Latent Class Models for Testing Monotonicity and Invariant Item Ordering for Polytomous Items

Rudy Ligtvoet[*]
University of Amsterdam

Jeroen, K. Vermunt
Tilburg University

[*] Correspondence should to Rudy Ligtvoet, Department of Pedagogical and Educational Sciences, Nieuwe Prinsengracht 130, 1018 VZ, Amsterdam. E-Mail: r.ligtvoet@uva.nl

Latent Class Models for Testing Monotonicity and Invariant Item Ordering for Polytomous Items

Abstract

Two assumptions that are relevant to many applications using item response theory are the assumptions of monotonicity (M) and invariant item ordering (IIO). A latent class model is proposed for ordinal items with inequality constraints on the class-specific item means. This model is used as a tool for testing for violations of M and IIO. A Gibbs sampling scheme is used for estimating the model parameters. It is shown that the deviance information criterion can be used as an overall test of M and IIO, while posterior predictive checks can be used to test these assumptions at the item level. A real-data application illustrates a model fitting strategy for detecting item that violate M and IIO.

1 Introduction

Item response theory (IRT) models are used to construct measures using multiple observed scores from tests or questionnaires. An example of a questionnaire item used to measure a subject's attitude towards women's liberation reads "*Women's liberation sets women against men.*", for which the subject scores 0 if he or she *agrees* with the statement, 1 if he or she is *neutral* towards the statement, and 2 if he or she *disagrees* with the statement (Heinen, 1996, p. 291). The score on each item i is considered to be a random variable X_i, for which each subject has the ordered score $x_i \in \{0, \ldots, m\}$, and where for the example $m = 2$. In IRT an unobservable *latent variable* θ is postulated to underly the items scores and the item scores are assumed to be mutually independent conditional on θ. The latter assumption is usually referred to as the assumption of *local independence* (Ip, 2001; Lord & Novick 1968, p. 361). Further, in IRT the relationship between the latent variable and the item scores is described by means of response functions, and it are these response functions about which specific assumptions are made. Let a_i be the slope parameter of item i and b_{x_i} its difficulty parameter corresponding to item score $X_i = x_i$. Samejima's (1969, 1997) graded response model, for example, specifies a logistic function for the cumulative probability of item i ($i = 1, \ldots, k$) and score $x_i = 1, \ldots, m$; that is,

$$\mathrm{logit} P(X_i \geq x|\theta) = a_i(\theta - b_{x_i}). \tag{1}$$

with $a_i > 0$. In this IRT model, the cumulative probabilities $P(X_i \geq x|\theta)$ are nondecreasing in θ, which implies that subjects with higher scores on θ (e.g., with a more extreme attitude related to women's liberation) are expected to score higher on each of the items. The nondecreasingness of the response functions is referred to as the assumption of *monotonicity* (M; e.g., Holland & Rosenbaum, 1986).

The logistic shape of the response function in Equation 1 is mathematical convenient, but it may also be too restrictive, in which case model-data misfit may occur while M may still hold. When solely interested in the assumption M, more flexible models may be considered which still allow for an ordering of the subjects on θ (e.g., Junker & Sijtsma, 2001; Molenaar, 1997). These subject ordering models are based on the assumption M, which states that for all i,

$$E(X_i|\theta), \text{ is nondecreasing in } \theta. \tag{M}$$

In contrast to most IRT models which define m response functions for each item, taking the conditional expectation $E(X_i|\theta)$ as the point of departure yields an IRT model with only one response function per item.

In addition to assumption M relating to the ordering of subjects based on θ, a second assumption relating to the ordering of the items is sometimes considered. The latter assumption, which is known as an *invariant item ordering* (IIO; Sijtsma & Hemker, 1998), implies that the same ordering of items, in terms of attractiveness (or difficulty), holds for all subjects. Let the items be indexed so that $i < i'$ indicates that $E(X_i) \leq E(X_{i'})$, then IIO states that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \cdots \leq E(X_k|\theta), \text{ for all } \theta. \tag{IIO}$$

Sijtsma and Junker (1996) provided various examples of practical applications of IRT requiring IIO. However, only few IRT models imply IIO (Sijtsma & Hemker, 1998), and those models that do so are often too restrictive to fit real life data well. For example, while the IRT model described in Equation 1 does not imply IIO, a version with equal slope parameters across items and difficulty parameters restricted by $b_{x_i} = d_i + c_x$ yields a model that is consistent with the IIO assumption. The resulting rating scale version of the grade response model (Sijtsma & Hemker, 1998) is defined as

$$\text{logit} P(X_i \geq x|\theta) = a(\theta - d_i - c_x). \tag{2}$$

This parametric IRT model will typically be too restrictive to show an acceptable model-data fit, whereas in reality the M and IIO assumptions may still hold. The reason for such a misfit is that the model imposes many more constraints than M and IIO alone. The contribution of this article is that it proposes a class of less restricted models that in addition to assumption M can be used to test assumption IIO without imposing parametric restrictions such as logistic response functions. The resulting non-parametric models can be considered additions to existing tools for IIO research (see also, Ligtvoet, Van der Ark, Te Marvelde, & Sijtsma, 2010).

Constrained latent class models (LCMs) have been shown to be useful tools for obtaining approximations of IRT models, where the underlying continuous latent variable is discretized by assuming that subjects belong to q homogenous latent classes (Heinen, 1996). By imposing linear constraints on the logistic parameters of a LCM, discretized versions of parametric IRT models, such as the graded response model, can be obtained (Vermunt, 2001). The approach proposed in this paper is to assess the assumptions M and IIO using latent class models with inequality restrictions. Such LCMs with inequality constraints on the model parameters have been used for testing model assumptions in the context of nonparametric IRT models (e.g., Croon, 1990, 1991; Karabatsos & Sheu, 2004; Vermunt, 2001). Also Hoijtink and Molenaar (1997; see also Hoijtink, 1998; Van Onna, 2002) showed how to formulate nonparametric IRT models as LCMs with ordinal constraints, and illustrated how to estimate and test such models using Bayesian methods. These Bayesian methods are not just confined to LCMs, but have, for example, also been applied in the context of ANOVA model selection (Klugkist, Laudy, & Hoijtink, 2005a). Here, we apply a similar type of Bayesian procedure to formulate a non-parametric model for the conditional expected item scores to test for violations of the assumptions of M and IIO. As an alternative, likelihood based methods developed by Bartolucci and Forcina (2005), and Vermunt (1999, 2001) could be adapted to the models of interest in this paper, but this approach is not further pursuit here.

The remainder of this paper is organized as follows. First, we present the unrestricted LCM and explain how to estimate its parameters using a Gibbs sampler. Then we describe the restricted LCMs corresponding to M and IIO and show how to incorporate the implied inequality restrictions in a Gibbs sampler. Subsequently, Bayesian tests for violations of M and IIO are described. And finally, the proposed LCM procedure for testing M and IIO is illustrated with an application using five questionnaire items measuring respondents' attitudes toward women's liberation.

2 Unrestricted Latent Class Models

To test the assumptions M and IIO, we approximate the continuous latent variable in IRT using a finite number of latent classes (i.e., $\theta = 1, \ldots, q$). The resulting latent class model is a model for $P(\mathbf{x})$; that is, a model for a particular response pattern (note that $\mathbf{x}$ refers to the vector of k item scores). LCMs are finite mixture models (Agresti, 1990; McLachlan & Peel, 2000), where subjects belonging the same latent class are homogeneous with respect to the probability $P(\mathbf{x}|\theta)$ of obtaining a certain response pattern. In addition, we assume that within each latent class, the item scores are mutually independent (e.g., Lazarsfeld & Henry, 1968, p. 22), which is equivalent to the IRT assumption of local independence (see Clogg, 1988). Let $\pi_{x_i|\theta}$ denote the probability of a score $X_i = x_i$ given θ and letting π_θ be the class proportion; the probability $P(\mathbf{x})$ is expressed as

$$P(\mathbf{x}) = \sum_\theta \pi_\theta P(\mathbf{x}|\theta) = \sum_\theta \pi_\theta \prod_i \pi_{x_i|\theta}. \tag{3}$$

Apart from choosing a fixed number of latent classes q and the local independence assumption, Equation 3 is yet unconstrained and referred to as the unconstrained LCM.

The Gibbs sampler is an iterative algorithm that can be used for obtaining samples from the posterior distribution of the parameters of a statistical model given a set of data, a likelihood function linking the data to the model of interest, and a prior distribution for the unknown parameters (e.g., Zeger & Karim, 1991). Let t denote the tth iteration of the Gibbs sampler algorithm. When applied to the model described in Equation 3, the algorithm starts by assigning initial values to the class proportions

$$\pi_\theta^{(0)} = \pi_1^{(0)}, \ldots, \pi_q^{(0)}$$

and the conditional item probabilities

$$\begin{aligned} \pi_{x_i|\theta}^{(0)} &= \pi_{0_1|1}^{(0)}, \ldots, \pi_{m_1|1}^{(0)}, \cdots, \pi_{0_k|1}^{(0)}, \ldots, \pi_{m_k|1}^{(0)}, \\ &\quad \cdots, \pi_{0_1|q}^{(0)}, \ldots, \pi_{m_1|q}^{(0)}, \cdots, \pi_{0_k|q}^{(0)}, \ldots, \pi_{m_k|q}^{(0)}. \end{aligned}$$

The subsequential three steps are passes iteratively. The first step is a data augmentation step, which involves assigning each subject to one of the latent classes (Tanner & Wong, 1987). The second and third steps consist of sampling values to the class proportions $\pi_\theta^{(t)}$ and the conditional item probabilities $\pi_{x_i|\theta}^{(t)}$, respectively. The algorithm is repeated until a criterium of convergence is reached. Ones this criterium is reached, the parameters are sampled in successive iterations as if sampled from their posterior distribution (Gelfand, Smith, & Lee, 1992). Based on these samples from the posterior, inferences can then be made about the parameter values. Our Gibbs sampler algorithm consists of the following three steps:

Step 1: Given the parameters values at the previous iteration, we derive from Equation 3 for each subject j with a score pattern $\mathbf{x}$ the probability of belonging to latent class θ as

$$P\big(\theta^{(t)}\big|\mathbf{x}_j\big) = \frac{\pi_\theta^{(t-1)}}{C} \prod_i \pi_{x_i|\theta}^{(t-1)},$$

with

$$C = \sum_{\theta=1}^{q} \pi_\theta^{(t-1)} \prod_i \pi_{x_i|\theta}^{(t-1)}.$$

Each subject is assigned to a latent class by a single draw from a multinomial distribution. This augmentation step yields values of the number of subjects in latent class θ, denoted $n_\theta^{(t)}$, and values of the number of subjects in latent class θ with the item score x_i on item i, denoted $n_{x_i|\theta}^{(t)}$. The $n_\theta^{(t)}$ and $n_{x_i|\theta}^{(t)}$ are used in the next two steps of the algorithm.

Step 2: The number of subjects the latent class θ ($n_\theta^{(t)}$) is defined by a multinomial distribution, which combined with a (conjugate) Dirichlet prior distribution yields a Dirichlet posterior distribution for ${\pi_\theta}^{(t)}$:

$${\pi_\theta}^{(t)} \sim \text{Dir}(n_1^{(t)} + \alpha, \ldots, n_q^{(t)} + \alpha).$$

Here, the α are hyper-parameters which are chosen to equal unity; reflecting ignorance concerning the information of the prior.

Step 3: Likewise, for a given item i and θ the parameters $\pi_{0_i|\theta}^{(t)}, \ldots, \pi_{m_i|\theta}^{(t)}$ are sampled form a Dirichlet distribution

$$\pi_{0_i|\theta}^{(t)}, \ldots, \pi_{m_i|\theta}^{(t)} \sim \text{Dir}(n_{0_i|\theta}^{(t)} + \alpha, \ldots, n_{m_i|\theta}^{(t)} + \alpha).$$

3 Latent Class Models with M and IIO Constraints

We wish to test whether there are violations of the assumptions M and IIO by making use of LCMs. Recall that these assumptions imply particular inequality constraints on $E(X_i|\theta)$ across θ values and across items, respectively. Because

$E(X_i|\theta) = \sum_x x_i \pi_{x_i|\theta}$ and $\pi_{x_i|\theta}$ are the LCM parameters, imposing the M and IIO constraints in a LCM implies that the inequality constraints on $E(X_i|\theta)$ should be translated into inequality constraints on the joint outcome space of $(\pi_{0_i|\theta}, \ldots, \pi_{m_i|\theta})$. In the Gibbs sampler, this corresponds to sampling from truncated Dirichlet distributions; more specifically, from distributions in which the $(\pi_{0_i|\theta}, \ldots, \pi_{m_i|\theta})$ parameters can only attain values which at the aggregated level of the expected item scores are in agreement with M and IIO. The sampling of $\pi_{x_i|\theta}$ under the relevant constraints is, however, complicated by the fact that the $m+1$ probabilities for item i and class θ are not independent of one another. To deal with this dependency, the $m+1$ probabilities can be sampled at once from a Dirichlet distribution and retained only if they agree with the constraints imposed by M and IIO, and rejected otherwise. A downside of this procedure is that it may become inefficient when the admissible outcome space is small. Van Onna (2002) proposed resolving this issue by a sequential sampling scheme in which the probabilities for categories 0 to $m-1$ are sampled from truncated Beta distributions and the probability for category m is obtained by $\pi^{(t)}_{m_i|\theta} = 1 - \sum_{x=0}^{m-1} \pi^{(t)}_{x_i|\theta}$. A small simulation (not reported here) revealed that this procedure does not yield correct samples from the posterior distribution. Instead we follow Hoijtink (1998, see also Laudy & Hoijtink, 2007) suggestion and re-parameterize $\pi_{x_i|\theta}$ as

$$\pi_{x_i|\theta} = \frac{\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}}, \tag{4}$$

where $\gamma_{x_i|\theta} \sim \text{Gamma}(n_{x_i|\theta} + \alpha, 1)$. It is important to note that also $E(X_i|\theta)$ can be expressed in terms of these $\gamma_{x_i|\theta}$ parameters; that is,

$$E(X_i|\theta) = \frac{\sum\limits_x x\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}}. \tag{5}$$

The advantage of this re-parametrization is that in contrast to the $\pi_{x_i|\theta}$'s, the $\gamma_{x_i|\theta}$'s can be sampled independently of one another (e.g., Ferguson, 1973; Narayanan, 1990). It is well-known that sampling $m+1$ $\gamma_{x_i|\theta}$ parameters from independent

Gamma distributions is equivalent to sampling the $m+1$ $\pi_{x_i|\theta}$ from a $m+1$ dimensional Dirichlet distribution. For the Gibbs sampler this means that the $\gamma_{0_i|\theta}, \ldots, \gamma_{m_i|\theta}$ parameters are sampled one by one from a truncated Gamma distribution with bounds depending on the other parameters. The latter implies that the constraints by M and IIO need to be reevaluated at each step.

3.1 Monotonicity Constraints

Implementation of the assumption M in a LCM means that in addition to the model formulated in Equation 3, for each item i, $E(X_i|\theta-1) \leq E(X_i|\theta)$ for $\theta = 2, \ldots, q$ and $E(X_i|\theta) \leq E(X_i|\theta+1)$ for $\theta = 1, \ldots, q-1$. Substitution of $E(X_i|\theta)$ by its definition in Equation 5 yields

$$E(X_i|\theta-1) \leq \frac{\sum\limits_x x\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}} \tag{6}$$

$$\frac{\sum\limits_x x\gamma_{x_i|\theta}}{\sum\limits_x \gamma_{x_i|\theta}} \leq E(X_i|\theta+1). \tag{7}$$

for $2 \leq \theta \leq q$ and $1 \leq \theta \leq q-1$, respectively.

In the Gibbs sampler, $\gamma_{x_i|\theta}$ is sampled at each iteration given the values of all the other parameters. This means that the restrictions on $\gamma_{x_i|\theta}$ implied by M should be derived from Equations 6 and 7; that is, the bounds for the admissible values of $\gamma_{x_i|\theta}$ are obtained by isolating the term $\gamma_{x_i|\theta}$ from these equations. More specifically, we derive the first bound for $\gamma_{x_i|\theta}$ using the equality $E(X_i|\theta-1) = E(X_i|\theta)$. Denoting this bound by u^{M}, we obtain

$$u^{\mathrm{M}}_{x_i|\theta} = \frac{\sum\limits_{y\neq x} \gamma_{y_i|\theta}\Big(E(X_i|\theta-1) - y\Big)}{x - E(X_i|\theta-1)},$$

for $\theta \geq 2$. The second bound corresponds to the value of $\gamma_{x_i|\theta}$ for which $E(X_i|\theta) =$

$E(X_i|\theta+1)$. Denoting this bound by v^{M}, for $\theta < q$

$$v^{\mathrm{M}}_{x_i|\theta} = \frac{\sum\limits_{y \neq x} \gamma_{y_i|\theta}\Big(E(X_i|\theta+1) - y\Big)}{x - E(X_i|\theta+1)}.$$

Though at first glance one may think that $u^{\mathrm{M}}_{x_i|\theta}$ and $v^{\mathrm{M}}_{x_i|\theta}$ are the lower and upper bounds for $\gamma_{x_i|\theta}$, respectively, this is not correct. To illustrate this point, consider $\gamma_{0_i|1}$, the parameter for the first category of item i at $\theta = 1$. Its bound is derived from the equality $E(X_i|\theta = 1) = E(X_i|\theta = 2)$. Here, v^{M} is not an upper but a lower bound for $\gamma_{0_i|1}$, because gamma values smaller than v^{M} yield higher $E(X_i|\theta = 1)$ values, which not allowed according to the M restriction. It turns out that the bounds define the domain of admissible values for $\gamma_{x_i|\theta}$, which can lie either inside or outside the interval defined by $u^{\mathrm{M}}_{x_i|\theta}$ and $v^{\mathrm{M}}_{x_i|\theta}$. This means that not only the bounds should be computed at each step of the Gibbs sampler, but it should also be checked whether the domain of admissible values lies inside or outside the bounds. Denoting the admissible domain under assumption M by $\mathcal{A}^{\mathrm{M}}_{x_i|\theta}$, Step 3 of the Gibbs sampler can be implemented as follows:

Step 3*: For each $\gamma_{x_i|\theta}$ at each iteration t compute the bounds $u_{x_i|\theta}$ and $v_{x_i|\theta}$ and sample a new value $\gamma^{(t)}_{x_i|\theta}$ from a truncated Gamma distribution:

$$\gamma^{(t)}_{x_i|\theta} \sim \mathrm{Gamma}(n^{(t)}_{x_i|\theta} + \alpha, 1 \big| \gamma^{(t)}_{x_i|\theta} \in \mathcal{A}_{x_i|\theta}).$$

After each sample of $\gamma^{(t)}_{x_i|\theta}$, $E(x|\theta)^{(t)}$ is recomputed using Equation 5.

The method of inverse probability sampling is used to obtain samples from the relevant truncated gamma distributions (e.g., Gelfand, Smith, & Lee, 1992).

3.2 Invariant Item Ordering Constraints

The IIO constraints on the LCM are similar to the ones for M, but with the role of the items and latent classes reversed; that is, for any class θ $E(X_{i-1}|\theta) \leq E(X_i|\theta)$

for $i = 2, \ldots, k$ and $E(X_i|\theta) \leq E(X_{i+1}|\theta)$ for $i = 1, \ldots, k-1$. The bounds on $\gamma_{x_i|\theta}$ under IIO are

$$u^{\mathrm{IIO}}_{x_i|\theta} = \frac{\sum_{y \neq x} \gamma_{y_i|\theta}\Big(E(X_{i-1}|\theta) - y\Big)}{x - E(X_{i-1}|\theta)},$$

for $i \geq 2$, and

$$v^{\mathrm{IIO}}_{x_i|\theta} = \frac{\sum_{y \neq x} \gamma_{y_i|\theta}\Big(E(X_{i+1}|\theta) - y\Big)}{x - E(X_{i+1}|\theta)},$$

for $i < k$. At each iteration of the Gibbs sampler, both bounds are computed and the admissible domain $\mathcal{A}^{\mathrm{IIO}}_{x_i|\theta}$ is determined. To constrain the LCM by both M and IIO, the $\gamma^{(t)}_{x_i|\theta}$ parameters are sampled from truncated Gamma distributions with admissible ranges $\mathcal{A}_{x_i|\theta}$ defined as the intersection of $\mathcal{A}^{\mathrm{M}}_{x_i|\theta}$ and $\mathcal{A}^{\mathrm{IIO}}_{x_i|\theta}$.

3.3 Assessing Convergence

The values of the parameters obtained from the Gibbs sampler can be considered samples from their posterior distribution as the number of iterations t of the Gibbs sampler approaches infinity. In practice however, we are interested in t to be sufficiently large for our samples to correctly approximate the posterior distributions. The first samples of the model parameters are drawn after discarding the initial 10000 parameter values corresponding to the burn-in period. Sequential samples are drawn at intervals of 10 iterations. For these samples, convergence is first assessed by comparing the samples from the posteriors between successive samples of size 5000 (e.g., Hoijtink & Molenaar, 1997). If the differences between the posteriors are small, it is concluded that convergence is reached. If the differences are large, the samples are discarded and new samples are taken until it can be concluded that convergence is reached. A second assessment for convergence we used is by inspection of the trace lines for the likelihood function across iterations of the Gibbs sampler (e.g., Gamerman, 1997, pp. 134-137). A nearly flat trace line across the samples indicates convergence.

3.4 Parameter Estimation

With the 10000 samples of parameter values taken from their posterior, the parameter values are estimated by the median value. Likewise, the 95% credibility interval is taken between the 2.5th percentile and the 97.5th percentile. Values of $E(X_i|\theta)$ were computed for each of these samples, the corresponding posterior expectations and credibility intervals are reported. In the application described below, the 10000 samples were obtained after 10000 burn-in iterations by running the Gibbs sampler another 100000 iterations and retaining each 10th draw.

3.5 Assessing Model Adequacy

Three statistics are considered to assess the fit of the M and IIO constrained LCMs. The first one is the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002), which can be used to compare the overall goodness-of-fit of competing models. Here, an increase in DIC after adding the M or IIO constraints to the model indicates an overall violation of the model assumptions. In addition, two statistics were developed to test for violations of the assumptions M and IIO, respectively, at the item level.

3.5.1 Bayesian Deviance Statistic

Under inequality constraints, the number of free parameters of a model (or model complexity) is not easily defined. However, Spiegelhalter, Best, Carlin, and Van der Linde (2002) proposed a Bayesian measure p_D for the effective number of parameters of a model that can easily be obtained with our Gibbs sampler. Let $P(\mathbf{X}|\pi)$ denote the likelihood function, and let $D(\pi) = -2 \ln P(\mathbf{X}|\pi)$ denote the deviance. From the sampled parameter values from the Gibbs sampler, the following two quantities can be computed: $P_E\big(D(\pi)\big)$ which is the posterior expectation of the deviance, and $D\big(P_E(\pi)\big)$ which is the deviance given the posterior expectation of the parameters.

The effective number of parameters is expressed as

$$p_D = P_E\big(D(\pi)\big) - D\big(P_E(\pi)\big).$$

Spiegelhalter, Best, Carlin, and Van der Linde (2002) suggested using DIC as a measure for comparing model-data fit across competing models. Similar to Akaike's (1973) information criterion, DIC equals the deviance penalized by the complexity of the estimated model. It defined as

$$\mathrm{DIC} = P_E\big(D(\pi)\big) + p_D.$$

Lower DIC values indicate a better model-data fit.

3.5.2 Posterior Check for Monotonicity

Now we define the statistic expressing the deviation from M for item i. For two items i and i', let the corresponding *rest score* be defined as

$$Y_{ii'} = \sum_{i'' \neq i, i'} X_{i''},$$

which is the sum score excluding items i and i', and let n_y denote the number of observations with rest score $Y_{ii'} = y$. Let $S(X_i, X_{i'}|y)$ be the covariance between X_i and $X_{i'}$ conditional on their rest score $Y_{ii'} = y$. Holland and Rosenbaum (1986; Rosenbaum, 1984) showed that M implies that these covariances are nonnegative. We define the item specific statistic for M as

$$M_i(\mathbf{X}) = \frac{1}{n} \sum_{i' \neq i} \sum_{y} n_y S(X_i, X_{i'}|y), \tag{8}$$

where the weighing factor n_y accounts for small observations yielding less reliable estimates of the covariances.

To obtain a distribution of the statistic in Equation 8 under the null-hypothesis that the constrained LCM holds, new data sets $\mathbf{X}^{(t)}$ are generated using the parameter values of the tth sample. For each of these data sets, statistic $M_i(\mathbf{X}^{(t)})$

is computed, which yields a sampling distribution for the test statistic under the null-hypothesis of the constrained LCM. We define the (one-sided) p-value for assessing for each item whether it fits the constrained LCM as the proportion of times $M_i(\mathbf{X}^{(t)})$ is smaller than the observed $M_i(\mathbf{X})$ (cf. Gelman, Meng, & Stern, 1996; Meng, 1994). A small p-value undermines the credibility of assumption M for that particular item (i.e., M is likely to be violated for that particular item).

3.5.3 Posterior Check for Invariant Item Ordering

The last statistic expresses the deviation from IIO for item i. It is derived in a similar manner as $M_i(\mathbf{X})$. For two items i and i', where $i < i'$, IIO implies that $E(X_{i'} - X_i|y) \geq 0$, for all $Y_{ii'} = y$ (Ligtvoet, Van der Ark, Bergsma, & Sijtsma, in press). As a statistic for goodness of fit of item i, we propose

$$IIO_i(\mathbf{X}) = \frac{1}{n}\sum_{i' \neq i}\sum_{y} n_y E(X_{i'} - X_i|y). \tag{9}$$

In Equation 9, we only consider for i' the items adjacent of i. Similar to $M_i(\mathbf{X})$, a posterior sample distribution of $IIO_i(\mathbf{X})$ with the corresponding p-value can be computed by replicating new data sets $\mathbf{X}^{(t)}$. A small p-value undermines the credibility of the IIO assumption with the particular item involved in the most severe violations.

3.5.4 Model Fitting Strategy

We propose using a three-step model fitting strategy for assessing whether assumptions M and IIO hold. The steps involve: 1) determining the number of latent classes $\hat{q}$, 2) determining for which items M holds, and 3) determining for which items IIO holds. This procedure yields a LCM for which M and IIO holds for a certain number of items. We refer to the number of items for which M holds as $\hat{k}^M$ and for which IIO holds as $\hat{k}^{IIO}$.

Step 1 involves estimating LCMs with different numbers of classes, where the model with the lowest DIC is retained for the next step. Step 2 starts by fitting a M restricted $\hat{q}$-class model and comparing its DIC value to the unrestricted $\hat{q}$-class model. In case the constrained LCM fits worse, the posterior check for M is used to determine for which items assumption M is violated. The M constraint is then removed for the item with the largest misfit, and the model is reestimated with M constraints on the remaining items. This is repeated until the DIC indicates that the LCM constrained by M for the remaining $\hat{k}^M$ items fits at least as well as the unconstrained LCM. Step 3 starts with the estimation of a LCM constrained by both M and IIO, with constraints imposed only on those $\hat{k}^M$ items for which M held. The fit of this LCM is compared to the fit of the final model from step 2. As long as the DIC values indicate that the IIO restricted model fits worse, the IIO constraint is relaxed for the item for which the posterior check indicates the largest misfit due to a violation of IIO. The three-step strategy gives information on the items involved in misfit due to violations of M and IIO. On the basis of this information, the researcher can choose to discard items from the test.

4 Application

To illustrate the procedure for testing M and IIO by means of constrained LCMs, we use an application to a set of items from a study on sociocultural developments in The Netherlands (Felling, Peters, & Schreuder, 1987; Heinen, 1996, chap. 2 and 3). These were self-ratings of 1134 subjects to five statements related to women's liberation, each with three ordered score categories.

Because DIC values showed that an unrestricted four-class model did not improve the model-data fit compared to a three-class model, we retained the three-class model for testing M and IIO. The unconstrained LCM with three latent classes had a DIC equal to 8539.62. The estimates of the conditional expectations revealed no violations of the assumption M for any item, whereas the item ordering of items 1 and 2 was different at the first latent class than at the remaining two classes, and

Table 1: Median Values Constrained LCM (IIO Relaxed for Item 2).

Latent	Class	Conditional Expectation				
Class	Proportion	Item 1	Item 2	Item 3	Item 4	Item 5
1	.16	.23	.17	.55	.69	1.49
2	.34	.72	.73	1.30	1.65	1.90
3	.51	1.61	1.79	1.69	1.93	1.97

the item ordering of items 2 and 3 was different at the third latent class.

Then a LCM was fitted constrained by M for all five items. This constrained LCM fitted as well as the unconstrained LCM, with DIC = 8539.56. The posterior check for M also indicated that all items fitted the model (i.e., no violation of M was detected). So there is no reason to relax the M assumption for any of the items.

With the LCM constrained by both M and IIO for all five items we obtained DIC = 8544.52, which indicated that the model fitted the data less well than the LCM constrained by only M. The posterior check for IIO indicated that the lack of model-data fit might be due to item 2, for which $IIO_2 = 0.284$ ($p = .001$). All other items had p-values larger than .10. The posterior checks for M remained similar to those of the LCM constrained by only M. Then we released the IIO constraint for item 2, which means that item 1 is now constrained by IIO from above by item 3 and item 3 from below by item 1. This model showed a good model-data fit based on DIC = 8538.93 (lowest value of DIC for all models tested). None of the posterior checks for M and IIO showed any remaining violations of M or IIO. Table 1 contains the median values of the samples of the class proportions and conditional expectations under the LCM constrained by M for all five items and IIO for items 1, 3, 4, and 5.

5 Discussion

It was shown how the M and IIO assumptions in IRT can be translated into LCMs with inequality constraints on the class-specific item means. A Gibbs sampling procedure was presented which uses a re-parametrization of the model probabilities to make the implementation of the complex inequality constraints feasible. Our data application illustrated the proposed three-step model fitting strategy for detecting items that violate M and IIO. This procedure makes use of the overall fit statistic DIC, as well as item-specific posterior checks.

In case the DIC indicates that the specified model does not hold, posterior checks can be used for assessing which items violate M or which items were involved in violations of IIO. In the data application with five items and 1134 subjects, the posterior checks seem to work well. However, the sample size required for these checks to achieve a reasonable power for detecting violations in larger tests remains a question for future research. Alternative statistics for these checks may also be considered.

Other methods for assessing model-data fit may also be considered. For example, Kluglist, Laudy, & Hoijtink (2005a) considered the Bayesian factor for model selection. A drawback of the Bayesian factor is that it depends on which other models are included in the comparisons (Stern, 2005; see also Kluglist, Laudy, & Hoijtink, 2005b). This means that to obtain the best fitting model, many models with different numbers of latent classes and different sets of items for which M and IIO holds, would need to be considered. In our three-step model fitting strategy we first assess the number of latent classes; second, we assess for which items assumption M is violated; and third, we assess for which items assumption IIO is violated. This sequential strategy is efficient when it comes to identifying a set of items for which both M and IIO holds. However, other strategies may be proposed, which may identify another model as best fitting. For example, one could start with testing both M and IIO for many latent classes and step-by-step relaxing some constraints or reducing the number of latent classes. Which kind of strategy is optimal with respect to finding the best fitting model is a topic for future research.

Also, in our analysis the number of latent classes was fixed to three based on the goodness-of-fit of the unrestricted LCM. An alternative strategy for testing M and IIO using LCMs might be to use a model with a large number of latent classes as the starting point. However, a downside of this alternative strategy is that some of the classes will contain only a few subjects, hence reducing the reliability of the estimates. Another possibility may be to treat the number of latent classes as an additional parameter in the Bayesian estimation procedure. This would yield a Gibbs sampler with in addition to the augmentation and the two sampling steps, a step in which the number of latent classes is updated (e.g., Neal, 1991; Richardson & Green, 1997).

Because most IRT models are used with the goal of obtaining an ordering of subjects, it is most logical to assume IIO only in addition to M, which is also what we did in this chapter. However, in theory it is also possible to test IIO without assuming M. This could be done using a LCM with only the IIO constraints. A practical problem one may encounter when estimating such a model with a Gibbs sampler is what is referred to as the label switching problem (e.g., Redner & Walker, 1984; Stephens, 2000). Because the classes are not ordered, their labeling is arbitrary and may thus change during the Gibbs sampling iterations (see Hoijtink, 1998, for an example of this phenomena). Note that this may also occur in an unconstrained LCM. Stephens (1997, 2000) developed an algorithm that can be used to reorder the classes so that their order is same across Gibbs sample iterations (see also Jasra, Holmes, & Stephens, 2005).

A last extension of the proposed procedures for testing M and IIO we would like to mention is the possibility of imposing constraints that correspond to more common IRT models implying M and IIO. For example, Van Onna (2002) defined a LCM with order restrictions similar to ours but imposed on the cumulative probabilities

$P(X_i \geq x_i|\theta)$ instead of the expected values $E(X_i \geq x_i|\theta)$. These restrictions imply both M and IIO to hold (cf. the strong double monotonicity model, Sijtsma & Hemker, 1998). Ligtvoet, Van der Ark, Bergsma and Sijtsma (in press) suggested models with similar order restrictions for continuation ratios and adjacent category odds (see also Vermunt, 2001).

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki (Eds.), *Procedings 2nd International Symposium on Information Theory*, pp. 267-281. Budapest: Akadémiai Kiadó.

Bartolucci, F. & Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika, 70*, 31-43.

Clogg, C. C. (1988). Latent class models for measuring. In R. Langeheine & J. Rost (Eds.), *Latent traits and latent class models* (pp. 173-205). New York: Plenum Press.

Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171-192.

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.

Felling, A., Peters, J., & Schreuder, O. (1987). *Religion in Dutch society 85; documentation of a national survey on religious and secular attitudes in 1985.* Amsterdam: Steinmetz Archive.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics, 1*, 209-230.

Gamerman, D. (1997). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* London: Chapman & Hall.

Gelfand, A. E., Smith, A. F. M., & Lee, T. -M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association, 87*, 523-532.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.

Heinen, T. (1996). *Discrete latent variable models.* Thousand Oaks, CA: Sage.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and postrior predictive p-values: Applications to educational testing. *Statistica Sinica, 8*, 691-711.

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.

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

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66*, 109-132.

Jasra, A., Holmes, C. C., Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science, 20*, 50-67.

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

Karabatsos, G., & Sheu, C. -F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement, 28*, 110-125.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005a). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10*, 477493.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005b). Bayesian eggs and bayesian omelettes: Reply to Stern (2005). *Psychological Methods, 10*, 500503.

Laudy, O., & Hoijtink, H. (2007). Bayesian methods for the analysis of inequality constrained contingency tables. *Statistical Methods in Medical Research, 16*, 123-138.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (in press). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika.*

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.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

McLachlan, G. J. & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*, 1142-1160.

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.

Narayanan A. (1990). Computer generation of Dirichlet random vectors. *Journal of Statistical Computation and Simulation, 36*, 19-30.

Neal, R. (1991). Bayesian mixture modeling. In C. R. Smith, G. J. Erickson, & P. O. Neudorfer (Eds.), *Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods* (pp. 197-211). Dordrecht, The Netherlands: Kluwer.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26*, 195-239.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B, 59*, 731-792.

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

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

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63*, 183-200.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79-105.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, 64*, 583-639.

Stephens, M. (1997). Bayesian methods for mixtures of normal distributions. *PhD Thesis.* University of Oxford.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B, 62*, 795-809.

Stern, H. S. (2005). Model inference or model selection: Discussion of Klugkist, Laudy, and Hoijtink (2005). *Psychological Methods, 10*, 494499.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528-540.

Van Onna, H. J. M. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519-538.

Vermunt, J. K. (1999). A general non-parametric approach to the analysis of ordinal categorical data. *Sociological Methodology, 29*, 197-221.

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.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79-86.