Nonparametric item response theory and Mokken scale analysis, with relations to latent class models and cognitive diagnostic models

van der Ark, L.A.; Rossi, G.; Sijtsma, K.

DOI
10.1007/978-3-030-05584-4_2

Publication date
2019

Document Version
Final published version

Published in
Handbook of diagnostic classification models

Chapter 2
Nonparametric Item Response Theory and Mokken Scale Analysis, with Relations to Latent Class Models and Cognitive Diagnostic Models

L. Andries van der Ark, Gina Rossi, and Klaas Sijtsma

Abstract As the focus of this chapter, we discuss nonparametric item response theory for ordinal person scales, specifically the monotone homogeneity model and Mokken scale analysis, which is the data-analysis procedure used for investigating the compliance between the monotone homogeneity model and data. Next, we discuss the unrestricted latent class model as an even more liberal model for investigating the scalability of a set of items, producing nominal scales, but we also discuss an ordered latent class model that one can use to investigate assumptions about item response functions in the monotone homogeneity model and other nonparametric item response models. Finally, we discuss cognitive diagnostic models, which are the core of this volume, and which are a further deepening of latent class models, providing diagnostic information about the people who responded to a set of items. A data analysis example, using item scores of 1210 respondents on 44 items from the Millon Clinical Multiaxial Inventory III, demonstrates how the monotone homogeneity model, the latent class model, and two cognitive diagnostic models can be used jointly to understand one's data.

L. A. van der Ark (✉)
Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands
e-mail: L.A.vanderArk@uva.nl

G. Rossi
Research Group Personality and Psychopathology, Vrije Universiteit Brussel, Brussels, Belgium
e-mail: grossi@vub.be

K. Sijtsma
Department of Methodology and Statistics, TSB, Tilburg University, Tilburg, The Netherlands
e-mail: k.sijtsma@tilburguniversity.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_2

2.1 Introduction

Nonparametric item response theory (IRT; Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & van der Ark, 2017; van Schuur, 2011), which is the focus of this chapter, is a set of psychometric measurement models implying ordinal person measurement using the sum score on a set of items. The sum score provides a summary of the ability or the trait the items measure but does not inform us about sub-attributes needed for one or more subsets of items in the test. Latent class models (LCMs) aim at classifying persons in unordered or nominal classes based on the set of scores on the items that comprise the test (Hagenaars & McCutcheon, 2002; Heinen, 1996); in LCMs the sum score does not play a role. Although parametric LCMs exist (e.g., Goodman, 1974; Formann & Kohlmann, 2002), the typical LCM is nonparametric. Nonparametric IRT and LCMs may appear to be different, but Croon (1990) and Vermunt (2001) showed how imposing an ordering of the latent classes renders LCM analysis suitable for assessing the fit of a nonparametric IRT model to the data. This application identifies an interesting link between LCMs and nonparametric IRT. Haertel (1989) argued that LCMs are stepping-stones to what later became known as cognitive diagnostic models (CDMs; Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010; von Davier, 2010, 2014). CDMs constitute the core of this volume. The models classify persons based on a set of skills, abilities or attributes the researcher hypothesizes persons need to solve a set of items. In this sense, CDMs provide information about persons' proficiency to solve particular sets of items that is finer-grained than the summary sum score that nonparametric IRT models provide. Junker and Sijtsma (2001) demonstrated how nonparametric IRT and CDMs are related.

The three types of models—nonparametric IRT models, LCMs, and CDMs—have in common that they rely on assumptions about the data that are sufficiently strong to classify or order persons. On the other hand, the assumptions are not so demanding that they all too easily lead to the rejection of items that may not satisfy stronger models but contribute to reliable person classification or person ordering. In this sense, the models are "item-preserving", asking as little as possible from the data and still being able to measure people's attributes at the nominal and ordinal levels (Michell, 1999; Stout, 2002). Although there is much to say about the relationships between the three types of models and much more work that remains to be done to further unravel these relationships, given the goal of this volume we will only briefly discuss the models' assumptions and main ideas. We focus on nonparametric IRT, specifically the version Mokken (1971) introduced and many other researchers further developed. We conclude the chapter discussing a real-data example of a Mokken scale analysis, with brief reference to LCM and CDM data analysis.

2.2 Three Types of Models, Their Properties, and Their Relations

We focus on tests and questionnaires that use a set of K items to measure an attribute, such as a cognitive ability or a personality trait (psychology), an educational achievement and skills (educational measurement), quality of life or pain experience (health sciences), an attitude (sociology) or an opinion (political science). The measurement of a certain type of attributes is not the privilege of a particular research area, hence educational measurement may also measure cognitive abilities (e.g., verbal ability), psychology may also measure attitudes (e.g., towards significant others), health science may also measure personality traits (e.g., introversion), et cetera. Random variable X_k ($k = 1, \ldots, K$) represents the score on item k, and attains ordered scores $x_k = 0, \ldots, m_k$. For simplicity, we assume that within one measurement instrument all items are scored similarly, so that $m_k = m$. Items are often scored dichotomously, for example, incorrect/correct, no/yes, or disagree/agree, in which case $x = 0, 1$. The test score or the sum score summarizes the performance on the K items, $X_+ = \sum_{k=1}^{K} X_k$.

Nonparametric IRT models have in common that they use X_+ to order persons on a scale for the attribute. Each of the models does this by ordinally restricting the relation between the score on an item and the scale of measurement represented by one or more latent variables, but without the use of a parametric function such as the normal ogive or the logistic; see van der Linden (2016) for examples. Mokken (1971) considered the use of IRT models based on parametric functions for the relation between the item score and the latent variable, called item response functions (IRFs), for short, prohibitive of successful measurement of attributes for which foundational theory often was absent or poorly developed, and proposed his nonparametric IRT models (also, see Sijtsma & Molenaar, 2016). Nonparametric IRT models differ with respect to the assumptions they posit to describe the structure of the data, such that they imply an ordinal person scale. Stout (1990, 2002) developed assumptions that were as weak as possible, that is, imposing as few restrictions as possible on the data, and still enabling the ordering of persons. Ramsay (1991, 2016) used kernel smoothing and spline regression to arrive at an ordinal scale for person measurement. Holland and Rosenbaum (1986) derived a broad class of what one might call nonparametric IRT models and studied the mathematical properties of these models. Other work is due to, for example, Junker (1993), Douglas (2001), and Karabatsos and Sheu (2004), and recent work is due to, for example, Straat, van der Ark, and Sijtsma (2013), Tijmstra, Hessen, van der Heijden, and Sijtsma (2013), Ellis (2014) and Brusco, Köhn, and Steinley (2015). Each of the nonparametric approaches has their merits, but in this chapter, we focus on Mokken's approach and present the state of the art of this line of research.

2.2.1 Monotone Homogeneity Model

Mokken's model of monotone homogeneity (Mokken, 1971, pp. 115–169) for ordering persons using the sum score X_+, is based on three assumptions:

1. *Unidimensionality (UD)*. One latent variable denoted Θ stands for the attribute the K items measure.
2. *Monotonicity (M)*. The probability of obtaining a score of at least x on item k, $X_k \geq x$, increases or remains constant but cannot decrease as latent variable Θ increases: $P(X_k \geq x|\,\Theta)$ is non-decreasing in Θ, for $x = 1, \ldots, m$, while $P(X_k \geq 0|\,\Theta) = 1$ by definition; hence, it is uninformative about the relation between the item score and the latent variable. Conditional probability $P(X_k \geq x|\,\Theta)$ is called the item step response function (ISRF), and Fig. 2.1 shows an example of two items each with $x = 0, \ldots, 3$; hence, both items have three IRSFs for $x = 1, 2, 3$. For dichotomous items, $P(X_k = 1|\,\Theta)$ is non-decreasing in Θ, while $P(X_k = 0|\,\Theta) = 1 - P(X_k = 1|\,\Theta)$ and thus is uninformative when $P(X_k = 1|\,\Theta)$ is known.
3. *Local Independence (LI)*. When latent variable Θ explains the relations between the K items and no other latent variables or observed variables such as covariates explain the relations between at least two of the other items, conditioning on Θ renders the K-variate distribution of the item scores equal to the product of the K marginal item-score distributions. This property is called local independence (LI),

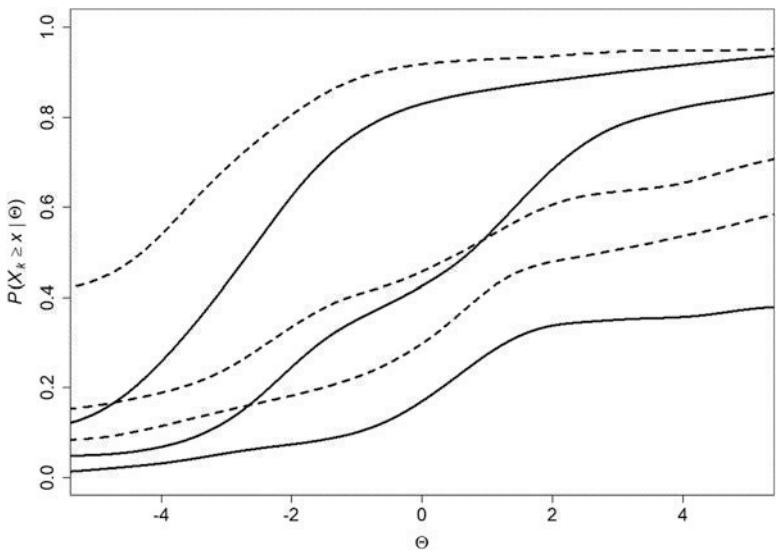

Fig. 2.1 Three nondecreasing item step response functions for two items; solid lines for one item, dashed lines for the other item

$$P(X_1 = x_1, \ldots, X_K = x_K|\Theta) = \prod_{k=1}^{K} P(X_k = x_k|\Theta)\,. \tag{2.1}$$

LI implies weak LI, meaning that the conditional covariance between any pair of items j and k equals 0 (Stout, 1990); that is, Eq. (2.1) implies $\sigma(X_j, X_k|\,\Theta) = 0$.

The three assumptions UD, M, and LI together do not enable direct estimation of Θ. However, for dichotomous items, the monotone homogeneity model implies that sum score X_+ orders persons stochastically on latent variable Θ; that is, for any value θ of Θ and any pair of values x_{+a} and x_{+b} of X_+ such that $0 \leq x_{+a} < x_{+b} \leq K$,

$$P(\Theta > \theta|X_+ = x_{+a}) \leq P(\Theta > \theta|X_+ = x_{+b}) \tag{2.2}$$

(Grayson, 1988). Hemker, Sijtsma, Molenaar and Junker (1997) called the property in Eq. (2.2) stochastic ordering of the latent trait by means of the sum score (SOL). SOL is important, because it shows that if one orders persons by their sum scores, they are also stochastically ordered by Θ. Hence, the monotone homogeneity model implies that Θ is an ordinal scale for persons, and that one can use X_+ to order persons on this scale. An interesting and insightful implication of Eq. (2.2) pertains to conditional expectations; that is,

$$\mathcal{E}(\Theta|X_+ = x_{+a}) \leq \mathcal{E}(\Theta|X_+ = x_{+b})\,, \tag{2.3}$$

meaning that sum score X_+ orders persons by expectation, that is, subgroups characterized by increasing mean Θs. Obviously, because random error affects measurement, one cannot unambiguously conclude at the level of individuals n_1 and n_2 that when one observes $x_{+n_1} < x_{+n_2}$, then $\theta_{n_1} \leq \theta_{n_2}$. Random measurement error may obscure the real ordering $\theta_{n_1} > \theta_{n_2}$, but for homogeneous sum-score groups, identified by θ_a and θ_b, Eq. (2.3) assures an ordering by mean Θs.

For polytomous items, Hemker et al. (1997) showed that the monotone homogeneity model does not imply SOL, hence the model does not produce an ordinal person scale; also see Hemker, van der Ark and Sijtsma (2001). They further showed that among the class of parametric IRT models for polytomous items, only the parametric partial credit model (Masters, 1982), and its special cases such as the rating scale model (Andrich, 1978), implies SOL. Other well-known polytomous IRT models, such as the generalized partial credit model (Muraki, 1992) and the graded response model (Samejima, 1969) do not possess the SOL property. This result suggested that sum score X_+ may not be useful for ordering people on Θ in most polytomous IRT models, but one may also argue that this is not a problem because such models allow the assumption of a real-valued variable Θ and its estimation, thus enabling person measurement using Θ and without X_+.

However, two additional results provide hope for X_+. First, based on multiple simulated latent variable distributions and ISRFs, van der Ark (2005) found that, as a rule, X_+ correctly orders people on Θ. When reversals with respect to Θ happen, they mostly concern X_+ values that are close, often just one unit apart. When K and

m decrease, and ISRFs are more similar, the proportions of person pairs showing ordering violations decrease. Reversely, short tests containing items with, say, five ordered scores, and ISRFs that vary greatly produced more ordering violations than long tests containing items with, say, three ordered scores and ISRFs that are similar.

One may argue that SOL must hold for models to justify ordinal person scales and that failure of SOL is unacceptable, thus rendering X_+ useless as a statistic that orders persons on Θ. Two arguments mitigate this position. One argument is that for realistic K, say, $K \leq 40$, measurement of psychological attributes suffers greatly from measurement error in any measurement value including X_+, which probably causes many accidental ordering reversals that cannot be distinguished from systematic reversals caused by failure of SOL (Eq. 2.2), and may even have a greater impact on ordering. Hence, irrespective of whether the IRT model implies SOL, random measurement error probably overshadows the damage a violation of SOL does to person ordering. The other argument concerns an ordering property van der Ark and Bergsma (2010) called weak SOL, which is an implication of Eq. (2.2), SOL, and which the authors proved holds for all polytomous IRT models assuming UD, M, and LI. Hence, weak SOL provides some relief when a model fails to imply the stronger SOL.

Weak SOL is defined as follows. Assume polytomous items, a fixed integer value x_{+c}, such that $1 \leq x_{+c} \leq Km$, and assume UD, M, and LI; then weak SOL means

$$P(\Theta > \theta | X_+ < x_{+c}) \leq P(\Theta > \theta | X_+ \geq x_{+c}) . \tag{2.4}$$

It may be noted that for $x_{+c} < 1$ and $x_{+c} > Km$, Eq. (2.4) is undefined. Weak SOL does not imply Eq. (2.2), SOL, and is thus a weaker ordering property; see van der Ark and Bergsma (2010) for a computational example showing that SOL can fail while weak SOL is satisfied. SOL Eq. (2.2) implies Eq. (2.3) concerning expected values, and weak SOL Eq. (2.4) implies a similar ordering property concerning expected values,

$$\mathcal{E}(\Theta | X_+ < x_{+c}) \leq \mathcal{E}(\Theta | X_+ \geq x_{+c}) . \tag{2.5}$$

Equation (2.5) shows that, for $x_{+c} = 1, \ldots, Km$, weak SOL enables the ordering of two groups defined by $X_+ < x_{+c}$ and $X_+ \geq x_{+c}$ on Θ. For example, if one selects the 20% best students from a sample using the test scores as a selection criterion, then weak SOL implies that the expected Θ value for the selected respondents is at least as high as the expected Θ value for the respondents who were not selected. However, weak SOL does not allow the ordering of more than two mutually exclusive groups (e.g., three groups defined by $X_+ < x_{+c}$, $x_{+c} \leq X_+ < x_{+c} + u$, and $X_+ \geq x_{+c} + u$; $u \in \{1, 2, \ldots, Km - x_{+c} - 1\}$); two non-exclusive groups (e.g., two groups defined by $X_+ < x_{+c} + u$ and $X_+ \geq x_{+c}$) or two non-exhaustive groups (e.g., two groups defined by $X_+ < x_{+c}$ and $X_+ \geq x_{+c} + u$; $u \geq 1$) (van der Ark & Bergsma, 2010, proposition; also see Douglas, Fienberg, Lee, Sampson, & Whitaker, 1991). One can check that for three persons, n_1, n_2, and n_3, with $x_{+n_1} < x_{+n_2}$, $x_{+n_2} < x_{+n_3}$, and consequently, $x_{+n_1} < x_{+n_3}$, for each person pair

one can always find cut scores x_{+c}, such that for each person pair weak SOL implies a pairwise ordering, but one can also check that an ordering of all three persons is not possible because three subgroups based on two cut scores always overlap.

We conclude that, based on theoretical considerations, the monotone homogeneity model for polytomous items only allows pairwise person ordering but not complete person ordering. Van der Ark's (2005) computational results give us enough confidence to use sum scores X_+ to order people on Θ in practical applications of tests and questionnaires.

2.2.2 *Latent Class Model*

The LCM assumes a discrete latent variable but refrains from specifying its dimensionality, thus defining unordered measurement values that represent latent classes. The model can be used to identify subgroups characterized by the same pattern of scores on K observables, and here we assume that N persons provide discrete integer scores on K items, just as with the discussion of the monotone homogeneity model. Like the monotone homogeneity model, the LCM assumes LI, but now given class membership. Let latent variable Φ have W discrete values denoted $w = 1, \ldots, W$; then LI is defined as in Eq. (2.1), but for $\Phi = w$.

Only assuming a discrete latent variable and LI would provide too little structure to restrict the probability structure governing the data (Suppes & Zanotti, 1981). Assuming every observation falls into one of just a few latent classes, W, restricts the LCM and makes it a feasible approach. One can write the probability of a particular pattern of item scores, denoted $\mathbf{X} = (X_1, \ldots, X_K)$ with realization $\mathbf{x} = (x_1, \ldots, x_K)$, and being in class $\Phi = w$, $P(\mathbf{X} = \mathbf{x} \wedge \Phi = w)$, as the product of the probability of being in class $\Phi = w$, $P(\Phi = w)$, and the probability of obtaining score pattern $\mathbf{X} = \mathbf{x}$ conditional on class membership, $P(\mathbf{X} = \mathbf{x} | \Phi = w)$; that is,

$$P(\mathbf{X} = \mathbf{x} \wedge \Phi = w) = P(\Phi = w)\, P(\mathbf{X} = \mathbf{x} | \Phi = w)\,. \tag{2.6}$$

Applying LI to conditional probability $P(\mathbf{X} = \mathbf{x} | \Phi = w)$ in Eq. (2.6), and summation across discrete classes yields the foundational equation of LCM analysis,

$$P(\mathbf{X} = \mathbf{x}) = \sum_{w=1}^{W} P(\Phi = w) \prod_{j=1}^{J} P\big(X_j = x_j | \Phi = w\big)\,. \tag{2.7}$$

The discrete IRFs (rather, for each item, W separate response probabilities), $P(X_k = x_k | \Phi = w)$, appear on the right-hand side in Eq. (2.7). The model is typically used in an exploratory fashion, because the classes are unknown, hence latent, and the quest is for the number W that explains the data structure best using the model in Eq. (2.6). Conducting an analysis involves estimating the class weights, $P(\Phi = w)$, for each w, and the item response probabilities, $P(X_k = x_k | \Phi = w)$, for each k and w.

One can use these probabilities in conjunction with Bayes theorem to assign people to the class with best fit; that is, for person n, one finds the class w for which

$$P(\Phi = w|\mathbf{X}_n = \mathbf{x}_n) = \frac{P(\mathbf{X}_n = \mathbf{x}_n|\Phi = w)\, P(\Phi = w)}{\sum_{w=1}^{W} P(\Phi = w) \prod_{k=1}^{K} P(X_k = x_k|\Phi = w)}, \tag{2.8}$$

is maximized and assigns person n to this class. This application of the model assigns individuals to latent classes, thus producing a nominal scale. Another application of LCM analysis is to identify latent classes in an effort to understand the structure of the data. Different applications use the LCM to impute scores (Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008), to model population ability distributions (Wetzel, Xu, & von Davier, 2015), to smooth large sparse contingency tables (Linzer, 2011), and to estimate the reliability of sum scores on tests (van der Ark, van der Palm, & Sijtsma, 2011) and of the scores on individual items (Zijlmans, van der Ark, Tijmstra, & Sijtsma, 2018).

LCMs have been extended with explanatory structures, such as regression models, multilevel models and factor models (Hagenaars & McCutcheon, 2002), but also IRT models such as the partial credit model (e.g., Bouwmeester, Vermunt, & Sijtsma, 2007), and led to numerous applications in a variety of research areas. In an effort to tie the LCM to the monotone homogeneity model, we briefly focus on Ligtvoet and Vermunt (2012; also, see Croon 1990, 1991; Hoijtink & Molenaar, 1997; van Onna, 2002; Vermunt, 2001) who used ordered LCM analysis to investigate assumption M of the monotone homogeneity model.

The unconstrained LCM (Eq. 2.7) is typically estimated using an EM algorithm, but can be estimated using a Gibbs sampler. Both methods yield estimates for the class weights $P(\Phi = w)$ and the item-response probabilities $P(X_k = x_k|\,\Phi = w)$. Ligtvoet and Vermunt (2012) explain how to use the LCM to test assumption M of the monotone homogeneity model by rephrasing that assumption as follows. Replace continuous latent variable Θ with discrete latent variable $\Phi = w$, $w = 1, \ldots, W$, and define the expectation

$$\mathcal{E}(X_k|\Phi = w) = \sum_{x=1}^{m} x \cdot P(X_k = x|\Phi = w)\,, \tag{2.9}$$

(Sijtsma & Hemker, 1998). We assume that $\mathcal{E}(X_k|w)$ is non-decreasing in Φ. The conditional expected item score, $\mathcal{E}(X_k|\Phi = w)$, summarizes the m item step response functions, $P(X_k \geq x|\,\Phi = w)$, for each item, while losing information present at the lower aggregation level, but simplifying the investigation of assumption M. Because for one item, conditional probabilities are dependent, in the Gibbs sampler, investigating assumption M by means of $\mathcal{E}(X_k|\Phi = w)$ entails sampling transformations of conditional probabilities, $P(X_k \geq x|\,\Phi = w)$, that are independent of one another, and together satisfy assumption M at the higher aggregation level of conditional expected item scores. Parameter estimates can be generated after convergence of the algorithm from the posterior distributions of the parameters.

A standard goodness of fit statistic is available for assessing the overall fit of the constrained LCM relative to competing models, and specialized fit statistics assess the fit of individual items. A model fitting strategy first entails choosing a value for W, the number of latent classes based on the best overall fit, and in the second analysis round determining for which items assumption M is satisfied. This is done by comparing the fit of the constrained W-class LCM to the unconstrained W-class LCM. The constrained model fits worse by definition but if the discrepancy between models is large, item fit statistics may be used to suggest which badly fitting items should be removed to improve the overall fit (rather than removing the item, the constraint M on the item is removed). Because inactivating constraint M for one or two items probably affects overall fit, the first analysis round is redone and depending on the fit, other items may be flagged for removal. After some iterations, the result is a W-class LCM for K^* items ($K^* \leq K$) for which assumption M holds, if applicable.

2.2.3 Cognitive Diagnostic Model

CDMs allow the assessment of mastery or non-mastery of multiple attributes or skills needed to solve items. CDMs have been applied most frequently to cognitive items in an educational context, but applications are also known to the evaluation and diagnosis of pathological gambling (Templin & Henson, 2006) and the understanding and scoring of situational judgment tests (Sorrel et al., 2016). Several models are available that have in common that they assume that the solution of an item depends on the availability of a set of latent attributes, and for different items different albeit partly overlapping subsets of latent attributes may be required. The most important difference between the two models we discuss here is that one is conjunctive or non-compensatory, and the other disjunctive or compensatory. Conjunctive models assume the tested person needs to master all attributes necessary to solve an item, and non-mastery of a required attribute cannot be compensated by mastery of another attribute. Disjunctive models require a subset of attributes to solve an item but not all attributes, and non-mastery of one or more attributes can be compensated by mastery of others. The models have in common that they compare a person's ideal item-score pattern with her observed item-score pattern, and posit an IRF that relates the two patterns and allows persons lacking attributes the item requires for its solution to solve it correctly (guessing), and likewise persons in possession of the necessary attributes to fail the item (slipping). Von Davier (2014) studied the relationship between non-compensatory and compensatory models, and showed mathematically how their differences may be understood in more detail.

Some notation needed for both models is the following. Let $\mathbf{X}_n = (X_{n1}, \ldots, X_{nK})$ be the vector of binary (incorrect/correct) scores for person n's responses on the K items, and let $\mathbf{A}_n = (A_{n1}, \ldots, A_{nD})$ be the binary latent attribute vector, where $A_{nd} = 1$ means that person n possesses attribute d, and $A_{nd} = 0$ that the person does not possess the attribute. Another building block of CDMs is the Q-matrix, which contains for each item (rows) and attribute (columns) elements q_{kd} indicating whether item k requires attribute d for its solution ($q_{kd} = 1$) or not ($q_{kd} = 0$). We discuss two ways in which $\mathbf{A}_n$ and matrix $\mathbf{Q}$ can be combined to produce a latent item-score vector, $\boldsymbol{\Xi}_n = (\Xi_{n1}, \ldots, \Xi_{nK})$, with realization $(\xi_{n1}, \ldots, \xi_{nK})$. $\boldsymbol{\Xi}_n$ may be considered ideal and can be compared to the observed item-score vector, $\mathbf{X}_n$, to determine how well the model fits the data. The two models we discuss are representatives of conjunctive and disjunctive approaches, and we discuss the models for didactical reasons, but notice that other, more flexible models are available. These alternative models are discussed elsewhere in this book.

The deterministic inputs, noisy "and" gate model (DINA; Junker & Sijtsma, 2001; Haertel, 1989; Macready & Dayton, 1977) is a conjunctive model that defines binary latent response variable, Ξ_{nk}, to indicate whether person n possesses all the attributes needed for solving item k ($\Xi_{nk} = 1$) or not ($\Xi_{nk} = 0$). The ideal responses are defined as

$$\Xi_{nk} = \prod_{d=1}^{D} A_{nd}^{q_{kd}}. \tag{2.10}$$

Equation (2.10) shows that if item k requires an attribute d (i.e., $q_{kd} = 1$) that person n lacks (i.e., $A_{nd} = 0$), then $A_{nd}^{q_{nd}} = 0^1 = 0$, yielding $\Xi_{nk} = 0$; otherwise, $A_{nd}^{q_{nd}} = 1$, and only if all power terms equal 1 we obtain $\Xi_{nk} = 1$. The IRFs relate the ideal latent item-score vector to the fallible real-data item-score vector by allowing masters ($\Xi_{nk} = 1$) to fail an item accidentally, called slipping, and quantified by the slipping parameter,

$$s_k = P(X_{nk} = 0 | \Xi_{nk} = 1), \tag{2.11}$$

and non-masters ($\Xi_{nk} = 0$) to succeed accidentally, quantified by the guessing parameter,

$$g_k = P(X_{nk} = 1 | \Xi_{nk} = 0). \tag{2.12}$$

Using the definitions in Eqs. (2.10), (2.11), and (2.12), the IRF of the DINA model is defined as

$$P(X_{nk} = 1 | \mathbf{A}_n, s_k, g_k) = (1 - s_k)^{\xi_{nk}} g_k{}^{1-\xi_{nk}}. \tag{2.13}$$

Equation (2.11) shows that for non-masters ($\xi_{nk} = 0$), we have $P(X_{nk} = 1 | \mathbf{A}_n, s_k, g_k) = g_k$ and for masters ($\xi_{nk} = 1$), we have $P(X_{nk} = 1 | \mathbf{A}_n, s_k, g_k) = 1 - s_k$.

Hence, the class of non-masters has a probability at the guessing level to solve the item correctly, and the class of masters has a probability reflecting non-slipping or, indeed, mastery. A feature of the IRF in Eq. (2.13) is that it is coordinate-wise monotone in $\mathbf{A}_n$ if and only if $1 - s_k > g_k$. One can check this monotonicity property by checking that changing zeroes in $\mathbf{A}_n$ in ones can change $\xi_{nk} = 0$ into $\xi_{nk} = 1$, but not vice versa; hence, by adding attributes, a non-master can become a master, but this makes sense only if scoring $X_{ij} = 1$ becomes more likely, i.e., if $1 - s_k > g_k$.

We briefly consider the disjunctive deterministic input, noisy "or" gate model (DINO; Templin & Henson, 2006) to illustrate a disjunctive process model. The DINO model assumes that the person needs to master only one attribute, A_{nd}, and the latent response variable is defined as

$$\Psi_{nk} = 1 - \prod_{d=1}^{D} (1 - A_{nd})^{q_{kd}}. \tag{2.14}$$

From Eq. (2.14) it can be seen that the combination of the item requiring an attribute that the person masters ($A_{nd} = q_{kd} = 1$), is the only combination that produces $(1 - A_{nd})^{q_{kd}} = 0$, hence a product equal to 0 and latent response, $\Psi_{nk} = 1$. Thus, one needs to master at least one attribute necessary for item k to produce a latent response $\Psi_{nk} = 1$. Several authors have suggested flexible frameworks that include the DINA and DINO models and several other CDMs (e.g., de la Torre, 2011; von Davier, 2008). This volume witnesses the wealth of CDMs and we therefore refrain from further discussion, except for two notes.

First, the joint distribution of the data conditional on the latent variables, here the D binary attributes, is the product of the conditional distributions of the item scores; that is, LI is assumed,

$$P(\mathbf{X}_n|\mathbf{a}_n) = \prod_{k=1}^{K} P(X_{nk}|\mathbf{a}_n)\,. \tag{2.15}$$

Also assuming that the data records of different persons are independent, the conditional likelihood of the data matrix $\mathbf{X}$ is written as

$$L(\mathbf{X}|\mathbf{a}) = \prod_{n=1}^{N} L(\mathbf{X}_n|\mathbf{a}_n)\,. \tag{2.16}$$

This joint likelihood can be maximized for the parameters $\mathbf{g} = (g_1, \ \ldots \ , g_K)$ and $\mathbf{s} = (s_1, \ \ldots \ , s_K)$, but because they are known to have unfavorable statistical properties, alternatively one rather uses the marginal likelihood approach,

$$L(\mathbf{X}) = \prod_{n=1}^{N} L(\mathbf{X}_n) = \prod_{n=1}^{N} \sum_{h=1}^{H} L(\mathbf{X}_n|\mathbf{a}_h)\, P(\mathbf{a}_h)\,, \tag{2.17}$$

where the number of possible skills patterns equals $H = 2^D$. The latent class structure is apparent on the right-hand side of Eq. (2.17). De la Torre (2009) discussed two estimation algorithms based on Eq. (2.17). One estimation method uses EM, which is labor-intensive due to the huge number of different latent attribute

vectors $\mathbf{A}_h$. The other estimation method avoids this problem by assuming that the elements of vector $\mathbf{A}$ are locally independent given a continuous higher-order latent variable Θ, having the structure of Eq. (2.1),

$$P(\mathbf{A}|\theta) = \prod_{d=1}^{D} P(a_d|\theta)\,, \tag{2.18}$$

and $P(a_d|\,\theta)$ is modeled as a two-parameter logistic model,

$$P(a_d|\theta) = \frac{\exp(\lambda_{0d} + \lambda_1\theta)}{1 + \exp(\lambda_{0d} + \lambda_1\theta)}, \tag{2.19}$$

in which λ_{0d} is the intercept, $\lambda_1 > 0$ is the slope, and $\theta \sim \mathcal{N}(0, 1)$ by assumption. Equation (2.19) renders the probability monotone in θ and dependent on an intercept parameter and a slope parameter that is equal across attributes. This is the higher-order DINA (de la Torre & Douglas, 2004), and an MCMC algorithm is used to estimate $D - 1$ intercept and 1 slope parameter. Yang and Embretson (2007) discussed an equation similar to Eq. (2.18) for inferring a person's most likely latent class $\mathbf{a}_h$ given her item-score pattern $\mathbf{X}_n$, the item parameters $\mathbf{g}$ and $\mathbf{s}$, and design matrix $\mathbf{Q}$.

Second, Junker and Sijtsma (2001) studied the properties of the DINA model from the perspective of the monotone homogeneity model and focused on monotonicity properties. Before we consider their result, we first notice a stochastic ordering result different from SOL, which reverses the roles of latent and manifest variables, and therefore is called stochastic ordering of the manifest variable by the latent variable (SOM). Starting from UD, M, and LI, and Eq. (2.4), for any pair of persons with $\theta_{n_1} < \theta_{n_2}$, Hemker et al. (1997) defined SOM as

$$P\left(X_+ \geq x_{+c}|\theta_{n_1}\right) \leq P\left(X_+ \geq x_{+c}|\theta_{n_2}\right). \tag{2.20}$$

The monotone homogeneity model thus supplies a latent structure justifying ordering people on the observable X_+ total score. Older approaches, such as classical test theory, did not supply such a justification, but simply recommended the use of X_+. With the exception of the Rasch model, modern IRT approaches based on UD, M, and LI missed that they also justify SOM and even the more useful SOL, which allows one making inferences about latent, explanatory structures—an ordinal latent scale—from observable data. Holland and Rosenbaum (1986) introduced the notion of non-decreasing summaries of the item scores, denoted $g(\mathbf{X}_n)$, non-decreasing coordinate-wise in X_k ($k = 1, \ldots, K$), and Junker and Sijtsma (2001) noticed that in the DINA model,

$$P[g(\mathbf{X}_n)\,|\mathbf{a}_n] \text{ is coordinate-wise non-decreasing in } \mathbf{a}_n \tag{2.21}$$

Obviously, this is a SOM property, meaning that the mastery of more attributes yields a higher summary score. The authors were unable to derive similar SOL properties for the DINA model.

2.3 Example

2.3.1 Data: Millon Clinical Multiaxial Inventory-III

We used the item scores of 1210 Caucasian patients and inmates in Belgium (61% males) on 44 items of the Dutch version of the Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, Millon, Davis, & Grossman, 2009; Dutch version by Rossi, Sloore, & Derksen, 2008). For more details about the sample, see Rossi, Elklit, and Simonsen (2010), and for a previous data analysis, see de la Torre, van der Ark, and Rossi (2018). The MCMI-III consists of 175 dichotomous items defining 14 personality scales, 10 clinical syndrome scales, and 5 correction scales. The 44 items we used pertain to the clinical syndrome scales anxiety (A), somatoform (H), thought disorder (SS) and major depression (CC). Several items are indicative for more than one clinical disorder (Table 2.1). For example, a positive response to Item 148 ("Few things in life give me pleasure") is believed to be an indicator for somatoform, thought disorder, and major depression. The 44 × 4 Q-matrix (Table 2.2) reflects the contributions of each item to each scale.

Table 2.1 The number of items per scale measuring one, two, or three disorders

Scale	Number of disorders			Total
	1	2	3	
A	9	5	0	14
H	2	9	1	12
SS	11	5	1	17
CC	6	11	1	18

Note: *A* Anxiety, *H* Somatoform, *SS* thought disorder, *CC* major depression

Table 2.2 Q-matrix of 44 items by four clinical disorders

	Disorder					Disorder					Disorder					Disorder			
k	A	H	SS	CC	*k*	A	H	SS	CC	*k*	A	H	SS	CC	*k*	A	H	SS	CC
1	0	1	0	1	61	1	0	1	0	108	1	0	0	0	147	1	0	0	0
4	0	1	0	1	68	0	0	1	0	109	1	0	0	0	148	0	1	1	1
11	0	1	0	0	72	0	0	1	0	111	0	1	0	1	149	1	0	0	1
22	0	0	1	0	74	0	1	0	1	117	0	0	1	0	150	0	0	0	1
34	0	0	1	1	75	1	1	0	0	124	1	0	0	0	151	0	0	1	1
37	0	1	0	0	76	1	0	1	0	128	0	0	0	1	154	0	0	0	1
40	1	0	0	0	78	0	0	1	0	130	0	1	0	1	162	0	0	1	0
44	0	0	0	1	83	0	0	1	0	134	0	0	1	0	164	1	0	0	0
55	0	1	0	1	102	0	0	1	0	135	1	0	0	0	168	0	0	1	0
56	0	0	1	0	104	0	0	0	1	142	0	0	1	1	170	1	0	0	0
58	1	0	0	0	107	0	1	0	1	145	1	1	0	0	171	0	0	0	1

Note: *k* Item number, *A* Anxiety, *H* Somatoform, *SS* thought disorder, *CC* major depression

We screened the data for outliers using the number of Guttman errors as an outlier score (Zijlstra, van der Ark, & Sijtsma, 2007), which identifies inconsistent item-score patterns given the item ordering based on item-total scores. Two respondents had an unexpectedly large number of Guttman errors, well beyond the cutoff value suggested by the adjusted boxplot (Hubert & Vandervieren, 2008). These respondents were removed from the data, yielding a final sample size of $N = 1208$. The data contained no missing values.

2.3.2 Analysis of the Data

Nonparametric IRT Analysis. We investigated the assumptions of the monotone homogeneity model by means of Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & van der Ark, 2017), using the R package `mokken` (van der Ark, 2007, 2012). First, we conducted a confirmatory Mokken scale analysis assuming all items belonged to the same scale. For all 946 item-pairs, scalability coefficient H_{jk} was significantly greater than 0 ($.10 < H_{jk} < .88$). Except for item 117 ($H_{117} = .28$), item 154 ($H_{154} = .25$), and item 168 ($H_{168} = .29$), for each of the other 41 items, item-scalability coefficient H_k was significantly greater than .30 ($.25 < H_k < .60$). These results support the fit of the monotone homogeneity model and suggest that the 44 items form a unidimensional scale. Total-scale scalability coefficient $H = .42$ ($se = .01$), a value which Mokken labeled as a medium scale. Hence, based on scalability coefficients alone, we did not find support that the four scales represent different clinical disorders. This first result might imply that a CDM with four attributes is superfluous.

Second, we conducted an exploratory Mokken scale analysis. For lower-bound values $c \in \{.00, .05, .10, .15, \ldots, .60\}$, we partitioned the 44 items into scales requiring that items admitted to a scale have $H_k > c$. This means that items may drop out of scales and remain unscalable. For $c \leq .20$, all 44 items constituted a single scale. For $.20 < c \leq .35$, some items were unscalable (i.e., item 154 was unscalable at $c = .25$, items 22, 117, 154, and 168 were unscalable at $c = .35$) but the remaining items constituted a single scale. For $c > .35$, the one-scale structure fell apart into multiple scales (3 scales at $c = .40$ to 11 scales at $c = .60$) and up to 7 unscalable items. At first glance, these results also support the hypothesis that the data are approximately unidimensional suggest. However, when one applies stricter criteria for scalability, the items represent a smaller number of attributes, and a closer look may be in order.

Third, we inspected local independence using the W indices (notation W not to be confused with the number of latent classes) proposed by Straat, van der Ark, and Sijtsma (2016). Space limitations do not permit a discussion of these indices; hence, we refer the interested reader to Straat et al. (2016). Index W_1, which is used for the detection of positive locally dependent item pairs, flagged 106 of the 946 item pairs; index W_3, which is used for the detection of negative locally dependent item pairs, flagged two of the 946 item pairs. Because we did not have benchmarks

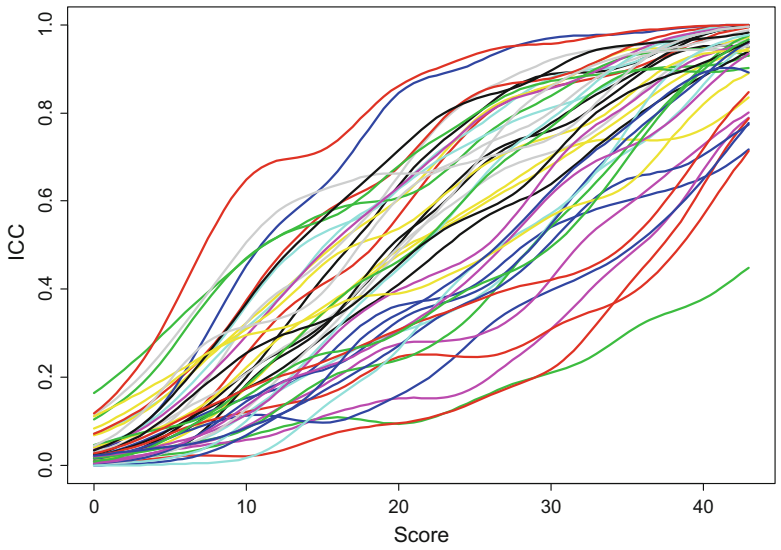

Fig. 2.2 Forty-four item response functions estimated by means of kernel smoothing ($h = 2.5$)

for W indices for so many items, we tentatively concluded that some items may be positive locally dependent, which suggests that within the unidimensional scale a more refined structure may be present.

Fourth, another in-depth analysis concerned the investigation of monotonicity using the property of manifest monotonicity (Junker & Sijtsma, 2000; Sijtsma & van der Ark, 2017). We did not find violations. This finding supports the fit of the monotone homogeneity model. Figure 2.2 shows the 44 IRFs estimated by means of kernel smoothing (Ramsay, 1991, 1996), using smoothing parameter $h = 2.5$.

To conclude, using the monotone homogeneity model, different clinical disorders remained unidentified and the data were unidimensional, but the unidimensionality signal was moderate. When we used stricter scaling criteria, the item set did not break down into four smaller scales that related to the four clinical disorders. To find out whether this result was a nonparametric-method effect, we also computed Yen's (1981) Q_1 statistic based on ten groups for testing goodness of fit of the parametric two-parameter logistic model (Table 2.3). The two-parameter logistic model is a special case of the monotone homogeneity model. Using Bonferroni correction (i.e., $p \approx .001$), based on the Q_1 statistic, none of the items showed misfit. Thus, the global goodness of fit measure for the two-parameter logistic model produced a result similar to that obtained from the confirmatory analysis in the context of the monotone homogeneity model, so that we could exclude a method effect.

Latent Class Analysis. We estimated twelve LCMs with $W = 1, 2, \ldots, 12$ classes using the R package poLCA (Linzer & Lewis, 2011), and computed information indices AIC3 (Andrews & Currim, 2003) and BIC (Schwarz, 1978). For large sample sizes and modest numbers of latent classes, both AIC3 and BIC are known

Table 2.3 Q_1 statistic for fit of two-parameter logistic model to each of 44 items

k	χ^2	p	k	χ^2	p	k	χ^2	p	k	χ^2	p
1	4.3	.828	61	14.3	.075	108	13.9	.085	147	16.3	.038
4	10.6	.226	68	6.8	.561	109	12.8	.120	148	8.9	.347
11	5.2	.735	72	9.7	.289	111	5.3	.727	149	16.5	.036
22	17.5	.025	74	15.5	.049	117	16.6	.035	150	16.9	.031
34	6.8	.561	75	10.6	.228	124	11.8	.162	151	6.6	.575
37	17.6	.025	76	10.0	.267	128	13.6	.093	154	2.2	.973
40	9.3	.320	78	23.1	.003	130	10.2	.252	162	18.2	.020
44	5.5	.703	83	10.3	.244	134	8.6	.378	164	10.0	.263
55	15.1	.057	102	11.8	.162	135	2.8	.948	168	9.9	.269
56	9.1	.337	104	12.5	.128	142	6.6	.580	170	19.9	.011
58	15.8	.046	107	13.8	.088	145	10.8	.213	171	9.5	.299

Note: k Item number, χ^2 chi-squared statistic with 10 degrees of freedom, p p value

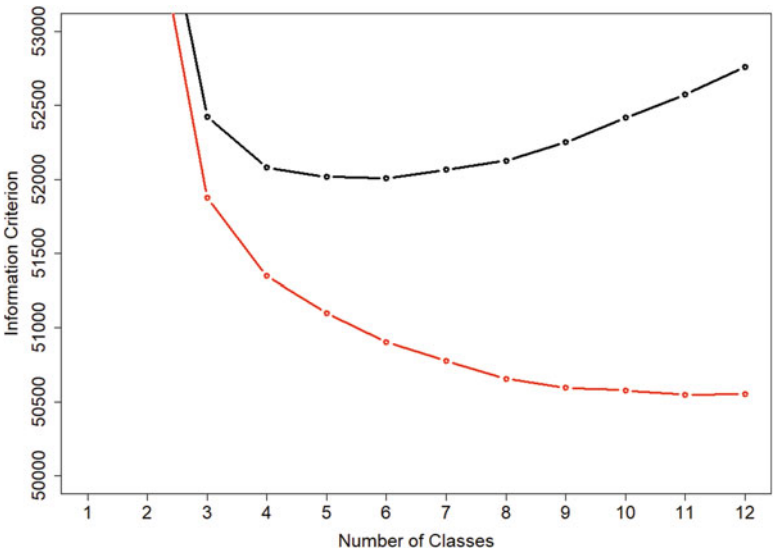

Fig. 2.3 BIC (black) and AIC3 (red) values for LCMs with 1, 2, . . . , 12 latent classes

to identify the correct number of classes reasonably well, but BIC tends to be conservative (Yang & Yang, 2007). To decrease the risk of local maxima, we estimated each model 10 times. We discuss results for the four-class LCM, because the number of classes is conveniently small, its interpretation relatively easy, while the fit of the model in terms of BIC seems adequate, and for the six and eleven-class LCMs, because these models provided the smallest values of AIC3 and BIC, respectively (Fig. 2.3).

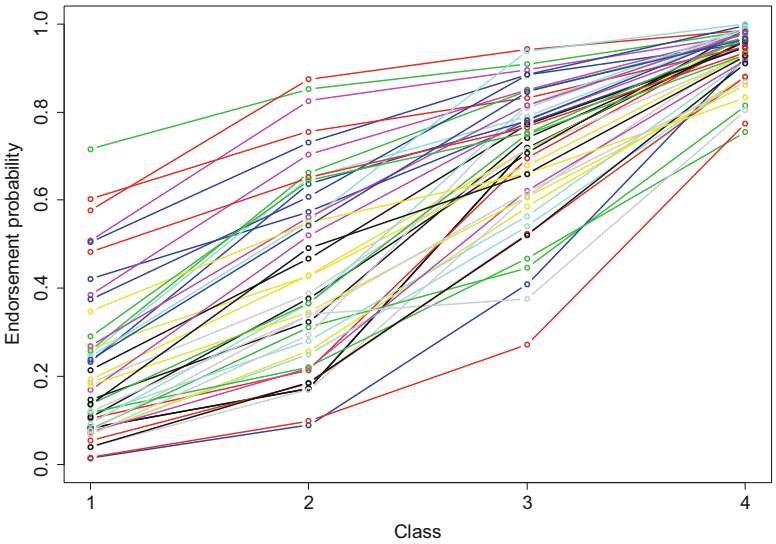

Fig. 2.4 *Endorsement probabilities for the four-class LCM*

Because it did not provide the smallest AIC3 and BIC values, the four-class LCM may have missed some of the heterogeneity in the data but the model corroborated the results from the nonparametric IRT data analysis. The four classes, with class probabilities $P(\Phi = w)$ equal to .288, .266, .159, and .291, are strictly ordinal, because for all 44 items, the estimated endorsement probability $P(X_i = 1 | \Phi = w)$ increased as w increased (Fig. 2.4).

Except for classes 3 and 4, the six-class LCM showed increasing endorsement probabilities (Fig. 2.5). The class probabilities $P(\Phi = w)$ equal .112, .264, .194, .111, .134, and .184. Figure 2.5 shows absence of consistent ordering between classes 3 and 4: For 17 of the 44 items (solid lines) the endorsement probability was larger in Class 4 than in Class 3. Fourteen of the 17 items relate to major depression or somatoform disorder. Hence, in addition to the ordinal trend, the six-class LCM seemed to distinguish a class with moderate endorsement probabilities leaning towards major depression and somatoform disorders (Class 3) and a class with moderate endorsement probabilities leaning towards anxiety disorders and thought disorders (Class 4). The eleven-class LCM was too difficult to interpret without an a priori hypothesized structure. Next, we investigated whether CDMs can provide additional information about the data structure.

Cognitive Diagnosis Models. Because this chapter discusses CDMs relative to nonparametric IRT and nonparametric LCMs, we compared these models with nonparametric CDMs. Our ambition was not to be complete with respect to the discussion of CDMs, but to discuss the general idea using a few simple models. The choice of two nonparametric CDMs, the basic DINA and DINO, reflect this modest ambition. Because these models are rather restrictive, we did not expect them to fit the data but used them instead for didactical purposes. We estimated the models

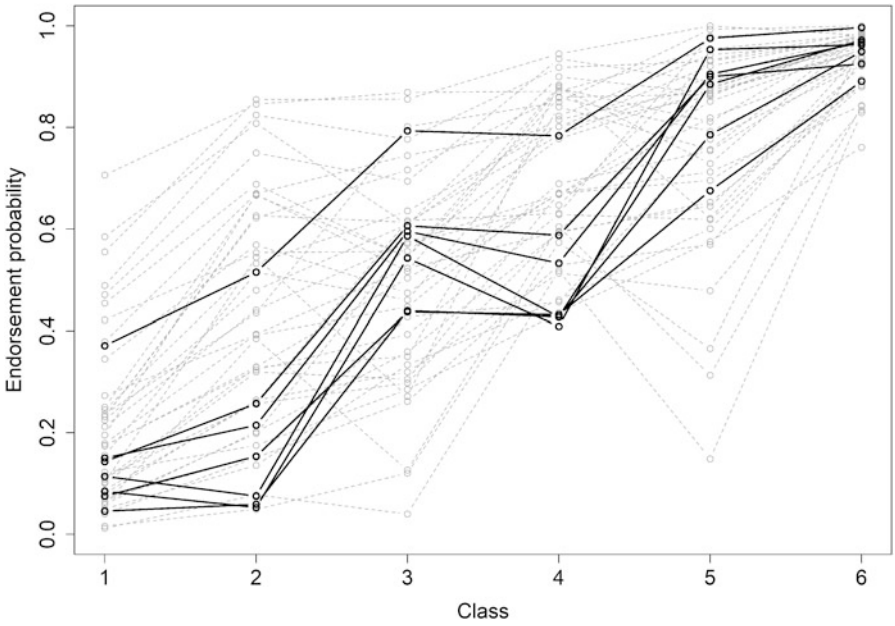

Fig. 2.5 *Endorsement probabilities for the six-class LCM. Black solid lines pertain to 17 items that have higher endorsement probabilities in Class 3 than in Class 4*

using the R package `NPCD` (Zheng & Chiu, 2016). First, the attribute profiles $\mathbf{A}$ were estimated using a nonparametric algorithm minimizing the plain Hamming distance (Chiu & Douglas, 2013), and given estimate $\widehat{\mathbf{A}}$, maximum likelihood estimates of the guessing parameters and slip parameters were obtained. Table 2.4 shows DINA results and Table 2.5 shows DINO results. Items printed in boldface had a high slipping or a high guessing parameter estimate, and the models fitted worse for these items. For global model fit, R package `NPCD` provides AIC and BIC but not AIC3. For the DINA model, AIC = 50,705 and BIC = 51,154, and for the DINO model, AIC = 50,296 and BIC = 50,745. One may notice that one cannot compare the AIC and BIC values of the these CDMs to the AIC and BIC values of the LCMs in Fig. 2.3. The reason is that for the CDMs, the likelihood is derived under the assumption that the Hamming distance-based class assignments ($\widehat{\mathbf{A}}$) are fixed, whereas for the LCMs, the class assignments are part of the likelihood (von Davier, personal communication). Comparing AIC and BIC of the CDMs and LCMs would be unfair and in favor of the DINO and the DINA given fixed $\widehat{\mathbf{A}}$.

For this example, $\Xi_{nk} = 1$ (DINA) means that respondent n suffers from all the disorders item k assesses, and $\Psi_{nk} = 1$ (DINO) means that respondent n suffers from at least one of the disorders item k assesses. Slipping parameter $s_k = P(X_{nk} = 0 | \Xi_{nk} = 1)$ (Eq. 2.11) is the probability that respondent n does not endorse item k, even though respondent n suffers from all the disorders related to item k; and guessing parameter $g_k = P(X_{nk} = 1 | \Xi_{nk} = 0)$ (Eq. 2.12) is the probability that respondent n endorses item k, even though respondent n does not

Table 2.4 Slipping and guessing parameters for the 44 items estimated from the DINA model

k	s_k	g_k	k	s_k	g_k	k	s_k	g_k	k	s_k	g_k
1	.08	.26	61	.03	.40	108	.34	.13	147	.05	.42
4	.10	.26	68	.17	.26	109	.16	.30	148	.15	.24
11	**.73**	**.03**	72	.15	.26	111	.28	.19	**149**	**.38**	**.13**
22	**.44**	**.11**	74	.22	.19	**117**	**.64**	**.08**	150	.38	.04
34	.11	.35	75	.20	.21	**124**	**.51**	**.05**	151	.33	.13
37	**.63**	**.03**	76	.18	.18	**128**	**.47**	**.06**	**154**	**.39**	**.22**
40	.19	.23	**78**	**.72**	**.02**	130	.16	.16	162	.19	.25
44	.07	.25	**83**	**.13**	**.38**	134	.33	.12	164	.31	.13
55	.15	.21	**102**	**.60**	**.07**	135	.24	.19	**168**	**.52**	**.12**
56	.12	.33	**104**	**.45**	**.11**	142	.12	.19	**170**	**.48**	**.08**
58	.19	.30	**107**	**.48**	**.09**	145	.08	.36	171	.35	.09

Note: k Item number, s_k slipping parameter, g_k guessing parameter. If $s_k + g_k > .5$, the values are printed in boldface

Table 2.5 Slipping and guessing parameters for the 44 items estimated from the DINO model

k	s_k	g_k	k	s_k	g_k	k	s_k	g_k	k	s_k	g_k
1	.09	.19	61	.07	.24	108	.34	.16	147	.05	.45
4	.13	.20	68	.14	.31	109	.15	.32	148	.28	.14
11	**.47**	**.06**	72	.14	.32	111	.32	.15	**149**	**.48**	**.07**
22	**.40**	**.14**	74	.26	.15	117	**.59**	**.09**	150	.29	.08
34	.16	.29	75	.32	.14	124	**.48**	**.06**	**151**	**.43**	**.10**
37	**.44**	**.09**	76	.29	.08	128	**.42**	**.10**	**154**	**.34**	**.25**
40	.18	.25	78	**.69**	**.03**	130	.21	.12	162	.14	.28
44	.04	.32	83	**.10**	**.42**	134	.27	.15	164	.30	.14
55	.19	.16	102	**.55**	**.08**	135	.25	.23	**168**	**.47**	**.14**
56	.11	.38	104	**.41**	**.15**	142	.19	.11	**170**	**.47**	**.09**
58	.17	.31	107	**.52**	**.07**	145	.14	.23	171	.28	.12

Note: k Item number, s_k slipping parameter, g_k guessing parameter. If $s_k + g_k > .5$, the values are printed in boldface

suffer from all the disorders related to item k. Because in the clinical context, one assumes that one endorses an item if one possesses at least one of the disorders, the DINO model seems more in line with this assumption than the DINA model. Based on the BIC, the DINO model fitted better than the DINA model, but for both models, proportions of slipping and guessing were high; see Tables 2.4 and 2.5.

For the DINO model, slipping parameter estimates were generally higher than guessing parameter estimates, and for 14 items, slipping parameter estimates exceeded .40 (Table 2.5). Hence, respondents suffering from a relevant disorder did not always endorse the item. An explanation could be that some items refer to rare circumstances. An example is item 78, "Even when I'm awake, I don't seem to notice people who are near me" ($s_{78} = .69$) that even respondents suffering from thought disorder may find too unlikely to endorse. Some other questions were

Table 2.6 Class sizes based on the Hamming distance-based attribute profiles for the DINO model in percentages

Class	Prevalence	Class	Prevalence
No disorder	45.3%	H and SS	1.6%
CC	1.9%	A and SS	4.7%
SS	2.4%	A and H	1.7%
H	2.1%	All but A	0.5%
A	8.7%	All but H	12.6%
CC and SS	1.3%	All but SS	1.1%
CC and H	0.2%	All but CC	3.4%
CC and A	6.5%	All disorders	6.0%

Note *A* anxiety, *H* somatoform, *SS* thought disorder, *CC* major depression

double barreled, which may explain low endorsement. An example is item 107, "I completely lost my appetite and have trouble sleeping most nights". Only for item 83, "My moods seem to change a great deal from one day to the next", the guessing parameter estimate exceeded .40 ($g_{83} = .42$). Item 83 relates to thought disorder, but given its high popularity, respondents not suffering from thought disorder also seemed to endorse the item.

Based on the estimated attribute profiles $\widehat{\mathbf{A}}$ of the DINO model, four attribute profiles had substantial size (Table 2.6): no disorder (45.3%), only A (8.7%), CC and A (6.5%), and SS, CC and A (12.6%). Approximately 73% of the respondents belonged to one of these four classes. If one adds the percentages in Table 2.6 that pertain to A, one finds that the DINO model identified anxiety (44,7%) as the most common disorder, followed by thought disorder (32,5%), major depression (30.2%), and somatoform (16,6%).

2.4 Discussion

This chapter discussed the relation between nonparametric IRT models and CDMs. The two approaches are related via the LCM, and both IRT models and CDMs may be viewed as restricted LCMs with a large number of classes. For IRT models, the number of classes equals the number of distinct θ values, but IRT models are mainly used for measuring individuals on a scale for the attribute of interest, and for this purpose the IRFs or ISRFs are nonparametrically or parametrically restricted. For CDMs, the number of classes equals 2^D attribute profiles, and a parametric functional form restricts the response probabilities within a class.

The nonparametric IRT models, LCMs, and CDMs are related, but researchers use the models in different situations. Nonparametric IRT models are useful for ordinal measurement and as a preliminary analysis for measurement using parametric IRT models, whereas LCMs are useful for nominal measurement; that is, identifying prototypes of respondents in the data, but also as a density estimation tool. CDMs are also used for nominal measurement. Yet by identifying the presence

or the absence of cognitive skills or clinical disorders, CDMs provide insight into the attribute of interest.

Comparing the fit of the nonparametric IRT models to an LCM or a CDM is not straightforward. Nonparametric IRT models have many methods to investigate the local fit, but a global goodness of fit statistic is unavailable. If one uses an ordinal LCM to investigate goodness of fit of a nonparametric IRT model, relative fit measures such as AIC or AIC3 are available. However, these measures suffer from the problem that they indicate which of the models the researcher compares fits best to the data, but not whether the best fitting of these models actually fits the data well. The data analysis using local fit methods showed that the nonparametric IRT model fitted well, and the nonparametric CMDs fitted well relative to the LCM.

The interpretation of the nonparametric IRT model and the CDMs was different. Using the IRT models, one uses a single continuous attribute to explain the responses to four comorbid disorders. The CDMs provide additional information. First, one could argue that the CDM analyses corroborated the conclusion from nonparametric IRT and LCA that the data were largely unidimensional, because classes showed a cumulative structure. That is, Class 0000 represents no disorders (45.3%), Class 1000 represents only an anxiety disorder (8.7%), Class 1010 represents anxiety and thought disorder (6,5%), Class 1110 represents all disorders but major depression (12.6%), and Class 1111 represents all disorders (6%). The classes can be considered ordered. Because over 73% of the sample belonged to these classes, a unidimensional scale may be appropriate for the majority of the sample (also, see von Davier & Habermann, 2014). However, to obtain a finer-grained picture, CDMs can retrieve information from the data that IRT models cannot.

References

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research, 40*, 235–243. https://doi.org/10.1509/jmkr.40.2.235.19225

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Bouwmeester, S., Vermunt, J. K., & Sijtsma, K. (2007). Development and individual differences in transitive reasoning: A fuzzy trace theory approach. *Developmental Review, 27*, 41–74. https://doi.org/10.1016/j.dr.2006.08.001

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

Chiu, C. Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*, 225–250. https://doi.org/10.1007/s00357-013-9132-9

Croon, M. (1990). Latent class analysis with ordered latent classed. *British Journal of Mathematical and Statistical Psychology, 43*, 171–192. https://doi.org/10.1111/j.2044-8317.1990.tb00934.x

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315–331. https://doi.org/10.1111/j.2044-8317.1991.tb00964.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J, van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development, 51*, 281–296. https://doi.org/10.1177/0748175615569110

Douglas, J. A. (2001). Asymptotic identifiability of nonparametric item response models. *Psychometrika, 66*, 531–540. https://doi.org/10.1007/BF02296194

Douglas, R., Fienberg, S. E., Lee, M.-L. T., Sampson, A. R., & Whitaker, L. R. (1991). Positive dependence concepts for ordinal contingency tables. In H. W. Block, A. R. Sampson, & T. H. Savits (Eds.), *Topics in statistical dependence* (pp. 189–202). Hayward, CA: Institute of Mathematical Statistics. Retrieved from http://www.jstor.org/stable/4355592

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

Formann, A. K., & Kohlmann, T. (2002). Three-parameter linear logistic latent class analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.),. *Applied latent class analysis* (pp. 183–210). Cambridge, UK: Cambridge University Press.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231. https://doi.org/10.1093/biomet/61.2.215

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Heinen, T. (1996). *Latent class and discrete latent trait models. Similarities and differences*. Thousand Oaks, CA: Sage.

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. https://doi.org/10.1007/BF02294555

Hemker, B. T., van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66*, 487–506. https://doi.org/10.1007/BF02296191

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171–189. https://doi.org/10.1007/BF02295273

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543.. Retrieved from https://projecteuclid.org/euclid.aos/1176350174

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

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359–1378. Retrieved from http://www.jstor.org/stable/2242199

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

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272. https://doi.org/10.1177/01466210122032064

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement, 28*, 110–125. https://doi.org/10.1177/0146621603260678

Leighton, J. A., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education. Theory and applications*. Cambridge, UK: Cambridge University Press.

Ligtvoet, R., & Vermunt, J. K. (2012). Latent class models for testing monotonicity and invariant item ordering for polytomous items. *British Journal of Mathematical and Statistical Psychology, 65*, 237–250. https://doi.org/10.1111/j.2044-8317.2011.02019.x

Linzer, D. A. (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis, 19*, 173–187. https://doi.org/10.1093/pan/mpr006

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1–29. https://doi.org/10.18637/jss.v042.i10

MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99–120. https://doi.org/10.3102/10769986002002099

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. https://doi.org/10.1007/BF02296272

Michell, J. (1999). *Measurement in psychology. A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCMI-III Manual* (4th ed.). Minneapolis, MN: Pearson Assessments.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. https://doi.org/10.1177/014662169201600206

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630. https://doi.org/10.1007/BF02294494

Ramsay, J. O. (2016). Functional approaches to modeling response data. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume one. Models* (pp. 337–350). Boca Raton, FL: Chapman & Hall/CRC.

Rossi, G., Elklit, A., & Simonsen, E. (2010). Empirical evidence for a four factor framework of personality disorder organization: Multigroup confirmatory factor analysis of the million clinical multiaxial inventory–III personality disorders scales across Belgian and Danish data samples. *Journal of Personality Disorders, 24*, 128–150. https://doi.org/10.1521/pedi.2010.24.1.128

Rossi, G., Sloore, H., & Derksen, J. (2008). The adaptation of the MCMI-III in two non-English-speaking countries: State of the art of the Dutch language version. In T. Millon & C. Bloom (Eds.), *The Millon inventories: A practitioner's guide to personalized clinical assessment* (2nd ed., pp. 369–386). New York, NY: The Guilford Press.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement. Theory, methods, and applications*. New York, NY: The Guilford Press.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. https://doi.org/10.2307/2958889

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

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

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

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*, 137–158. https://doi.org/10.1111/bmsp.12078

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*, 506–532. https://doi.org/10.1177/1094428116630065

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

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

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*, 75–99. https://doi.org/10.1007/s00357-013-9122-y

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12*, 117–123. https://doi.org/10.1027/1614-2241/a000115

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48*, 191–199. https://doi.org/10.1007/BF01063886

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Tijmstra, J., Hessen, D. J., van der Heijden, P. G. M., & Sijtsma, K. (2013). Testing manifest monotonicity using order-constrained statistical inference. *Psychometrika, 78*, 83–97. https://doi.org/10.1007/s11336-012-9297-x

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. https://doi.org/10.1007/s11336-000-0862-3

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

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

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

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. (Ed.). (2016). *Handbook of item response theory. Volume one. Models*. Boca Raton, FL: Chapman & Hall/CRC.

van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519–538. https://doi.org/10.1007/BF02295129

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

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25*, 283–294. https://doi.org/10.1177/01466210122032082

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology, 38*, 369–397.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71. https://doi.org/10.1111/bmsp.12003

von Davier, M., & Haberman, S. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models – A commentary. *Psychometrika, 79*, 340–346. https://doi.org/10.1007/s11336-013-9363-z

Wetzel, E., Xu, X., & Von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement, 75*, 739–763.

Yang, C.-C., & Yang, C.-C. (2007). Separating latent classes by information criteria. *Journal of Classification, 24*, 183–203. https://doi.org/10.1007/s00357-007-0010-1

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. A. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education. Theory and applications* (pp. 119–145). Cambridge, UK: Cambridge University Press.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262. https://doi.org/10.1177/014662168100500212

Zheng, Y., & Chiu, C.-Y. (2016). *NPCD: Nonparametric methods for cognitive diagnosis*. R package version 1.0–10 [computer software]. Retrieved from https://CRAN.R-project.org/package=NPCD

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research, 42*(3), 531–555.

Zijlmans, E. A. O., van der Ark, L. A., Tijmstra, J., & Sijtsma, K. (2018). Methods for estimating item-score reliability. *Applied Psychological Measurement, 42*, 553–570. https://doi.org/10.1177/0146621618758290.