Statistica Sinica **8**(1998), 691-711

CONSTRAINED LATENT CLASS ANALYSIS USING THE GIBBS SAMPLER AND POSTERIOR PREDICTIVE P-VALUES: APPLICATIONS TO EDUCATIONAL TESTING

Herbert Hoijtink

Utrecht University

Abstract: This paper will illustrate how a number of educational testing models may be formalized as constrained (using inequality and equality constraints) latent class models. The parameters of these models will be estimated using an application of the Gibbs sampler. The goodness of fit of these models will be determined using (pseudo) likelihood ratio tests evaluated via posterior predictive P-values. The feasibility of both the estimation and testing procedures will be illustrated via the analysis of a number of simulated data sets.

Key words and phrases: Discrepancy measure, Gibbs sampler, inequality constraints, latent class analysis, latent class models, posterior predictive P-values, pseudo likelihood ratio, robust goodness-of-fit test.

1. Constrained Latent Class Models

Unconstrained latent class models (ULCM) may be used to explain the responses $x_{ij} \in \{0, 1\}$ of $i = 1, \ldots, N$ persons to $j = 1, \ldots, J$ items. The ULCM assumes that $q = 1, \ldots, Q$ latent classes (it is unknown to which class each person belongs) with class-specific response probabilities explain the dependencies observed among the item responses. Goodman (1974) and Haberman (1974) were the first to present the ULCM and algorithms to obtain maximum likelihood estimates of its parameters.

The likelihood function of the ULCM is given by

$$L(\boldsymbol{\omega}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \prod_{i=1}^{N} \Big[\sum_{q=1}^{Q} P_q(\boldsymbol{x}_i) \omega_q \Big],$$

where, $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_Q]$ denotes the proportion of persons in each of the latent classes, $\boldsymbol{\pi}_q = [\pi_{q1}, \ldots, \pi_{qJ}]$, $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iJ}]$,

$$P_q(\boldsymbol{x}_i) = \prod_{j=1}^{J} \pi_{qj}^{x_{ij}} (1 - \pi_{qj})^{(1 - x_{ij})}$$

denotes the probability of response vector $\boldsymbol{x}_i$ in class q, and π_{qj} denotes the probability of response 1 to item j in class q. In this paper a flat prior (also

called constant, vague or uninformative prior) is used for the parameters of the ULCM, i.e., $\Pr(\boldsymbol{\omega}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q) = 1$. Consequently, the posterior distribution of the parameters of the ULCM is proportional to the likelihood function:

$$\begin{aligned}\text{Post}(\boldsymbol{\omega}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \;\alpha\; & L(\boldsymbol{\omega}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \Pr(\boldsymbol{\omega}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q) \\ & = \prod_{i=1}^{N} \Big[\sum_{q=1}^{Q} P_q(\boldsymbol{x}_i) \omega_q \Big]. \end{aligned} \tag{1.1}$$

In a sense ULCM are exploratory models. The user must search for and (if at all possible) interpret the structure in the estimates of the class-specific probabilities. Several authors have proposed constrained or confirmatory latent class models (CLCM). Heinen (1993), pp. 74-112 gives an overview of CLCM. To name but a few: Lindsay, Clogg and Grego (1991) restrict the class-specific probabilities using a logistic function to obtain the latent class equivalent of the Rasch model; Formann (1985) discusses constraining (some of) the class-specific probabilities to a constant or to each other, and, linear restrictions on the class-specific probabilities; and, Croon (1990) uses inequality constraints to construct a latent class model with ordered latent classes.

An advantage of CLCM is that the user can translate a theory about the response process into constraints on the class-specific probabilities, estimate the model parameters and, using a goodness-of-fit test, confront the theory with the data. In Section 2 of this paper it will be shown how several interesting models can be constructed using constraints of the following types (let c and d denote constants):

$$\pi_{qj} = c_{qj}; \tag{1.2}$$

$$\omega_q = d_q; \tag{1.3}$$

$$\pi_{qj} > c_{qj}, \text{ and, } \pi_{qj} < c_{qj}; \tag{1.4}$$

and, for $q \neq q'$ and/or $j \neq j'$,

$$\pi_{qj} > \pi_{q'j'} + c_{qj}, \text{ and, } \pi_{qj} < \pi_{q'j'} + c_{qj}; \tag{1.5}$$

$$\pi_{qj} > \pi_{q'j'} \times c_{qj}, \text{ and, } \pi_{qj} < \pi_{q'j'} \times c_{qj}; \tag{1.6}$$

and,

$$\omega_q > \omega_{q'}, \text{ and, } \omega_q < \omega_{q'}. \tag{1.7}$$

Subsequently the set of constraints and the value of Q used to construct a ULCM (here constraints will be used to make the model identifiable, see Section 3.2) or CLCM will be denoted by H, since this set constitutes a hypothesis with respect to the response process. In the software used for this paper all constraints may be combined with the exception of (1.3) and (1.7).

Croon (1990, 1991) uses (constrained) maximum likelihood to estimate the parameters of a CLCM built using constraint (1.5) with all constants equal to zero. Although his estimation procedure performs well, he reports problems with the computation of standard errors of the estimates. In Section 3 it will be illustrated that the Gibbs sampler is an excellent and easy-to-use tool for parameter estimation, the computation of posterior standard deviations, and the computation of the posterior distribution of class membership for each person, if constraints like (1.2) through (1.7) have to be taken into consideration.

The goodness of fit of ULCM and CLCM is almost always determined using likelihood ratio tests (see, for example, Goodman (1974), Formann (1985)). The likelihood ratio test has several drawbacks. Its distribution is not approximately chi-squared when models with different numbers of latent classes are compared (see, for example, Heinen (1993), p. 73), and is rather unpredictable with sparse data (Holt and Macready (1989)). Furthermore, using constraints (1.2) through (1.7), the appropriate number of degrees of freedom for a chi-squared approximation to its finite sample distribution is unclear.

Rubin and Stern (1993) present a likelihood ratio test for the comparison of models with different numbers of latent classes. The P-value for this test statistic is evaluated using the posterior predictive distribution of the statistic (Rubin (1984), Meng (1994), Gelman, Meng and Stern (1996)). In Section 4 a likelihood ratio and a pseudo likelihood ratio test will be presented that can be used to compare the fit of ULCM and CLCM to the fit of a saturated multinomial model for the frequencies (see Section 4.1) involved in the (pseudo) likelihood of the ULCM and CLCM. Both tests will be evaluated using posterior predictive P-values.

In Section 5 simulated data sets will be used to illustrate estimation of the parameters of the CLCM using the Gibbs sampler. Furthermore differences between likelihood ratio and pseudo likelihood ratio tests will be illustrated. An example will be presented in Section 6. The paper will be concluded with a short discussion in Section 7.

2. Applications to Educational Testing

Using (1.2) through (1.7) several models can be constructed that may be of use in educational testing. Croon (1991) constrained the class-specific probabilities using

$$\pi_{1j} < \cdots < \pi_{Qj}, \quad \text{for } j = 1, \ldots, J. \tag{2.1}$$

Since, for each item, the response probabilities are increasing with class number, the latent classes that result are ordered. The result is a unidimensional item response model with weak assumptions (the higher the number of the latent class,

the higher the probability of a positive response) about the response process, i.e., class 1 contains the persons of low ability, and class Q the persons of high ability. This model can be seen as the latent class formalization of the Mokken model assuming monotone homogeneity (MH) (Mokken and Lewis (1982), Croon (1991)).

Let the items be ordered according to the proportion of positive responses given in the sample, i.e., $j = 1$ denotes the item with the highest proportion and $j = J$ the item with the lowest proportion. Then, (2.1) in combination with

$$\pi_{q1} > \cdots > \pi_{qJ}, \text{ for } q = 1, \ldots, Q, \tag{2.2}$$

renders an item response model in which both the latent classes and the items are ordered. Within each latent class the item-specific probabilities are ordered according to the proportion of positive responses observed in the sample, i.e., item 1 is the easiest item and item J the most difficult. The resulting model can be seen as the latent class formalization of the Mokken model asuming double monotonicity (DM) (Mokken and Lewis (1982)). Further information about the Mokken model can be found in Mokken (1996) and Molenaar (1996).

Hoijtink and Molenaar (1997) obtain latent classes that are ordered in two dimensions. Since (for each dimension) the class-specific probabilities are increasing with the number of the class for the dimension at hand, the resulting model can be interpreted as a two-dimensional item response model with weak assumptions (the higher the class number on either dimension, the higher the probability of a positive response) on the response process.

Yamamoto (1989) presents a latent class model with two latent classes. The persons in the first class respond according to the Rasch model, the persons in the second class are unscalable and are assumed to respond randomly. Using constraints (2.1) and (2.2) for the first Q classes, and adding class $Q+1$ without restrictions on the class-specific response probabilities, Yamamoto's model can be translated into our framework.

Sometimes abilities consist of a number of components. Reading proficiency probably consists of three components: micro abilities, i.e., knowledge of words; meso abilities, i.e., interpretation of sentences; and, macro abilities, i.e., the ability to relate sentences. If it is assumed that none of the persons in the sample suffer from deficiencies with respect to either component of reading proficiency, constraints like (2.1) might be used to obtain a model that assigns persons to different ability levels. However, sometimes persons may be deficient with respect to one of the components. a deficiency with respect to macro abilities may be modelled using, for example, two latent classes in addition to (2.1). For items measuring micro or meso abilities the class-specific probabilities of the first extra class are restricted to be smaller than those of the second extra class, i.e., the

persons in the first class are of lower ability than the persons in the second class. For items measuring the deficient macro abilities the class specific probabilities in both extra classes may be restricted to be smaller than, for example, .20.

The last example concerns a latent class model that can be used to investigate the hypothesis of whether the population of students consists of three classes: those with a preference for languages and history ($q = 1$); those with a preference for natural sciences ($q = 2$); and, those with a preference for social and behavioral sciences ($q = 3$). Using ten statements ($j = 1, 2, 3$ with respect to languages and history, $j = 4, 5, 6, 7$ with respect to the natural sciences, and $j = 8, 9, 10$ with respect to social and behavioral sciences) the preferences of the students are measured. If the hypothesis is correct, the following CLCM should provide an adequate description of the data (the response 1 indicates a positive attitude to the science involved in the statement):

$$\pi_{1j} > \pi_{1j'}, \quad \text{for } j = 1, 2, 3, \text{ and, } j' = 4, \ldots, 10, \tag{2.3}$$

$$\pi_{2j} > \pi_{2j'}, \text{ for } j = 4, 5, 6, 7, \text{ and, } j' = 1, \ldots, 3, 8, \ldots, 10, \tag{2.4}$$

and,

$$\pi_{3j} > \pi_{3j'}, \text{ for } j = 8, 9, 10, \text{ and, } j' = 1, \ldots, 7. \tag{2.5}$$

The examples given above give only an impression of the CLCM that can be formulated using constraints (1.2) through (1.7). In the next section it will be explained how the Gibbs sampler may be used to estimate the parameters of the CLCM, and to obtain the posterior distribution of class membership for each person in the sample.

3.1. An algorithm for parameter estimation based on the Gibbs sampler

In this section it will be explained how the parameters of the ULCM and CLCM can be estimated using an application of the Gibbs sampler analogous to the one described in Hoijtink and Molenaar (1997), which is based on an algorithm presented by Zeger and Karim (1991). See Casella and George (1992) for an introduction to the Gibbs sampler, and Gelfand and Smith (1990), Tanner (1993), and Tierney (1994) who discuss and describe Markov chain Monte Carlo methods, the general family to which the Gibbs sampler belongs.

The Gibbs sampler can be used to obtain a (dependent) sample from the posterior distribution of the parameters of the ULCM and CLCM. This sample is obtained using a three-step iterative procedure (iterations will be numbered $m = 1, \ldots, M$) in which each parameter (or group of parameters) is sampled from its posterior distribution conditional on the current values of all the other parameters. The sample will be used for three purposes. It is needed for the

computation of posterior predictive P-values, it will be used to compute expected a posteriori estimates (also known as posterior means) and posterior standard deviations for each of the parameters in the model, and it will be used to compute the posterior distribution of class membership for each person.

3.2. Initial values and identification issues

The user has to provide initial values (indexed $m = 0$) for the class weights and the class-specific probabilities. Any set of values that is in agreement with the constraints imposed upon the class-specific probabilities and the class weights can be used. Note that the latter is only possible if the constraints are not mutually conflicting.

Two conditions are necessary (see, Goodman (1974) for a sufficient condition) to avoid identification problems in ULCM. Goodman (1974) (see also Heinen (1993), p. 71) shows that the number of parameters (with dichotomous data $Q(J + 1) - 1$) to be estimated should not be larger than the number of independent frequencies (different response vectors) observed in the data matrix (with dichotomous data at most 2 to the power J minus 1). Furthermore, each latent class should be uniquely labelled, i.e., the classes should not be mutually exchangeable. For ULCM the latter can be achieved using for example

$$\omega_1 < \cdots < \omega_Q, \tag{3.2.1}$$

which ensures that the latent classes are ordered according to the size of the class weights, or,

$$\pi_{11} < \cdots < \pi_{Q1}, \tag{3.2.2}$$

which ensures that the latent classes are ordered according to the size of the class-specific probabilities for item 1.

Note that the ULCM is only barely identified if any of the elements in either (3.2.1) or (3.2.2) are approximately equal in size (this implies that the corresponding latent classes are virtually exchangeable). In such a situation one is well advised to fix one of the parameters involved at a reasonable value, or, add additional constraints (for example, one might combine (a part of) (3.2.1) with (a part of) (3.2.2)). In Section 3.4 an example of this phenomenon will be given.

To avoid identification problems in CLCM the labelling of the classes has to be unique. The same kind of problems and solutions noted for ULCM apply to CLCM. The restriction that the number of parameters has to be smaller than the number of independent frequencies in the data matrix does not necessarily hold for CLCM. It should be noted, however, that a restriction on the number of parameters is still rather sensible (if only to obtain a model that gives a parsimonious description of the data).

3.3. The three steps of the Gibbs sampler

Step 1. Sample the class memberships

For $i = 1, \ldots, N$, sample class membership θ_i^m, which can attain the values $1, \ldots, Q$ (θ_i is a discrete random variable), from

$$\text{Post}(\theta_i|\boldsymbol{\omega}^{m-1}, \boldsymbol{\pi}_1^{m-1}, \ldots, \boldsymbol{\pi}_Q^{m-1}, \boldsymbol{x}_i) = P_{\theta_i}^{m-1}(\boldsymbol{x}_i)\omega_{\theta_i}^{m-1} / \Big[\sum_{\theta_i=1}^{Q} P_{\theta_i}^{m-1}(\boldsymbol{x}_i)\omega_{\theta_i}^{m-1} \Big],$$

which is a multinomial distribution with probabilities $\text{Post}(\theta_i|\cdot)$. Note that $P_{\theta_i}(\boldsymbol{x}_i)$ denotes the probability of response vector $\boldsymbol{x}_i$ given membership of class θ_i.

Step 2. Sample the class-specific probabilities

The class-specific probabilities have to be sampled in a fixed order. For each latent class (the classes are ordered according to the size of q) the class-specific probabilities will be sampled according to the size of j. Each π_{qj}^m is sampled from $\text{Post}(\pi_{qj}|\boldsymbol{\theta}^m, x_{1j}, \ldots, x_{Nj}, L, U)$, where $\boldsymbol{\theta}^m = [\theta_1^m, \ldots, \theta_N^m]$. The admissible range of values for the probability at hand depends on lower bound L and upper bound U. These bounds are determined after an inspection of the constraints to which the class-specific probability that has to be sampled is subjected. Constraints involving another class-specific probability are evaluated using the current value of that class-specific probability (the value from iteration $m-1$ if the probability has not yet been sampled in iteration m, the value from iteration m otherwise). More specifically, L and U are the largest lower bound and the smallest upper bound, respectively, resulting from constraints of the types (1.2), (1.4), (1.5), and (1.6), involving the class-specific probability at hand.

This posterior is given by a truncated beta distribution with parameters $s_{qj}^m + 1$ and $N_q^m - s_{qj}^m + 1$, where N_q^m denotes the number of persons allocated to class q in Step 1 of iteration m, and s_{qj}^m denotes the number of persons allocated to class q in Step 1 of iteration m that respond positively to item j. Note, that the conditional posterior depends only on L, U, the current class membership of each person, and the responses to item j. Note furthermore, that the prior distribution of all parameters, and thus of the class-specific probability at hand is constant (see Section 1).

Using inverse probability sampling (see, Gelman, Carlin, Stern and Rubin (1995), pp. 302-303), it is easy to sample from a truncated beta distribution:

(a) Sample a random number ν from a uniform distribution on the interval $[0, 1]$,
(b) Compute the proportions α and γ of the posterior of π_{qj} that are not admissible due to L and U:

$$\alpha = \int_0^L \text{Beta}(\pi_{qj}|s_{qj}^m + 1, N_q^m - s_{qj}^m + 1) d\pi_{qj},$$

and,

$$\gamma = \int_U^1 \text{Beta}(\pi_{qj}|s_{qj}^m + 1, N_q^m - S_{qj}^m + 1)d\pi_{qj},$$

where $\text{Beta}(\cdot|\cdot)$ is the density of a β-distributed random variable,

(c) Compute π_{qj}^m such that it is the deviate associated with the νth percentile of the admissible part of the posterior of π_{qj}:

$$\alpha + \nu(1 - \alpha - \gamma) = \int_0^{\pi_{qj}^m} \text{Beta}(\pi_{qj}|s_{qj}^m + 1, N_q^m - s_{qj}^m + 1)d\pi_{qj}.$$

Step 3. Sample the class weights

Without constraints (1.3) and (1.7) $\boldsymbol{\omega}^m$ is sampled from $\text{Post}(\boldsymbol{\omega}|\boldsymbol{\theta}^m)$ subject to the constraint

$$\sum_{q=1}^{Q} \omega_q = 1.0. \tag{3.3.1}$$

This posterior is given by a Dirichlet distribution with parameters $N_1^m + 1, \ldots, N_Q^m + 1$. Note that the conditional posterior of the class weights depends only on the current class membership of each person. Note also that the prior of the parameters of the latent class model is constant, and thus that the prior of $\boldsymbol{\omega}$ is constant (see Section 1). The class weights are sampled simultaneously using algorithm DIR-2 from Narayanan (1990) which automatically accounts for (3.3.1). In the first step of DIR-2 for $q = 1, \ldots, Q$, a random variable z_q^m is sampled from a gamma distribution with parameters $N_q^m + 1$ and 1. In the second step $\boldsymbol{z} = [z_1, \ldots, z_Q]$ is normed to obtain $\boldsymbol{\omega}$:

$$\omega_q^m = z_q^m / \sum_{q=1}^{Q} z_q^m,$$

for $q = 1, \ldots, Q$.

It is relatively easy to adjust Narayanan's procedure such that (1.7) is accounted for, i.e., sample z_q^m for $q = 1, \ldots, Q$ from a truncated Gamma distribution with parameters $N_q^m + 1$ and 1, lower bound L' and upper bound U', where L' and U' are the largest lower bound and the smallest upper bound, respectively, implied by constraint (1.7) involving the class weight at hand. If for example $\omega_1 < \omega_2$, then z_1^m is sampled with upper bound z_2^{m-1} and z_2^m is sampled with lower bound z_1^m. Using an inverse probability sampling procedure as described in Step 2 it is easy to sample from a truncated Gamma distribution.

To account for constraint (1.3) the first step of DIR-2 is only executed for the classes whose weights are not fixed at some value. To obtain the corresponding class weights, each z_q has to be divided by $\sum_{q=1}^{Q} z_q^m$. However, this quantity

is unknown since the z_q corresponding with the constrained class weights are unknown. The required summation can be written as a sum of an unknown and a known part:

$$\sum_{q=1}^{Q} z_q^m = \sum_{q\in\text{con}} z_q^m + \sum_{q\in\text{unc}} z_q^m, \tag{3.3.2}$$

where 'con' denotes the set of weights that are constrained using (1.3), and 'unc' denotes the set of weights that are unconstrained. The unknown part (the first summation on the right hand side of (3.3.2)) can be be computed from

$$\sum_{q\in\text{con}} \omega_q + \sum_{q\in\text{unc}} \Big[z_q^m / \big(\sum_{q\in\text{con}} z_q^m + \sum_{q\in\text{unc}} z_q^m \big) \Big] = 1.0, \tag{3.3.3}$$

which is an equation with only one unknown:$\sum_{q\in\text{con}} z_q^m$. Note that (3.3.3) states that the sum of the constrained and unconstrained class weights equals 1.0.

3.4. Convergence

A comprehensive review and evaluation of convergence diagnostics is given by Cowles and Carlin (1996). They conclude that none of the diagnostics is perfect and advise using a combination of them. It appears that further developments are needed before convergence diagnostics can be completely relied upon. In each example to be presented in this paper, the Gibbs sampler was run for $M = 110000$ iterations. Since this is a huge number of iterations, it is possible that the Gibbs sampler has converged and visited all modes of the posterior. However, since latent class models are complicated, it may be that there are still some important modes that have been missed.

The first 10000 iterations of the Gibbs sampler are discarded (these serve as a burn-in period). Furthermore, to save time during the computation of posterior estimates (see Section 3.5) and P-values (see Section 4.2), only every 50th of the subsequent iterations is retained. This leaves $c = 1, \ldots, C$, or more precise, $c = 1, \ldots, 2000$ iterations.

To give an indication of the behavior of the Gibbs sampler, the remaining 2000 iterations will be summarized in two ways: the expected a posteriori estimate of each parameter and its posterior standard deviation, computed for four consecutive sequences of 500 iterations of the remaining 2000 iterations, will be presented; and, the marginal posterior density of each parameter computed from $c = 1, \ldots, 1000$ will be compared with the density computed from $c = 1001, \ldots, 2000$. In Section 5 examples will be given.

Identification problems of the kind discussed in Section 3.2 are easily detected if the Gibbs sampler is used as described above. Suppose that the latent class model presented in Table 1 holds in the population. If the restriction $\omega_1 < \omega_2$

is used, both latent classes are virtually exchangable (since $\omega_1 = \omega_2$) and the ULCM is barely identified. This phenomenon may be detected via an inspection of the output rendered by the Gibbs sampler. As can be seen in Table 1, the class-specific probabilities are initially on the average .2 for class 1 and .8 for class 2, but at some point in the iterative process this will change, i.e., in class 1 probabilities of approximately .8 and in class 2 probabilities of approximately .2. In this case it would be better to make the model identifiable using, for example, the restriction $\pi_{11} < \pi_{21}$.

Table 1. Hypothetical example of virtually exchangeable latent classes.

Parameters	ω_1	π_{11}	π_{12}	π_{13}	ω_2	π_{21}	π_{22}	π_{23}
Population Values	.50	.20	.20	.20	.50	.80	.80	.80
$c = 1$	.45	.19	.22	.22	.55	.82	.81	.75
$c = 2$	.46	.22	.23	.19	.54	.82	.75	.78
$c = 3$	.43	.16	.25	.19	.57	.81	.78	.77
$c = 4$	.44	.24	.18	.18	.56	.83	.81	.81
$c = 5$	.49	.18	.18	.22	.51	.78	.82	.85
$\vdots$								
$c = 350$	.42	.83	.77	.78	.58	.17	.22	.21
$c = 351$	.44	.81	.79	.82	.56	.19	.21	.17
$c = 352$	.49	.78	.82	.79	.51	.22	.18	.21
$c = 353$	.47	.79	.81	.78	.53	.21	.18	.20
$c = 354$	.43	.80	.85	.79	.57	.20	.22	.19
$\vdots$								

3.5. Posterior classifications, estimates and covariance matrix

Using ξ and ξ' as generic symbols to represent any of the parameters in (1.1), the expected a posteriori (EAP(ξ)) estimates, and each of the elements from the posterior covariance matrix of the parameters (Cov (ξ, ξ')) are given by

$$\text{EAP}(\xi) \approx \sum_{c=1}^{C} \xi^c / C, \tag{3.5.1}$$

and

$$\text{Cov}\,(\xi, \xi') \approx \sum_{c=1}^{C} (\xi^c \xi'^c / C) - \sum_{c=1}^{C} (\xi^c / C) \sum_{c=1}^{C} (\xi'^c / C), \tag{3.5.2}$$

respectively. The $\approx$ in (3.5.1) and (3.5.2) reflects that the accuracy, with which the summations in (3.5.1) and (3.5.2) approximate the integrals over uni- and bivariate marginals of (1.1), depends on the size of C.

For $c = 1, \ldots, 2000$, Step 1 of the procedure described in Section 3.3 assigns each person to one of the Q latent classes. The resulting frequency distribution

can, for each person, be used for classification purposes. One might conclude for example that the posterior probability that a person prefers natural sciences is .90, or, that it is not clear whether a person prefers natural sciences, languages and history, or psychology and sociology, since the corresponding posterior probabilities are .32, .35 and .33 respectively.

4.1. (Pseudo) Likelihood ratio tests

The fit of the ULCM and the CLCM will be evaluated using a likelihood ratio test and a pseudo likelihood ratio test. Both tests compare the fit of the CLCM with the fit of a multinomial model for the frequencies that are used in the (pseudo) likelihood of the CLCM.

The likelihood ratio (LR) test (see, for example, Formann (1985)) is given by

$$\mathrm{LR}(\boldsymbol{X}, \boldsymbol{\xi}) = -2 \sum_{p=1}^{P} N(\boldsymbol{x}_p) \log[M(\boldsymbol{x}_p|\boldsymbol{\xi})/N(\boldsymbol{x}_p)], \tag{4.1.1}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$, P denotes the number of different response vectors $\boldsymbol{x}_p$ that are possible (with dichotomous data 2 to the power J), and $N(\cdot)$ the number of response vectors $\boldsymbol{x}_p$ observed in the data matrix for which LR is computed. The expected number of response vectors $\boldsymbol{x}_p$ in the data matrix for which LR is computed is given by

$$M(\boldsymbol{x}_p|\boldsymbol{\xi}) = N \sum_{q=1}^{Q} P_q(\boldsymbol{x}_p)\omega_q. \tag{4.1.2}$$

Since $\boldsymbol{\xi}$ is unknown, (4.1.2) cannot be computed. However, in the next section it will be shown that this problem can be solved if LR is evaluated as a discrepancy measure. For a more general discussion of discrepancy measures see Meng (1994) and Gelman, Meng, and Stern (1996).

As will be illustrated in Section 5, LR is rather sensitive to the presence of outliers in the data even if the set of restrictions H provide a good description of the response process. The posterior predictive P-values (see, Section 4.2) used for the evaluation of LR might indicate misfit due to the presence of outliers (persons whose response vectors are rather unlikely given the class-specific probabilities of each latent class). Sometimes one can incorporate outliers in the model, an example is the model of Yamamoto (1989) discussed in Section 2. On other occasions one just wants to know if the model holds for most of the sample without having to bother about a (relatively) small number of outliers. The latter is achieved using the following pseudo likelihood ratio (PLR) test which is sensitive with respect to misspecifications of the restrictions in H and the number

of latent classes Q used, but is robust with respect to outliers:

$$\begin{aligned}\mathrm{PLR}(\boldsymbol{X},\boldsymbol{\xi}) &= -2\log(\mathrm{PL}_H/\mathrm{PL}_M)\\ &= -2\sum_{j\neq j'}\sum_{v=0}^{1}\sum_{w=0}^{1} N(X_j=v, X_{j'}=w)\log\Big[\frac{M(X_j=v, X_{j'}=w|\boldsymbol{\xi})}{N(X_j=v, X_{j'}=w)}\Big],\end{aligned} \tag{4.1.3}$$

where PL_H denotes the pseudo likelihood (see, for example, Gourieroux, Monfort and Trognon (1984)) of the ULCM or CLCM, and PL_M the pseudo likelihood of the corresponding multinomial model:

$$\mathrm{PL}_H = \prod_{j\neq j'}\prod_{v=0}^{1}\prod_{w=0}^{1}[M(X_j=v, X_{j'}=w|\boldsymbol{\xi})/N]^{N(X_j=v,X_{j'}=w)};$$

and

$$\mathrm{PL}_M = \prod_{j\neq j'}\prod_{v=0}^{1}\prod_{w=0}^{1}[N(X_j=v, X_{j'}=w)/N]^{N(X_j=v,X_{j'}=w)}.$$

The pseudo likelihood ratio test involves, for each pair of items, a comparison of $N(X_j=v, X_{j'}=w)$ (the observed number of persons responding v to item j and w to item j' in the data matrix) with the corresponding expected number of persons:

$$M(X_j=v, X_{j'}=w|\boldsymbol{\xi}) = N\sum_{q=1}^{Q} P_q(X_j=v, X_{j'}=w)\omega_q. \tag{4.1.4}$$

Note that the comment following (4.1.2) also applies to (4.1.4). Note also that

$$P_q(X_j=v, X_{j'}=w) = \pi_{qj}^{v}(1-\pi_{qj})^{(1-v)}\pi_{qj'}^{w}(1-\pi_{qj'})^{(1-w)}.$$

In contrast to the likelihood ratio test which uses all the information in the response vectors $\boldsymbol{x}$, the pseudo likelihood ratio test uses only the information available in pairs of item responses. The result is a test statistic that is more robust against outliers (illustrations follow in Section 5.2): it is virtually impossible to give responses to a pair of items that are outliers, there will always be many persons in the sample that respond 1 to both items, that respond 0 to both items, or respond 1 to one of the items; on the other hand a response vector can easily be an outlier, e.g., a person responding 1 to nine out of ten items while none of the other persons in the sample respond 1 to more than six items.

4.2. Posterior predictive P-values

Rubin (1984) presents a Bayesian method to investigate goodness of fit. Meng (1994), and Gelman, Meng and Stern (1996) extend the method presented

by Rubin. One of their results is the discrepancy measure. The posterior predictive P-value for a discrepancy measure is given by

$$P = \Pr\{D(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\xi}) \geq D(\boldsymbol{X}, \boldsymbol{\xi}) | \boldsymbol{X}, H\}, \tag{4.2.1}$$

i.e., the probability that (what Meng (1994) calls) the discrepancy measure $D(\cdot)$ (which is a function of data and parameters) computed for a replication of the data matrix $\boldsymbol{X}^{\text{rep}}$ and $\boldsymbol{\xi}$, is equal to or exceeds the value computed for the observed data matrix $\boldsymbol{X}$ and $\boldsymbol{\xi}$. Here $D(\cdot)$ may be the likelihood ratio test (4.1.1) which leads to $P = \Pr\{\text{LR}(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\xi}) \geq \text{LR}(\boldsymbol{X}, \boldsymbol{\xi}) | \boldsymbol{X}, H\}$, or the pseudo likelihood ratio test (4.1.3) which leads to $P = \Pr\{\text{PLR}(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\xi}) \geq \text{PLR}(\boldsymbol{X}, \boldsymbol{\xi}) | \boldsymbol{X}, H\}$. The P-value is computed over the joint posterior distribution of the replicated data and $\boldsymbol{\xi}$ conditional on the observed data $\boldsymbol{X}$ and H:

$$f(\boldsymbol{X}^{\text{rep}}, \boldsymbol{\xi} | \boldsymbol{X}, H) = f(\boldsymbol{X}^{\text{rep}} | \boldsymbol{\xi}) \text{Post}(\boldsymbol{\xi} | \boldsymbol{X}, H).$$

The interested reader is referred to Meng (1994) and Gelman, Meng and Stern (1996) for a discussion of discrepancy measures. In Section 5 some experiences with discrepancy measures based on the likelihood ratio tests introduced in the previous section will be presented.

The probability in (4.2.1) may be evaluated using a three-step simulation procedure:

Step 1: The procedure described in Section 3.3 is used to sample $c = 1, \ldots, C$ parameter vectors $\boldsymbol{\xi}^c$ from $\text{Post}(\boldsymbol{\xi} | \boldsymbol{X}, H)$.

Step 2: For $c = 1, \ldots, C$ a replication of the data matrix, denoted by $\boldsymbol{X}^c$, is simulated conditional upon each of the sampled parameter vectors $\boldsymbol{\xi}^c$. Each simulation consists of two steps.

— First, each of N persons is assigned to one of the Q latent classes. This is achieved using a sample of size N from a multinomial distribution with probabilities $\boldsymbol{\omega}^c$.

— Second, for each person the item responses are simulated using the class-specific probabilities corresponding to the class q to which person i was assigned. To do this, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, π^c_{qj} is compared with ν^c_{ij} (a random number sampled from a uniform distribution on the interval [0,1]). If $\pi^c_{qj} > \nu^c_{ij}, x^c_{ij} = 1$, otherwise $x^c_{ij} = 0$.

Step 3: The discrepancy measure is computed for corresponding pairs of replicated and observed data matrices. The posterior predictive P-value can then be approximated using the proportion of times that $D(\boldsymbol{X}^c, \boldsymbol{\xi}^c)$ is equal to or exceeds the corresponding $D(\boldsymbol{X}, \boldsymbol{\xi}^c)$. The quality of the approximation depends on C.

5.1. The estimation procedure illustrated with simulated data

In this section a number of simulated data sets will be analysed to illustrate some features of the estimation and testing procedures proposed in Sections 3 and 4. The first three columns of Table 2 (repeated in Table 3) give the population parameters used to simulate two data sets. Table 2 presents the expected a posteriori estimates (3.5.1) obtained using the ULCM with restriction (3.2.1), the MH-model (2.1), and the DM-model (2.1) combined with (2.2), to analyse the first data set ($J = 10, N = 250$). Table 3 presents the estimates obtained using the ULCM and the MH-model to analyse the second data set ($J = 10, N = 500$).

Table 2. Expected a posteriori estimates for three models ($N = 250$). the first line with numbers gives the class weights, the other lines give the class-specific probabilities.

	Population			ULCM			MH			DM		
	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$
	.20	.35	.45	.17	.37	.46	.29	.29	.42	.38	.12	.50
$j=1$	.25	.80	.90	.72	.47	.90	.35	.76	.91	.49	.81	.92
$j=2$	.45	.50	.90	.53	.51	.89	.47	.61	.89	.46	.74	.89
$j=3$	.30	.75	.85	.67	.45	.90	.31	.72	.91	.43	.69	.87
$j=4$	.40	.60	.80	.60	.47	.84	.38	.65	.85	.40	.64	.83
$j=5$	.35	.65	.70	.68	.46	.76	.35	.69	.78	.37	.58	.77
$j=6$	.25	.35	.70	.29	.28	.67	.25	.34	.67	.29	.44	.70
$j=7$	.10	.45	.60	.39	.34	.57	.24	.46	.58	.25	.39	.68
$j=8$	.05	.20	.80	.29	.20	.73	.15	.32	.75	.18	.32	.67
$j=9$	.10	.30	.75	.26	.20	.73	.14	.32	.74	.14	.26	.66
$j=10$	.05	.15	.85	.21	.04	.81	.03	.17	.85	.03	.16	.64

Looking at Table 2, it can be seen that the estimates of the class-specific probabilities obtained using the ULCM, are quite different from their population values. It appears that the data do not contain enough information to recover the population values of the class-specific probabilities using expected a posteriori estimates. The latter was confirmed after an inspection of the posterior standard deviation (SD) and 95% central credibility intervals (CI) for the parameters. Two examples: the SD and CI for the class-specific probability of item 1 in class 1 were .27 and [.20,.98] respectively; the same quantities for item 1 in class 2 were .15 and [.29,.77] respectively. Given the fact that these parameters are restricted to values between zero and one, both the SD and the CI are considered to be huge. The parameters of class 3 (the largest class) are much better determined (the SD and CI for the first item in class 3 were .03 and [.84,.94] respectively). Apparently, for the ULCM and the data generated using the population model presented in Table 2, the weight of a class influences the accuracy with which the corresponding class-specific probabilities are estimated. The estimates obtained

for the MH-model are more accurate. Through the use of contraint (2.1) extra information with respect to the class-specific probabilities was 'bought'. The SD and CI of item 1 in class 1 were .09 and [.19,.49], in class 2, .09 and [.59,.89]. This is an improvement with respect to the corresponding quantities obtained for the ULCM. Note that the estimates of the class-specific probabilities are in agreement with constraint (2.1), i.e., increasing with q.

Table 3. Expected a posteriori estimates for two models ($N = 500$). the first line with numbers gives the class weights, the other lines give the class-specific probabilities.

	Population			ULCM			MH		
	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$
	.20	.35	.45	.21	.33	.46	.22	.33	.44
$j=1$	.25	.80	.90	.27	.74	.89	.30	.74	.89
$j=2$	.45	.50	.90	.51	.54	.90	.48	.58	.91
$j=3$	.30	.75	.85	.29	.68	.92	.28	.71	.92
$j=4$	.40	.60	.80	.41	.60	.82	.40	.61	.82
$j=5$	.35	.65	.70	.31	.73	.72	.35	.68	.74
$j=6$	.25	.35	.70	.30	.25	.70	.25	.31	.70
$j=7$	.10	.45	.60	.17	.38	.58	.17	.39	.58
$j=8$	.05	.20	.80	.10	.25	.76	.10	.26	.78
$j=9$	.10	.30	.75	.10	.26	.76	.10	.29	.77
$j=10$	.05	.15	.85	.06	.17	.80	.05	.20	.81

The estimates obtained for the class weights and specific probabilities for the DM-model (see, Table 2) are quite different from the population values. This is not surprising since the population values are not in agreement with (2.1) and (2.2), whereas the estimates obtained for the DM-model are, i.e., increasing with q and decreasing with j.

In Table 3 the ULCM and MH-model were used to analyse a larger data set ($N = 500$ instead of $N = 250$) simulated using the same population as in Table 2. The estimates obtained are closer to the population values than the corresponding estimates presented in Table 2. The latter is an indication of consistency of the estimates.

For three of the parameters the behavior of the Gibbs sampler is presented in Table 4. The results are representative for all parameters in all models that were analysed. As can be seen there are slight fluctuations in the EAP and SD of the class-specific probability for the ULCM with $N = 250$ over four sequences of 500 iterations of the Gibbs sampler. The EAP and SD for the class-specific probability for the MH-model with $N = 250$ and $N = 500$ is very stable. In Figure 1 reconstructions of the marginal posterior density of item 1 in class 1 of the

Table 4. An indication of the behavior of the Gibbs sampler for the class-specific probability of item 1 in class 1 of the ULCM with $N = 250$, the MH-model with N=250, and the MH-model with $N = 500$.

Model		Iteration	EAP	SD
ULCM:	$N = 250$	$c = 1, \ldots, 500$	.72	.27
		$c = 501, \ldots, 1000$	.73	.26
		$c = 1001, \ldots, 1500$	.70	.28
		$c = 1501, \ldots, 2000$	.72	.27
MH:	$N = 250$	$c = 1, \ldots, 500$	.34	.09
		$c = 501, \ldots, 1000$	.35	.09
		$c = 1001, \ldots, 1500$	.34	.10
		$c = 1501, \ldots, 2000$	.35	.09
MH:	$N = 500$	$c = 1, \ldots, 500$	.30	.08
		$c = 501, \ldots, 1000$	.30	.08
		$c = 1001, \ldots, 1500$	.29	.08
		$c = 1501, \ldots, 2000$	.30	.08

ULCM with $N = 250$, are displayed for $c = 1, \ldots, 1000$ and $c = 1001, \ldots, 2000$ respectively. As can be seen, the densities are rather similar. Given these results, it is possible that the Gibbs sampler has converged and visited all modes of the posterior. However, as stated before in Section 3.4, since latent class models are complicated, it may be that there are still some important modes that have been missed.

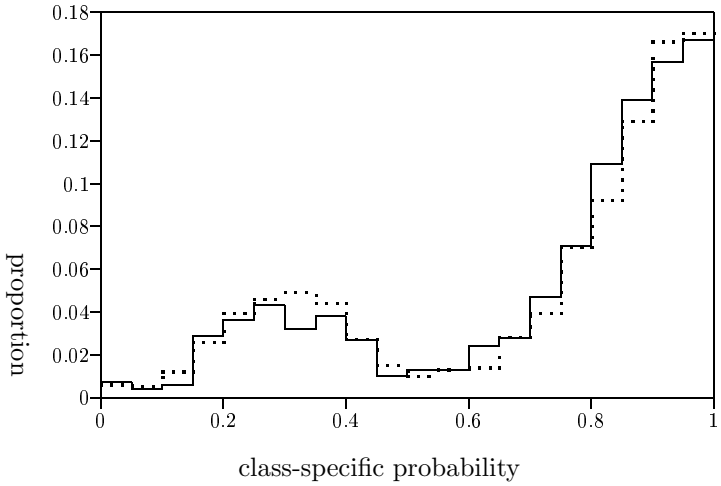

Figure 1. Posterior distribution of the class-specific probability of item 1 in class 1 of the ULCM with $N = 250$ computed for $c = 1, \ldots, 1000$ (solid line), and for $c = 1001, \ldots, 2000$ (dashed line).

5.2. Goodness of fit illustrated with simulated data

Table 5 displays posterior predictive P-values resulting from the analyses of data sets with a different size, simulated according to the population model displayed in Table 2, and analysed using the ULCM, MH-model, and DM-model with varying numbers of latent classes. If $N = 250$, the ULCM and MH-models with $Q = 3$, and $Q = 2$ fit almost equally well. With this sample size there is not enough power to distinguish among $Q = 3$ and $Q = 2$. Both models with $Q = 1$ are rejected, i.e., it is clear that the sample is not homogeneous, and that at least two latent classes have to be distinguished.

Table 5. Posterior predictive P-values for the (pseudo) likelihood ratio tests computed for data simulated according to the population model displayed in Table 2.

N	Q	Model	LR	PLR
250	3	ULCM	.54	.52
250	2	ULCM	.43	.50
250	1	ULCM	.00	.00
250	3	MH	.58	.52
250	2	MH	.48	.50
250	1	MH	.00	.00
500	2	MH	.22	.42
1000	2	MH	.07	.26
250	3	DM	.23	.10
500	3	DM	.27	.03
1000	3	DM	.00	.00

Increasing the sample size from 250 to 500 and 1000 (still for the same 3-class population model), the goodness-of-fit tests gain power. It can be seen that the fit of the MH-model with $Q = 2$ and $N = 1000$ is worse than the fit with $Q = 2$ and $N = 250$. Furthermore, where the DM-model has an acceptable fit for $N = 250$ it has to be rejected for $N = 1000$.

In Table 6 the population parameters for a preference model (see Section 2) with three latent classes are displayed. Table 7 presents the posterior predictive P-values resulting from the analyses of data sets with a different size, simulated according to this preference model, and analysed using the CLCM presented in (2.3), (2.4) and (2.5). To illustrate the robustness of the pseudo likelihood ratio tests against outliers, ten response vectors consisting of response 1 nine times and response 0 once (the first response vector contains a zero for item 1, the second for item 2 etc.) were added once, twice and three times to data sets with $N = 250$ and $N = 1000$. Note that these response vectors are considered to be outliers since response vectors containing response 1 nine times are unlikely given

Table 6. Population parameters for the preference model. The first line with numbers gives the class weights, the other lines give the class-specific probabilities.

Item	$q = 1$	$q = 2$	$q = 3$
	.20	.35	.45
$j = 1$	.80	.20	.20
$j = 2$	.80	.20	.20
$j = 3$	.80	.20	.20
$j = 4$	.20	.80	.20
$j = 5$	.20	.80	.20
$j = 6$	.20	.80	.20
$j = 7$	.20	.80	.20
$j = 8$	.20	.20	.80
$j = 9$	.20	.20	.80
$j = 10$	.20	.20	.80

the population parameters displayed in Table 6. For the CLCM with $N = 250$, the likelihood ratio test starts to indicate a lack of fit when 20 or more outliers are added to the data file, while the pseudo likelihood ratio test is still acceptable even if 30 outliers are added to the data file. For the CLCM with $N = 1000$, the likelihood ratio tests indicate a lack of fit when 30 outliers are added to the data, while the pseudo likelihood ratio tests do not. A general inspection of the pattern of P-values observed in Table 7, indicates that the pseudo likelihood ratio test is more robust to outliers than the likelihood ratio test.

Table 7. Posterior predictive P-values for the (pseudo) likelihood ratio tests computed for data simulated according to the preference model.

N	Q	Model	LR	PLR
250	3	Pref.	.41	.48
250+10	3	Pref.	.13	.37
250+20	3	Pref.	.01	.17
250+30	3	Pref.	.00	.07
1000	3	Pref.	.72	.50
1000+10	3	Pref.	.40	.49
1000+20	3	Pref.	.05	.43
1000+30	3	Pref.	.00	.32

6. Example: LSAT Section 7

Bock and Aitkin (1981) analysed 5 items from section 7 of the Law School Admission Test (LSAT), and concluded that a two-dimensional normal ogive

model provided a better fit than a one-dimensional normal ogive model. Note that the data can be found in Bock and Lieberman (1970). In this section it will be illustrated that the Mokken model assuming double monotonicity (a nonparametric alternative for the one-dimensional two-parameter normal ogive model), formalized in (2.1) and (2.2), provides an adequate description of section 7 of the LSAT. Stated otherwise, one underlying latent trait is sufficient to explain the responses to section 7 of the LSAT.

The results of the analyses are displayed in Table 8. As can be seen one latent class is clearly insufficient, but the fit of the two and three class solution is acceptable. The three class solution is preferred, it can be used to distinguish persons with low, medium and high abilies (the class-specific probabilities of the first class are clearly lower that those of the second class which in turn are clearly lower than those of the third class). Note that $j = 1$ refers to the first item in the table presented by Bock and Lieberman (1970), $j = 2$ to the second item etc.

Table 8. Results of the analysis of section 7 of the LSAT with the Mokken model assuming double monotonicity. The first line with numbers gives the class weights, the last two lines give posterior predictive P-values, and the lines in between give the class-specific probabilities.

	$Q=1$	$Q=2$		$Q=3$		
	$q=1$	$q=1$	$q=2$	$q=1$	$q=2$	$q=3$
	1.0	.29	.71	.11	.32	.57
$j=4$	.60	.32	.71	.15	.47	.73
$j=2$	.65	.36	.77	.23	.51	.81
$j=3$	.77	.44	.89	.31	.63	.91
$j=1$	.82	.61	.90	.48	.73	.92
$j=5$	.84	.68	.91	.61	.77	.94
LR	.00	.05		.10		
PLR	.00	.37		.34		

7. Discussion

This paper proposed estimation and testing procedures for constrained latent class models based on the Gibbs sampler and discrepancy measures, respectively. In Section 5 the results of the analyses of a number of simulated data sets were presented. The intention of the simulations was to illustrate the feasibility, and some interesting features, of the proposed estimation and testing procedures.

It was shown that parameter estimates obtained for the ULCM are rather inaccurate and undetermined for smaller samples ($N = 250$), but that the addition of constraints solves this problem. It was also shown that the estimates are indeed consistent with the constraints specified in H, and also appear to be

consistent in a statistical sense. The robustness with respect to outliers of the pseudo likelihood ratio test was illustrated.

This paper and the simulations provide handholds for the application of constrained latent class models. Lacking at this point in time are frequency evaluations (the analyses of repeated samples from the same population) of the procedures proposed. Since the frequentist properties are theoretically (probably) intractable, and simulations would take a long time (the average analysis described in this paper took about 15 hours on a pentium pc), these will probably not be available for some time.

References

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* **46**, 443-459.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika* **35**, 179-197.

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46**, 167-174.

Cowles, M. K. and Carlin B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.* **91**, 883-904.

Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British J. Math. Statist. Psych.* **43**, 171-192.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British J. Mathe. Statist. Psych.* **44**, 315-331.

Formann, A. K. (1985). Constrained latent class models: theory and applications. *British J. Math. Statist. Psych.* **38**, 87-111.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85**, 972-985.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall, London.

Gelman, A., Meng, X. and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6**, 733-807.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable model. *Biometrika* **61**, 215-231.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica* **54**, 681-700.

Haberman, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. *Ann. Statist.* **2**, 911-924.

Heinen, T. (1993). *Discrete Latent Variable Models.* Tilburg University Press, Tilburg.

Hoijtink, H. and Molenaar, I. W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika* **62**, 171-190.

Holt, J. A. and MacReady, G. B. (1989). A simulation of the difference chi- square statistic for comparing latent class models under violation of regularity conditions. *Appl. Psych. Measurement* **13**, 221-232.

Lindsay, B., Clogg, C. C. and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* **86**, 96-107.

Meng, X. L. (1994). Posterior Predictive p-Values. *Ann. Statist.* **22**, 1142-1160.

Mokken, R. J. (1996). Nonparametric models for dichotomous responses. In *Handbook for Modern Item Response Theory* (Edited by W. J. van der Linden and R. K. Hambleton), 351-368, Springer, New York.

Mokken, R. J. and Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Appl. Psych. Measurement* **6**, 417-430.

Molenaar, I. W. (1996). Nonparametric models for polytomous responses. In *Handbook for Modern Item Response Theory* (Edited by W. J. van der Linden and R. K. Hambleton), 369-380, Springer, New York.

Narayanan, A. (1990). Computer generation of Dirichlet random vectors. *J. Statist. Comput. Simulation* **36**, 19-30.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151-1172.

Rubin, D. B. and Stern, H. L. (1993). Testing in latent class models using a posterior predictive check distribution. In *Latent Variables Analysis. Applications for Developemental Research* (Edited by A. von Eye and C. Clogg), 420-438, SAGE, London.

Tanner, M. A. (1993). *Tools for Statistical Inference.* Springer, New York.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701-1762.

Yamamoto, K. (1989). Hybrid model of IRT and latent class models. Research report no. 89-41, Educational Testing Service, Princeton, NJ.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**, 79-86.

Department of Methodology and Statistics, FSW, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands.

E-mail: H.Hoijtink@fsw.ruu.nl

(Received February 1996; accepted December 1997)