251.

Williams, L., Curren, C., & Bruce G. (2011). Are alexithymia and Type D personality distinct or overlapping constructs? A confirmatory factor analysis of the Toronto alexithymia and Type D scales. *Personality and Individual Differences, 51*, 683-686.

Wismeijer, A. A. J., Sijtsma, K., Van Assen, M. A. L. M., & Vingerhoets, A. J. J. M. (2008). A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. *Journal of Personality Assessment, 90*, 323-334.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three- parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yu, D. S. F., Thompson, D. R., Yu, C. M., Pedersen, S. S., & Denollet J. (2010). Validating the Type D personality construct in Chinese patients with coronary heart disease. *Journal of Psychosomatic Research, 69*, 111-118.

Yu, X., Zhang, J., & Liu, X. (2008). Application of the Type D Scale (DS14) in Chinese coronary heart disease patients and healthy controls. *Journal of Psychosomatic Research, 65*, 595-601.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67*, 361-370.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72*, 69-91.

Zhang, J., & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-152.

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

Zohar, A. H., Denollet, J., Lev Ari, L., & Cloninger, C. R. (2011). The psychometric properties of the DS14 in Hebrew and the prevalence of Type D (distressed) personality in Israeli adults. *European Journal of Psychological Assessment, 24*, 74-281.

Summary

In Chapter 1, we introduced a general, nonparametric item response theory (IRT) model, known as the monotone homogeneity model. The monotone homogeneity model implies the measurement of individuals on an ordinal scale. If the monotone homogeneity model fits the data collected by means of a test or a questionnaire, the test or questionnaire is appropriate for ordinal measurement. Examples are personnel selection problems in which the applicants are selected that have the highest scores on, for example, an intelligence test, and the clinical context in which patients are selected for treatment that have the highest scores on, for example, an anxiety test.

The monotone homogeneity model assumes that the relation between response probabilities on individual items and the latent variable, also known as the item response function, is subject to order restrictions. The order restriction is that the item response function is monotone nondecreasing. Parametric IRT models assume a monotone, parametric item response function, such as the logistic, and are special cases of the monotone homogeneity model. Hence, misfit of monotone homogeneity implies that parametric IRT models that are special cases also show misfit. Monotone homogeneity is evaluated by means of observable consequences of the model, such as scalability coefficients that have values the model restricts to the interval $[0, 1]$, and conditional association. In this thesis, we discuss item selection methods that use scalability coefficients to identify clusters of items that measure one latent variable, and we discuss the property of conditional association to assess the assumption of local independence of the items in the test or the questionnaire. We also assess the usefulness of the methods for investigating the goodness of fit of the model of monotone homogeneity model to several real-data sets.

In Chapter 2, we discussed the bottom-up, automated item selection procedure that Mokken (1971, pp. 190-193) proposed. The procedure selects item sets that satisfy the definition of a Mokken scale. The procedure is known to suffer from two problems. First, due to its bottom-up character the procedure sometimes selects items that are inconsistent with the definition of a Mokken scale. Second, the procedure sometimes produces Mokken scales that are consistent with the definition of a scale but that are not optimal given the objective that selected Mokken scales should contain as many items as possible given the definition of a Mokken scale. Hence, the procedure may sometimes produce a local maximum.

We proposed a genetic algorithm that avoids the first problem and often avoids the second problem. Thus, each scale identified is a Mokken scale and often it contains the maximum number of items the side conditions of the selection problem allow. We used a simulation study to compare the automated item selection procedure and its genetic-algorithm version with respect to two questions: Which item selection method best produces item clusters consistent with the goal to find Mokken scales of maximum length, formalized as an objective function in an optimization problem; and which item selection method best retrieves the true dimensionality of simulated data. We found that the genetic-algorithm version of the item selection procedure performs better than the traditional bottom-up version with respect to both questions, in particular if the item-scalability values were close to the lower bound criterion for admitting items to a scale. We used the two item selection procedures to analyze the data collected by means of the communality scale of the Adjective Checklist, and found that the genetic-algorithm version resulted in one scale containing seven items and the bottom-up version resulted in two scales containing four items. We concluded that the genetic-algorithm version provides an improvement of the bottom-up version of Mokken's automated item selection procedure.

In Chapter 3, we used the bottom-up version and the genetic-algorithm version to investigate the dimensionality structure of the Type-D Scale 14 (DS14). The dimensionality structure of the DS14 is subject to debate. The DS14 has a three-level hierarchical structure: At the high level, the DS14 measures type D personality; at the medium level, the DS14 measures the personality traits of negative affectivity and social inhibition. At the low level, negative affectivity encompasses the attributes of dysphoria, anxiety, and irritability, and social inhibition encompasses the attributes of social discomfort, reticence, and lack of social poise. In the literature, three models that describe the internal structure of the DS14 are discussed. In addition to Mokken's automated item selection procedure and its genetic-algorithm version, we used exploratory factor analysis and confirmatory factor analysis to investigate which of the three models best describes the internal structure of the DS14. The results supported the three-level hierarchical model as a conceptual model for Type D personality, but only the genetic-algorithm version of Mokken's automated item selection procedure identified the expected six low-level scales. We concluded that the item structure of the DS14 reflects the theoretical three-level hierarchy.

In Chapter 4, we used Mokken's automated item selection procedure and the genetic-algorithm version to analyze data collected by means of the Hospital Anxiety and Depression Scale (HADS). A recent review found that researchers using different dimensionality assessment methods and investigating the HADS dimensionality in different populations produced different dimensionality structures. We showed that the different dimensionality structures are due to a methodological artifact that is effective as a result of the HADS' hierarchical structure. Using a methodology proposed by Hemker et al. (1995), both item selection procedures showed the different levels of the hierarchical structure of the HADS. Based on different analysis steps (Hemker et al., 1995), the two item selection procedures identified one scale at the high level, measuring psychological distress, and two scales at the low level, measuring anxiety and depression, respectively. Confirmatory Mokken scale analysis showed a partitioning of the 14 HADS items into three scales. We concluded that the Hemker et al. (1995) methodology is suited to identify the different levels of a hierarchical trait structure, but that several other psychometric methods, such as factor analysis, only identify one level of the hierarchy. Thus, other methods may provide misleading information because they miss the hierarchical structure altogether.

In Chapter 5, we investigated minimum sample-size requirements for the bottom-up and genetic-algorithm versions of Mokken's item selection procedure. In practice, researchers reported having used samples ranging from 133 to 15,022 respondents. We investigated the relevant factors that determine the sample size minimally required to assign items to the correct Mokken scales. We found that the most relevant factor was the difference between the item-scalability value and the lower bound value it must minimally attain for the item to be admitted to the scale. If the item-scalability value corresponded to this lower bound value, which suggests that the item is a borderline-case for selection, the minimally required sample size for accurate item selection ranged from 1250 to 1750 respondents. If the item-scalability values well exceeded the lower bound, which made them much more appropriate candidates for selection, the minimally required sample size for accurate item selection ranged from 250 to 500 respondents. Hence, for items that have item-scalability values well above the minimum value required for admittance to the scale, relatively small sample sizes suffice for accurate assignment of items to Mokken scales.

In Chapter 6, we investigated the observable property of the monotone homogeneity model known as conditional association (CA). We specialized conditional association into three covariances that must be nonnegative if the monotone homogeneity model is the true model for the data; investigated in a computational study whether the signs of the covariances were helpful to identify violations of the monotone homogeneity model; and used the results of the computational study to define the "CA procedure". CA procedure aims at identifying violations of local independence and removing items from locally dependent item pairs so as to obtain a locally independent set of items. We simulated data which we used to investigate whether CA procedure, Mokken's automated bottom-up item selection procedure and program DETECT were capable of identifying violations of the local independence assumption and violations of the monotonicity assumption. We found that CA procedure produced larger locally independent item sets than Mokken's automated item selection procedure and DETECT, but also that CA procedure was not sensitive for violations of monotonicity. We used CA procedure to analyze the DS14 data set that was investigated in Chapter 3, and found that CA procedure identified locally dependent items in two low-level scales. We concluded that CA procedure is sensitive to violations of local independence and specific for items that are consistent with the monotone homogeneity model.

Chapter 7 gives an overview of the results of this thesis. We discussed ideas for future research that the results suggested.

Samenvatting

In hoofdstuk 1 introduceren wij een algemeen nonparametrisch item-responstheorie (IRT) model, dat bekendstaat als het model van monotone homogeniteit. Het model van monotone homogeniteit laat meting van personen op een ordinale schaal toe. Indien het model van monotone homogeniteit bij de door middel van een test of vragenlijst verzamelde data past, is de test of vragenlijst daarmee geschikt voor ordinale meting. Voorbeelden zijn personeelsselectieproblemen waarbij de sollicitanten worden geselecteerd die de hoogste score hebben op, bijvoorbeeld, een intelligentietest, en de klinische context waarin patiënten voor een behandeling worden geselecteerd op basis van de hoogste score op, bijvoorbeeld, een test die angststoornis meet.

Het model van monotone homogeniteit veronderstelt dat de relatie tussen de responskansen op een item en de latente variabele, die bekendstaat als de itemresponsfunctie, een gerestricteerde ordening heeft. De orderestrictie is dat de itemresponsfunctie monotoon niet-dalend is. Parametrische IRT modellen veronderstellen een monotone, parametrische itemresponsfunctie, bijvoorbeeld een logistische functie, en zijn daarom speciale gevallen van het model van monotone homogeniteit. Dit betekent dat wanneer het model van monotone homogeniteit niet bij de data past, de parametrische modellen de data evenmin adequaat beschrijven. De passing van het model van monotone homogeniteit bij de data wordt onderzocht door middel van observeerbare eigenschappen van het model, zoals schaalbaarheidscoëfficiënten die onder monotone homogeniteit waarden hebben in het interval [0,1] en conditionele associatie. In dit proefschrift, bespreken wij itemselectiemethoden die gebruikmaken van schaalbaarheidscoëfficiënten om itemclusters die dezelfde latente variabele meten te identificeren, en ook bespreken wij de eigenschap van conditionele associatie waarmee bij de items in een test of vragenlijst de aanname van lokale onafhankelijkheid kan worden onderzocht. Wij onderzoeken ook de bruikbaarheid van deze methoden voor het nagaan van de passing van het model van monotone homogeniteit bij enkele echte datasets.

In hoofdstuk 2 beschouwen wij de door Mokken (1971, pp. 190-193) voorgestelde geautomatiseerde bottom-up itemselectieprocedure. De procedure selecteert verzamelingen van items die voldoen aan de definitie van een Mokkenschaal. De procedure kent twee problemen. Ten eerste selecteert de procedure, omdat het een bottom-up proces betreft, soms items die inconsistent zijn met de definitie van een Mokkenschaal. Ten tweede vindt de procedure soms een Mokkenschaal die wel voldoet aan de definitie van een schaal, maar die niet optimaal is gegeven het doel om Mokkenschalen te vinden die zoveel mogelijk items bevatten. De procedure vindt soms dus een lokaal maximum.

Wij stellen een genetisch algoritme voor dat het eerste probleem gegarandeert vermijdt en het tweede probleem bijna altijd vermijdt. Iedere geïdentificeerde schaal is dus een Mokkenschaal en bevat bijna altijd het grootst mogelijke aantal items die de randvoorwaarden van het selectieprobleem toestaan. Wij deden een simulatiestudie om Mokken's geautomatiseerde itemselectie procedure te vergelijken met de genetisch-algoritme variant met betrekking tot twee onderzoeksvragen: Welke itemselectiemethode vindt itemclusters die het best overeenkomen met het doel om Mokkenschalen met zoveel mogelijk items te vinden, geformaliseerd als een doelfunctie voor een optimalisatieprobleem; en welke itemselectiemethode vindt het beste de ware dimensionaliteit van de gesimuleerde data. Wij vonden met betrekking tot beide onderzoeksvragen dat de itemselectieprocedure gebaseerd op een genetisch algoritme beter presteert dan de traditionele bottom-up variant, en dit gebeurt vooral als de waarden voor itemschaalbaarheid ongeveer even groot zijn als het ondergrenscriterium voor het toelaten van items tot een schaal. Wij gebruikten beide varianten van Mokken's itemselectieprocedure om data die waren verzameld met de communaliteitsschaal van de Adjective Checklist te analyseren. Het genetisch algoritme vond een schaal die zeven items bevat en de bottom-up variant vond twee schalen die ieder vier items bevatten. Wij concludeerden dat de genetisch-algoritme variant een verbetering biedt van de bottom-up variant van Mokken's geautomatiseerde itemselectieprocedure.

In hoofdstuk 3 gebruiken wij de twee varianten van Mokken's geautomatiseerde itemselectieprocedure om de dimensionaliteitsstructuur van de Type-D Scale 14 (DS14) te onderzoeken. De dimensionaliteitsstructuur van de DS14 wordt in de vakliteratuur betwist. Theoretisch heeft de DS14 een hiërarchische structuur met drie niveaus: Op het hoge niveau meet de DS14 type-D persoonlijkheid. Op het middenniveau meet de DS14 de persoonlijkheidstrekken van negatieve affectiviteit en sociale inhibitie. Op het lage niveau bestaat negative affectiviteit uit de attributen dysforie, bezorgdheid, en prikkelbaarheid en sociale inhibitie uit de attributen sociaal ongemak, geslotenheid, en een gebrek aan zelfverzekerdheid in sociale situaties. In de literatuur worden drie modellen besproken die de interne structuur van de DS14 beschrijven. Naast Mokken's geautomatiseerde itemselectieprocedure en de genetisch-algoritme variant, gebruikten wij exploratieve factoranalyse en confirmatorische factoranalyse om te onderzoeken welk model de interne structuur van de DS14 het best beschrijft. De resultaten ondersteunden de theoretische hiërarchische structuur met drie niveaus als een geschikt conceptueel model voor Type D persoonlijkheid, maar alleen de genetisch-algoritme variant identificeerde de zes schalen op het lage niveau. Wij concludeerden dat de itemstructuur van de DS14 de theoretische hiërarchische structuur met drie niveaus weerspiegelt.

In hoofdstuk 4 gebruiken wij Mokken's geautomatiseerde itemselectieprocedure en de genetisch-algoritme versie om door middel van de Hospital Anxiety and Depression Scale (HADS) verzamelde data te analyseren. In een recent overzichtsartikel werd geconcludeerd dat onderzoekers, die verschillende methoden voor dimensionaliteitsonderzoek gebruikten en die de dimensionaliteit van de HADS onderzochten in verschillende populaties, verschillende dimensionaliteitsstructuren vonden voor de HADS. Wij toonden aan dat de verschillende dimensionaliteitsstructuren kunnen worden opgevat als een methodologisch artefact. In feite heeft de HADS een hiërarchische structuur. Door gebruik te maken van de door Hemker et al. (1995) voorgestelde methodologie, lieten beide geautomatiseerde itemselectieprocedures verschillende niveaus van de hiërarchische structuur van de HADS zien. Op basis van verschillende stappen in de analyse (Hemker et al., 1995), vonden beide itemselectieprocedures een schaal op het hoge niveau, die psychologisch leed meet, en twee schalen op het lage niveau, die respectievelijk angststoornis en depressie meten. Confirmatorische Mokkenschaalanalyse bevestigde een indeling van de 14 HADS items in drie schalen. Wij concludeerden dat de methodologie van Hemker et al. (1995) de verschillende niveaus van een hiërarchische structuur in kaart kan brengen, terwijl diverse andere psychometrische methoden, waaronder factoranalyse, slechts een niveau van een hiërarchische structuur laten zien. Vanwege het negeren van de hiërarchische structuur kunnen deze andere methoden misleidende informatie geven.

Hoofdstuk 5 behelst een onderzoek naar de minimaal vereiste steekproefomvang voor de bottom-up en de genetisch-algoritme varianten van Mokken's geautomatiseerde itemselectieprocedure. In de praktijk passen onderzoekers Mokkenschaalanalyse toe op steekproeven in grootte variërend van 133 tot 15.022 respondenten. Wij onderzochten de relevante factoren die van invloed zijn op de minimaal vereiste steekproefomvang die nodig is om de items aan de juiste schalen toe te wijzen. De belangrijkste factor was het verschil tussen de waarde van de gevonden itemschaalbaarheidscoëfficiënt en de waarde die een itemschaalbaarheidscoëfficiënt minimaal moet hebben voordat het item in een schaal kan worden geselecteerd. Indien de waarde van itemschaalbaarheid overeenkwam met de waarde van de ondergrens, vonden wij dat de minimaal vereiste steekproefgrootte voor nauwkeurige itemselectie tussen 1250 en 1750 respondenten lag. Indien de waarde van de itemschaalbaarheidscoëfficiënt beduidend groter was dan de waarde voor de ondergrens, waardoor de items veel geschikter zijn voor selectie, vonden wij dat de minimaal vereiste steekproefgrootte voor nauwkeurige itemselectie tussen 250 en 500 respondenten lag. Wij concludeerden dat relatief kleine steekproefgroottes voldoen om items nauwkeurig aan Mokkenschalen toe te kennen indien de waarden van de itemschaalbaarheidscoëfficiënt duidelijk boven de ondergrens liggen.

Hoofdstuk 6 betreft onderzoek naar de observeerbare eigenschap van het model van monotone homogeniteit die bekendstaat als conditionele associatie. Wij selecteerden drie covarianties die speciale gevallen zijn van conditionele associatie. De drie covarianties zijn niet-negatief indien het model van monotone homogeniteit het correcte model is voor de data. In een rekenkundige studie onderzochten wij of covarianties inderdaad negatief zijn als het model niet bij de data past en hoe geschikt die eigenschap dus is om schendingen van monotone homogeniteit te identificeren. Op basis van de resultaten van dit onderzoek definieerden wij de “CA procedure”. CA procedure heeft tot doel om schendingen van lokale onafhankelijkheid te identificeren en items te verwijderen die deel uitmaken van een lokaal afhankelijk itempaar. Het resultaat is een verzameling van lokaal onafhankelijke items. Wij simuleerden data die wij gebruikten om te onderzoeken of CA procedure, Mokken’s geautomatiseerde bottom-up itemselectieprocedure en DETECT in staat waren om schendingen van lokale onafhankelijkheid en monotonie in de data te identificeren. De resultaten lieten zien dat CA procedure grotere verzamelingen van lokaal onafhankelijke items vond dan Mokken’s geautomatiseerde itemselectieprocedure en DETECT, maar niet gevoelig was voor schendingen van monotonie. Wij pasten CA procedure toe op de DS14 data die wij onderzochten in hoofdstuk 3 en vonden dat CA procedure lokaal afhankelijke items in twee schalen op het lage niveau identificeerde. Wij concludeerden dat CA procedure gevoelig is voor schendingen van lokale onafhankelijkheid en specifiek voor items die consistent zijn met het model van monotone homogeniteit.

Hoofdstuk 7 geeft een overzicht van de gevonden resultaten in dit proefschrift. Naar aanleiding van deze resultaten bespreken wij ideeën voor toekomstig onderzoek.

Dankwoord

Na vier prettige jaren in Tilburg is het schrijven van mijn proefschrift afgerond. Op deze plaats wil ik graag iedereen bedanken die mij in deze periode advies, begeleiding en steun heeft gegeven. Allereerst, dank ik mijn promotor Klaas Sijtsma en mijn copromotor Andries van der Ark. Jullie hebben mij de mogelijkheid geboden om dit proefschrift te schrijven. Ik heb de afgelopen vier jaar op een fijne manier met jullie samengewerkt en ik ben jullie dankbaar voor de goede begeleiding die uiteindelijk tot dit mooie resultaat heeft geleid.

Ik wil Susanne Pedersen, en Wobbe Zijlstra bedanken voor het beschikbaar stellen van de data sets voor de hoofdstukken 3 en 4. Johan Denollet en Wilco Emons, ik ben jullie dankbaar voor de discussies die ik met jullie heb gevoerd over eerdere versies van deze twee hoofdstukken. Met Rudy Ligtvoet, Iris Smits, en Jesper Tijmstra heb ik regelmatig aangename en diepgaande discussies gevoerd over Mokkenschaalanalyse. Deze discussies hebben mij geholpen om een beter proefschrift te schrijven. Marieke Timmermans en Liesbeth Bluekens, ik ben jullie dankbaar voor de fijne gesprekken en alle hulp die het schrijven van mijn proefschrift heeft vereenvoudigd.

Ik ben ook alle collega's bij het MTO departement dankbaar voor de prettige werksfeer en gezellige lunches. In het bijzonder dank ik Stéfanie André, Marcel van Assen, Margot Bennink, Maarten Kampert, Miloš Kankaraš, Natalia Kieruj, Ruud van Keulen, Marie-Anne Mittelhaeuser, Meike Morren, en Daniël van der Palm. Naast gesprekken over ons werk heb ik met jullie fijne boswandelingen, vermakelijke congresbezoeken, biertjes bij Kandinsky en mooie reizen gedeeld. Natuurlijk dank ik ook mijn paranimfen Peter Kruyen en Aafke Raaijmakers. Wij zijn tegelijkertijd begonnen aan het avontuur van het schrijven van onze proefschriften en ik ben blij dat jullie tot het eind achter mij staan.