Chapter 4

Bayesian Computational Methods for Inequality Constrained Latent Class Analysis

Olav Laudy, Jan Boom, and Herbert Hoijtink
Utrecht University

4.1 Introduction

Exploratory latent class analysis (ELCA) (Clogg, 1981; Goodman, 1974; Haberman, 1988; Vermunt, 1997) is used to group responses x_{ij} of persons $i = 1, \ldots, N$ to items $j = 1, \ldots, J$ into classes $q = 1, \ldots, Q$ such that persons with similar responses are assigned to the same class. In this chapter we restrict ourselves to dichotomous data $x_{ij} \epsilon \{0, 1\}$. Each class q is characterized by J class specific probabilities π_{qj} indicating the probability of the response '1' on item j in class q and a weight ω_q indicating the unconditional probability that a person's latent class membership τ equals q. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, $\boldsymbol{\theta} = [\boldsymbol{\omega}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q]$, $\boldsymbol{\pi}_q = [\pi_{q1}, \ldots, \pi_{qJ}]$, $\mathbf{x}_i = [x_{i1}, \ldots, x_{iJ}]$ and $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_Q]$. The density of the data given the

parameters of ELCA is then given by

$$P(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(\mathbf{x}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \Big[\sum_{q=1}^{Q} P(\mathbf{x}_i, \tau = q \mid \boldsymbol{\theta}) \Big] = \prod_{i=1}^{N} \left[\sum_{q=1}^{Q} \omega_q \prod_{j=1}^{J} \pi_{qj}^{x_{ij}} (1 - \pi_{qj})^{(1-x_{ij})} \right]. \quad (4.1)$$

A key question in ELCA is into how many homogeneous subgroups the sample should be divided? Usually fit measures (Everitt, 1988; Lin & Dayton, 1997) are used to determine which number of classes is optimal. Another question concerns the interpretation of the resulting classes. Sometimes classes can be interpreted independent of other classes. As is illustrated in the next section, one class may account for persons with highly developed emotional skills, while an other class accounts for persons with highly developed social skills. It can also be that classes can be ordered with respect to one or more underlying dimensions (Croon, 1990). An example of the latter is an ELCA resulting in three latent classes that can be used to order persons with respect to different levels of social skills (a one-dimensional ordering). It even might be the case that the persons can be ordered with respect to two dimensions, for example, the combinations of levels of social skills and the levels of emotional skills.

A researcher using exploratory analysis behaves as if his research field has not yet been explored very thoroughly, and theories are not yet fully developed. After the execution of an exploratory analysis, a researcher has to determine whether the outcome is in accordance with an existing theory, or that a new theory is emerging. This approach has two drawbacks. First of all, it may not at all be clear which theory corresponds best to the outcomes. This may lead to over-interpretation and guessing. Secondly, scientific progress may be larger if the current state of affairs (existing knowledge and theories) are properly accounted for in the statistical models used for the analyses.

ELCA has been done in areas that have been thoroughly explored, and where theories are well developed, for example, Boom, Hoijtink, and Kunnen (2001) and Jansen and Van der Maas (1997). They use ELCA to analyze data with respect to the Piagetian Balance Scale Task. In section 4.5.1 new data with respect to this task are analyzed using confirmatory latent class analysis (CLCA). There it is also shown how CLCA can be used to refine (the best of) the existing theories, that is, how a new theory can be generated using the old theory as the point of departure.

In this chapter, a specific form of CLCA is proposed (Hoijtink & Molenaar, 1997; Hoijtink, 1998; Hoijtink, 2001). The approach allows a theory

to be translated into a CLCA using inequality constraints among the parameters of the model. This can be done for several competing theories. Two fit measures are presented that can be used to select the model that receives the most support from the data.

4.2 Translation of Theories into CLCA

Several models can be constructed using constraints of the following types for $q \neq q'$ and/or $j \neq j'$

$$\begin{aligned} \pi_{qj} &> \pi_{q'j'}, \\ \pi_{qj} &< \pi_{q'j'}. \end{aligned}$$

To start with a simple example, suppose that persons have to respond to ten items. The first five items can be answered using skills related to social qualities (e.g. do you think you have a good understanding of other people?), the others using skills related to emotional qualities (e.g., do you easily succeed in managing yourself?). The answers to these questions are coded 1 (well-developed) and 0 (undeveloped). Thus, the response vector of each respondent has ten scores with realization 1 or 0. Suppose, several theories exist for these data. A researcher thinks skills related social and emotional are not distinct, leading to the conclusion that there are only two groups of persons: persons who have a higher (social/emotional) intelligence, and persons who have a lower intelligence. This *common intelligence theory* can be translated into a latent class model with two latent classes. The class specific probabilities for the first class are all high, the class specific probabilities for the second class are all low, thus meaning that the persons who have both well developed social and emotional skills are allocated in class one, and the less intelligent persons who have both less developed social and emotional skills are allocated in class two. In terms of restrictions (see Table 4.1): For the common intelligence theory, the class specific probabilities of all items in the first class are restricted to be larger than those of all items in the second class. Note that j is used to indicate item numbers, ω_1 denotes the proportion of persons in class 1, and π_{2j} denotes the probability of responding '1' to item j in class 2.

Another researcher might not agree with the common intelligence theory and states that there are indeed two groups of persons, but one group has higher social related skills, while the other group has higher emotional related skills. From this *specific (social/emotional) intelligence theory* it can be inferred that in one class the probabilities of responding 'developed'—that is, the response indicates that the person has well-developed skills—to the social items are higher than for the emotional items, while for the other

Table 4.1: Items and Restrictions on the Response Probabilities for Common (Social/Emotional) Intelligence Theory.

Item type	Items	Restrictions		
Social	1-5	π_{1j}	$>$	π_{2j}
Emotional	6-10	π_{1j}	$>$	π_{2j}

Table 4.2: Inequality Constraints for the Specific Intelligence Theory.

Item type	Items	Restrictions		
Social	1-5	π_{1j}	$>$	π_{2j}
		$>$		$<$
Emotional	6-10	π_{1j}	$<$	π_{2j}

class the probabilities of responding 'developed' to the emotional items are higher than for the social items. This theory can be translated into a CLCA as indicated in Table 4.2: the first five items in the first class have probabilities that are restricted to be larger than the first five items in the second class. The first five items in the first class are also restricted to be larger than the last five items in the first class. The last five items in the second class have probabilities that are restricted to be larger than the last five items in the first class. The last five items in the second class are restricted to be larger than the first five items in the second class.

An alterative display of the inequality constraints for the 'specific intelligence' theory is given in Table 4.3. Here the inequality constraints are implicit; for example, a minus sign indicates a class specific probability is restricted to be smaller than all the class specific probabilities corresponding to a plus sign. This type of display is used in section 4.5.1, where the display with inequality signs is too complicated or impossible. The inequality constrains are implicit: - < +. Note that a minus sign is not restricted with respect to any other minus sign, and a plus sign is not restricted with respect to any other plus sign.

Table 4.3: Alternative Display of the Inequality Constraints for the Specific Intelligence Theory.

Item type	Items	Restrictions	
Social	1-5	+	-
Emotional	6-10	-	+

4.3 Estimates for the CLCA

In this section we explain how estimates of the parameters are obtained. The general algorithm is described by Gelfand, Smith, and Lee (1992) and the direct application to CLCA can be found in Hoijtink (1998). The basic principle is to use the posterior distribution to obtain a sample of the model parameters. This sample can be seen as a discrete representation of the posterior distribution. With this sample, further calculations are easy, for example, the average of the sampled values is the expected a posteriori (EAP) estimate of a parameter, and the 2.5-th and 97.5-th percentile of the sampled values constitute a 95% central credibility interval. Since it is not trivial to obtain a sample from a multivariate posterior distribution, the Gibbs sampler is applied. This algorithm renders a sample from the joint posterior of the parameters by repeatedly sampling from conditional distributions, that is, the distribution of the parameter at hand given all the other parameters.

4.3.1 Posterior Distribution

The density of the data given the parameters of the model is given by Equation 4.1. For each model $k = 1, \ldots, K$, where K denotes the number of models under consideration, the set of inequality constraints is denoted by H_k. The latter will be included in the posterior distribution via the prior distribution. In this chapter, all the priors are chosen to be uniform for all combinations of parameter values allowed by H_k. Note that since information about the models is included in the prior distributions via inequality constraints, in that respect the priors are informative. The conjugate prior for a (constrained) class specific probability is a (truncated) Beta(1,1) distribution. The conjugate prior for the class weights is a Dirichlet distribution parameterized such that a priori all combinations of weight values summing to one are equally likely, that is, Dirichlet$(\alpha_1, \ldots, \alpha_Q)$, with $\alpha_q = 1$. The resulting posterior $P(\boldsymbol{\theta} \mid \mathbf{X}, H_k)$ is proportional to the product of the density of the data $P(\mathbf{X} \mid \boldsymbol{\theta})$ and the (truncated) proportional prior

$P(\boldsymbol{\theta} \mid H_k)$, that is

$$P(\boldsymbol{\theta} \mid \mathbf{X}, H_k) \propto P(\mathbf{X} \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta} \mid H_k),$$

where $P(\boldsymbol{\theta} \mid H_k)$ has the value 1 if $\boldsymbol{\theta}$ is in accordance with the constraints imposed by H_k, and 0 otherwise.

4.3.2 Gibbs Sampler

The Gibbs sampler is an iterative procedure. In iteration $r = 0$ initial values have to be provided for the class weights and the class specific probabilities. Any set of values that is in agreement with the constraints imposed upon the parameters can be used. Each iteration $r = 1, \ldots, R$ consists of three steps:

Step 1: For $i = 1, \ldots, N$, sample class membership $\tau_{i,r} \in \{1, \ldots, Q\}$ from its posterior distribution given the current values (i.e., the values sampled in iteration $r-1$) of the class weights, the class specific probabilities and the data. This conditional posterior is a Multinomial distribution with probabilities

$$P(\tau_{i,r} = q \mid \mathbf{x}_i, \boldsymbol{\theta}_{r-1}) = \frac{P(\mathbf{x}_i, \tau_{i,r} = q \mid \boldsymbol{\theta}_{r-1})}{P(\mathbf{x}_i \mid \boldsymbol{\theta}_{r-1})} \tag{4.2}$$

for $q = 1, \ldots, Q$. Note that both the numerator and the denominator in the right-hand side of Equation 4.2 are defined in Equation 4.1.

Step 2: For $q = 1, \ldots, Q$ and $j = 1, \ldots, J$, sample π_{qj} from its posterior distribution given the current values of τ_i for $i = 1, \ldots, N$, and the data and the constraints. This conditional posterior is a (truncated) Beta distribution with parameters $s_{qj,r} + 1$ and $N_{q,r} - s_{qj,r} + 1$, where $N_{q,r}$ denotes the number of persons allocated to class q in iteration r, and $s_{qj,r}$ denotes the number of persons allocated to class q in iteration r that respond 1 to item j. Note that the Beta distribution is truncated because the sampled value for π_{qj} is only acceptable if it is in accordance with the inequality constraints involving π_{qj}. The naive way to do so is: sample from the correct (non-truncated) Beta distribution until a deviate is sampled that satisfies the constraints. However, this is quite inefficient when only a small range of the distribution is admissible. Inverse probability sampling solves this problem. Let π_{qj} be the parameter that has to be sampled from the truncated Beta distribution. The lower bound a is given by the largest class specific probability that, according to the constraints imposed by the model at hand, must be smaller than π_{qj}. The upper bound b is the smallest class specific probability that, according to the constraints imposed by the model at hand, must be greater than π_{qj}. The sampling is achieved

using a uniform (0,1) deviate U and the computation of

$$\pi_{qj} = \Phi^{-1}_{\pi_{qj}}\{\Phi_{\pi_{qj}}(a) + U[\Phi_{\pi_{qj}}(b) - \Phi_{\pi_{qj}}(a)]\},$$

where $\Phi_{\pi_{qj}}(a)$ is the proportion of the conditional posterior distribution (a truncated Beta distribution) of π_{qj} below a and $\Phi_{\pi_{qj}}(b)$ is the proportion of conditional posterior distribution below b. $\Phi^{-1}_{\pi_{qj}}\{.\}$ denotes the inverse cumulative density evaluated at the argument. This procedure always renders a deviate from the conditional distribution at hand within the bounds a and b (Gelfand et al., 1992).

Step 3: Sample the class weights from their posterior distribution given the current values of τ_i for $i = 1, \ldots, N$. This posterior is a Dirichlet distribution with parameters $N_{1,r} + 1, \ldots, N_{Q,r} + 1$.

For all analyses executed in the chapter, the Gibbs sampler was run for 110,000 iterations. After a burn-in period of 10,000 iterations the values sampled in the second and third step of each 100-th iteration were saved (these iterations are denoted using the superscript $m = 1, \ldots, M$). The result is $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^m, \ldots, \boldsymbol{\theta}^{1,000}$. This sample can be used to obtain estimates of the model parameters and the corresponding credibility intervals, taking into account the prior constraints. The expected a posteriori (EAP) estimate of a parameter is simply the average of the 1,000 values of that parameter sampled from the posterior distribution. A 95% central credibility for this parameter is given by the 2.5-th and 97.5-th percentile of the distribution of these 1,000 sampled values. In the next section it is shown that it is easy to compute and evaluate fit measures using the sample from the posterior distribution.

4.4 Model Selection

After the translation of a number of competing theories into constrained latent class models, the support the data provide for each latent class model has to be determined. Three fit measures that can be evaluated using Bayesian computational methods (the marginal likelihood, posterior model probabilities, and the pseudo likelihood ratio test) have been proposed in the literature (Kass & Raftery, 1995; Hoijtink, 2001). For a discussion of the performance of these measures in the context of inequality constrained models, the interested reader is referred to Hoijtink (1998, 2001). These fit measures are discussed in the next sections.

4.4.1 Marginal Likelihood and Posterior Model Probabilities

Kass and Raftery (1995) present a comprehensive review of the marginal likelihood and posterior probability of a model. The basic idea behind the marginal likelihood factors is the same as the basic idea behind more familiar information criteria like AIC, CAIC, and BIC. It can, for example, be shown (see Kass & Raftery, 1995), that the Bayesian Information Criterion (Schwarz, 1978) is an approximation of minus twice the logarithm of the marginal likelihood. Although not explicit in its formulation, the marginal likelihood, like the information criteria, contains a trade off between the likelihood of the parameters given the data and the number of parameters in the model.

In the remainder of this chapter, minus twice the logarithm of the marginal likelihood is used

$$-2\log P(\mathbf{X}|H_k) = -2\log \int_{\boldsymbol{\theta}_k} P(\mathbf{X} \mid \boldsymbol{\theta}_k) P(\boldsymbol{\theta}_k \mid H_k)\, d\boldsymbol{\theta}_k, \tag{4.3}$$

which brings comparisons of different models on the same scale as the familiar deviance statistics (Kass & Raftery, 1995). Loosely formulated, minus twice the logarithm of the marginal likelihood can be interpreted as the distance between the model at hand and the true model: The smaller its value, the smaller the distance.

There are many ways to compute Equation 4.3. In this chapter the method proposed by Kass and Raftery (1995) is used. They suggest to sample 99% of the parameter vectors (in our case 990) from the posterior distribution parameter vectors, and to imagine that 1% of the parameter vectors (in our case 10) is sampled from an imaginary distribution where for each θ $P(\mathbf{X} \mid \theta)$ is equal to the marginal likelihood. An approximation of $-2\log P(\mathbf{X} \mid H_k)$ is denoted as $-2\log \hat{P}$ and obtained via a simple iterative algorithm based on the implicit equation

$$-2\log \hat{P} = -2\log \left[\frac{10\hat{P} + \sum_{m=1}^{990} \frac{P(\mathbf{X}|\theta_k^m)}{.01 + P\left(\mathbf{X}|\theta_k^m\right)/\hat{P}}}{\hat{P} + \sum_{m=1}^{990} \frac{1}{.01 + P\left(\mathbf{X}|\theta_k^m\right)/\hat{P}}} \right].$$

If the prior probabilities of the K models under investigation are equal, that is, $P(H_k) = 1/K$ for $k = 1, \ldots, K$, the posterior probability of each model can be computed as

$$P(H_k|\mathbf{X}) = \frac{P(\mathbf{X}|H_k)}{\sum_{k=1}^{K} P(\mathbf{X}|H_k)},$$

for $k = 1, \ldots, K$. The posterior model probability $P(H_k|\mathbf{X})$ denotes the support for model k in the total set of K models given by the data. In this chapter, both the marginal likelihood and the posterior probability of a model are reported.

4.4.2 Pseudo Likelihood Ratio Test

Hoijtink (2001) shows that the likelihood ratio test (Everitt, 1988; Lin & Dayton, 1997) is not performing very well if the goal is to select the best of a number of inequality constrained models. The performance is much better if the pseudo likelihood ratio statistic is used. This statistic is denoted by $D_k(\mathbf{X}, \boldsymbol{\theta}_k)$ and compares for each pair of items, the expected number of each possible pair of responses (i.e., 00, 10, 01, and 11, respectively) to the corresponding observed number. Let $\mathbf{n}_{gh}^{vw}$ denote for items g and h the observed frequencies of the response pattern $X_g = v, X_h = w$ where $v, w \in \{0, 1\}$. Furthermore, let $\mathbf{m}_{gh|k}^{vw}$ denote the expected frequencies of these response patterns given $\boldsymbol{\theta}_k$. The pseudo likelihood ratio statistic is then defined as

$$\mathrm{D}_k(\mathbf{X}, \boldsymbol{\theta}_k) = -2 \sum_{g=1}^{J-1} \sum_{h=g+1}^{J} \sum_{v=0}^{1} \sum_{w=0}^{1} \mathbf{n}_{gh}^{vw} \log \frac{\mathbf{m}_{gh|k}^{vw}}{\mathbf{n}_{gh}^{vw}}. \tag{4.4}$$

The expected frequencies conditional on $\boldsymbol{\theta}_k$ are computed using

$$\mathbf{m}_{gh|k}^{vw} = N \sum_{q=1}^{Q} \omega_q \left[\pi_{qg}^{v}(1 - \pi_{qg})^{1-v}\right] \left[\pi_{qh}^{w}(1 - \pi_{qh})^{1-w}\right].$$

The larger $D(\mathbf{X}, \boldsymbol{\theta}_k)$, the larger the discrepancy between the data $\mathbf{X}$ and model k.

Because $\boldsymbol{\theta}_k$ is unknown, Equation 4.4 cannot be computed. The classical solution is to substitute the unknown quantity with the maximum likelihood estimate of $\boldsymbol{\theta}_k$. The Bayesian solution uses the posterior distribution of $\boldsymbol{\theta}_k$ (Rubin, 1984; Meng, 1994; Gelman, Meng, & Stern, 1996). The posterior distribution summarizes the available information with respect to $\boldsymbol{\theta}_k$. The posterior can accurately be represented using a sample $\boldsymbol{\theta}_k^1, \ldots, \boldsymbol{\theta}_k^m, \ldots, \boldsymbol{\theta}_k^{1000}$ from this posterior. Each of the 1000 vectors $\boldsymbol{\theta}_k^m$ can be used to generate a replicated data matrix $\mathbf{X}_k^m$ that is in accordance with model k. The procedure is simple: for N persons class membership is sampled from a multinomial distribution with probabilities $\boldsymbol{\omega}_k^m$. Subsequently, the class specific probabilities $\pi_{q1,k}^m, \ldots, \pi_{qJ,k}^m$ of the class to which a person is assigned are compared with a vector of pseudo random numbers sampled from a U(0,1) distribution. If a class specific probability is larger than the

corresponding random number, a person gives the response 1, otherwise the response 0 is given. This procedure is repeated for $m = 1, \ldots, 1000$.

For each $\boldsymbol{\theta}_k$ two discrepancies can be computed: $D(\mathbf{X}, \boldsymbol{\theta}_k^m)$, which is a discrepancy between the observed data and the model; and, $D(\mathbf{X}_k^m, \boldsymbol{\theta}_k^m)$, which is a discrepancy between replicated data and the model. If $D(\mathbf{X}_k^m, \boldsymbol{\theta}_k^m) \geq D(\mathbf{X}, \boldsymbol{\theta}_k^m)$, the discrepancy between the observed data and the model is equal to or smaller than the discrepancy between the replicated data and the model. The posterior predictive p-value is the proportion of 1000 comparisons for which $D(\mathbf{X}_k^m, \boldsymbol{\theta}^m) \geq D(\mathbf{X}, \boldsymbol{\theta}^m)$. The posterior predictive p-value is formally defined as

$$p_k = Pr[D(\mathbf{X}_k, \boldsymbol{\theta}_k) \geq D(\mathbf{X}, \boldsymbol{\theta}_k) \mid \mathbf{X}, H_k],$$

that is, the probability that the discrepancy between model k and a data-matrix $\mathbf{X}_k$ generated using model k is equal to or larger than the discrepancy between model k and the observed data matrix $\mathbf{X}$. The p_k is an absolute fit measure, that can be used to test the pseudo likelihood ratio statistic, that is, which can be used to determine whether model k accurately describes the data. Stated otherwise, analogous to the interpretation of classical p-values, values smaller than .05 indicate a lack of fit of the model, and values larger than .05 indicate that the model at hand was able to accurately reproduce the observed frequencies.

4.5 Strategies to Solve the Piagetian Balance Scale Task

The balance scale task was recognized in the early eighties as a way of eliciting different rule-governed response patterns for proportionality reasoning (Siegler, 1981). A picture of a simplified balance scale is shown to children. While the beam is fixed, a number of identical weights are placed on each side at certain distances from the fulcrum. For each of a number of balances (the items) the children have to predict which side will tip, if any. The weights on the balance differ with respect to their number and distance to the center. The formal rule to obtain the correct answer is that the balance is in equilibrium when the product of the number of weights and the distance from the center is equal at both sides of the balance.

Applications of ELCA in the context of the balance task can be found in Boom et al. (2001) and Jansen and Van der Maas (1997). New balance scale data will be used to determine which of the existing theories that explain children's responses to the items of the balance scale task is the best. As the result is not conclusive, the best of these theories will be used as the point of departure for further theoretical developments.

Nearly 900 randomly selected Dutch children from 4- to 16-years-old participated, with a mean age of 10.35 (standard deviation 2.82). The children were tested individually at home by students and did not receive feedback until the task had been finished. The assessment was part of a training procedure for psychology students. The students were prepared for this specific assessment in small groups and had to follow a strict assessment protocol.

4.5.1 Theories and Hypotheses About the Data

Siegler (1981) distinguished six types of problems. In *balance* problems, weight and distance are equal on both sides. In *weight* problems, the distance is the same on both sides but the number of weights is different. These first two problem types were not used in this study, since they do not differentiate between the postulated rules and were expected to be answered correctly by all children. In *distance* problems, the weight is the same on both sides but the distance is different. In conflict problems, more weight is on one side and greater distance on the other, such that the side with more weight falls (*conflict-weight* problem), the side with the greater distance falls (*conflict-distance* problem), or the balance remains horizontal (*conflict-balance* problem).

Siegler (1981) described four strategies or rules. Each of these strategies can be characterized by a specific pattern of scores on the different item types.

rule 1 Children will only consider the number of weights on each arm. Therefore it can be expected that they have a higher probability of correctly responding to the weight and the conflict-weight items than to the other item types.

rule 2 Children get a grasp of distance: When the number of weights is equal on both sides, they judge the influence of distance correctly, otherwise they ignore distance and only consider the number of weights. For this strategy, it can be predicted that children have a higher probability of correctly responding to the weight, distance, and conflict-weight items than the other item types.

rule 3 Children will evaluate both the distance and the number of weights correctly, but if one side has more weights and the other side more distance they will be confused and guess. The probability of success will be at chance level (they make a random prediction) for all conflict type of problems.

Table 4.4: Inequality Constraints for Siegler's Model (for Explanation of the Notation, See Text).

		Restrictions			
Item type	Items	Rule 1	Rule 2	Rule 3	Torque
Distance	1-5	-	+	+	+
Conflict Weight	6-9	+	+	±	+
Conflict Distance	10-14	-	-	±	+
Conflict Balance	15-19	-	-	±	+

rule 4 Children will apply the correct (torque) rule. The probability of responding correctly is high for all item types.

As can be seen in Table 4.4, the test used in this chapter contains 19 items of the following types: five distance, four conflict-weight (originally 5 but one item had a printing error in the test booklet for half of the sample), five conflict-distance, and five conflict-balance. In Table 4.4 a translation of Siegler's model into CLCA is elaborated upon. Note that the inequality constraints are implicitly shown: - indicates a low probability of correctly responding to the item, + a high probability of correctly responding to an item, and ± indicates a random prediction. All the probabilities associated with the - signs have to be smaller than the probabilities associated with the ± signs, and all the probabilities associated with the + signs have to be larger than probabilities associated with ±.

Wilkenings and Anderson (1982) argue the existence of another strategy. The *addition rule* suggests that the number of weights and the number of distance intervals on the left are summed and compared to the sum of weights and distances on the right: the side with the greater sum is expected to tip. For the existing item types, we designed the items such that the addition rule could be detected because two conflict-weight items and two conflict-balance items evoke an incorrect response whereas the remaining conflict-weight and conflict-balance and all conflict-distance items evoke a correct response when this rule is applied to this set of items. Children applying the addition rule will have a low probability of correctly responding to the items that evoke an incorrect response, and a high probability of correctly responding to the remaining conflict-items. Normandeau, Larivee, Roulin, and Longeot (1989) argue that rule 3 of Siegler is not homogeneous. Their paper supports the existence of the addition rule and they introduce yet another rule: *qualitative proportion rule*. Children using this rule understand that more weights at a small distance from the fulcrum compensates

Table 4.5: Inequality Constraints for Normandeau's Model (for Explanation of the Notation, See Text).

		Restrictions				
Item Type	Items	Rule 1	Rule 2	Add	QP	T
Distance	1-5	-	+	+	+	+
Conflict Weight	6,9	+	+	-	-	+
Conflict Weight Add	7,8	+	+	+	-	+
Conflict Distance	–	-	-	-	-	+
Conflict Distance Add	10-14	-	-	+	-	+
Conflict Balance	16,19	-	-	-	+	+
Conflict Balance Add	15,17,18	-	-	+	+	+

Note. T = Torque.

for fewer weights at a far distance, resulting in a prediction of balance for all conflict problems. Thus, the qualitative proportion rule predicts that all conflict-weight and conflict-distance items have low probabilities of being answered correctly, and all the conflict-balance items have a high (or higher) probability of a correct response. The five resulting latent classes are displayed in Table 4.5. Note that this table is comparable to Table 4.4, but extended to differentiate between the addition and non-addition items. Moreover, there can be seen that rule 3 has been split up into a latent class accounting for the addition rule and a latent class accounting for the qualitative proportion rule. In the current item set all conflict distance items were solvable using the addition rule.

4.5.2 Results

The model selection measures have been computed for Siegler's model and for Normandeau's model. Note that the pseudo likelihood p-value indicates the absolute fit of the model. A p-value smaller than 0.05 indicates a lack of fit of the model, whereas the p-value is larger than 0.05, the model accurately reproduces the observed frequencies. The marginal likelihood can be interpreted as the distance between the model at hand and the true model: the smaller the value, the smaller the distance. The value of the marginal likelihood is on the same scale as the familiar deviance statistics. Two or more models can be compared using the value of the marginal likelihood. Since the marginal likelihood implicitly uses a parameter penalty, the model with the smallest marginal likelihood value has to be preferred. The marginal likelihoods of all models analyzed can also be transformed

Table 4.6: Fit Evaluation of the Siegler's Model and Normandeau's Model for the Balance Scale Data.

Model	Pseudo Likelihood	-2log Marginal Likelihood	Posterior Model Probability
Siegler	0.003	13421	0
Normandeau	0.019	13206	1.0

into the posterior model probability. This number indicates the support for each model in the total set of models given the data.

In Table 4.6, it can be seen from both the marginal likelihood and the posterior model probability that Normandeau's model is superior. However, as indicated by the p-value of the pseudo likelihood ratio test (smaller than .05), it is questionable whether Normandeau's model adequately reproduces the frequencies with which the response vectors are observed. This lack of fit could be due to existing strategies that are not predicted by the theory.

Figure 4.1 presents the class specific probabilities of Normandeau's model. On the horizontal axis the items are displayed, on the vertical axis the class specific probabilities. Classes one and two clearly represent rule 1 (only considering weight leads to a high probability of answering conflict-weight items (6-9) correct) and rule 2 (high probability of answering conflict-weight (6-9) and distance items (1-5) correct). These rule are dominant, since a substantial part of the sample belongs to these classes. The third class represents the addition rule, although the probabilities for items 15, 17 and 18 are lower than expected for this rule.

Class four represents the qualitative proportion rule. As can be seen only a small proportion of the children belong to this class. Furthermore, although the class specific probabilities are in accordance with the constraints, especially the probabilities for the first and the last five items should have been higher to obtain a convincing representation of a qualitative proportion rule. It can be a 'true' strategy, but maybe it should be specified in more detail than simply by "all conflict items except conflict balance items are answered incorrectly." It could be, for example, that children do have an intuitive idea how distance and weight work together, but only when there is a large difference between the products of weight and distance on both sides.

Knowing that there are very few children that can actually solve the balance scale task, a class size of 29% for rule 4 seems extremely large.

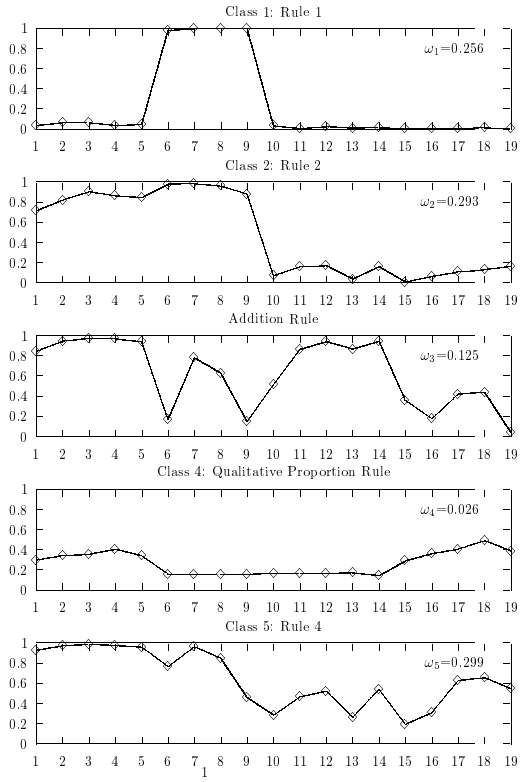

Figure 4.1: Class-Specific Probabilities of the Normandeau Theory.

Furthermore, the class specific probabilities for this correct strategy are all predicted to be high, but as can be seen in the figure, this is not the case. Stated otherwise, class five does not yet give a convincing representation of rule 4.

The results for Normandeau's model are not conclusive. The p-value of the pseudo likelihood ratio test indicates that the data are not adequately reproduced. Furthermore, for classes four and five the class specific probabilities do not give a clear representation of the presumed underlying rule.

It could have been an option to represent the torque rule by a class for which $\pi_{qj} > .90$ for $j = 1, \ldots, J$. However, this value of 0.90 seems rather arbitrary. From the theory it can be predicted that the probability of a correct response in class five has to be higher than the probabilities associated with a random prediction. Note that the method of testing models via the incorporation of inequality constraints on the model parameters explicitly shows that one chooses the best theory from a set of reasonable theories. This means that not all possible models are included, nor guarantees this procedure that the 'true' model is in the set. In the next section, it is shown how this best theory can be used as the point of departure for theory refinement.

4.5.3 Theory Refinement

In this section, Normandeau's model is extended with one and two unconstrained classes, respectively. This constitutes an example of scientific exploration using the current state of knowledge as the point of departure. Note that the results of this exploration are indefinite. To confirm the exploratory results, these findings have to be translated into inequality constrains and they have to be analyzed using new data.

As can be seen in Table 4.7, the Normandeau model with two unconstrained classes receives the most support from the data. Note, that the p-value of the pseudo likelihood ratio test now indicates that the data are adequately reproduced by this model. Furthermore, comparing the posterior probabilities of the three Normandeau models, it is clear that the variant with two extra unconstrained classes is superior.

As can be seen in Figure 4.2, the interpretation of the first four classes is similar to the interpretation given for the Normandeau model without extra classes. Note, however, that the probabilities for items 17 and 18 in class three have increased, that is, class three gives a better representation of the addition rule. The same holds for the first and last five items in class four, which now give a better representation of the Qualitative Proportion rule. Class five now represents rule 4 and contains only a small proportion of the children (as is expected).

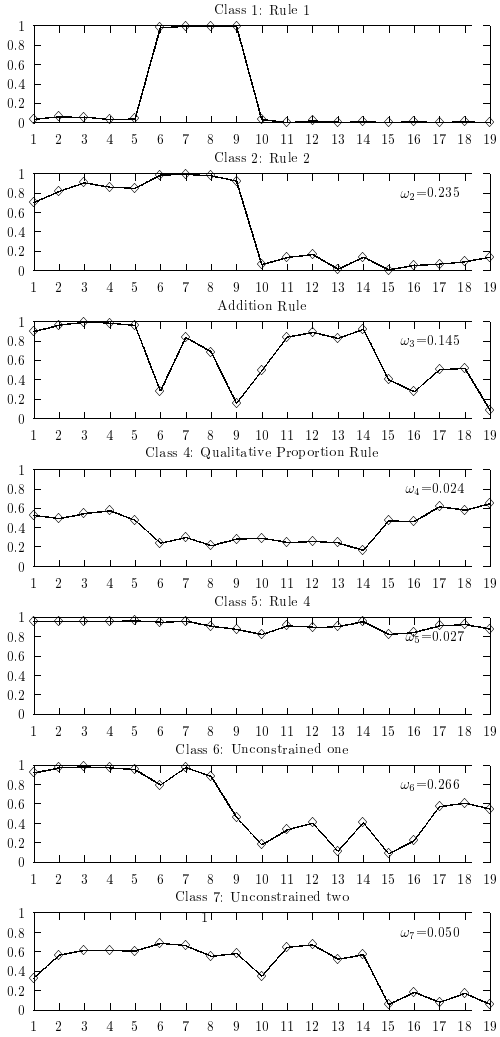

Figure 4.2: Class-Specific Probabilities of the Normandeau Theory Extended With Two Classes.

Table 4.7: Refined Fit Evaluation of Siegler's Model, Normandeau's Model, and Two Extended Models for the Balance Scale Data.

Model	Pseudo Likelihood	-2log Marginal Likelihood	Post
Siegler	0.003	13421	-
Normandeau	0.019	13206	0.0
Normandeau + 1 class	0.054	12909	0.0
Normandeau + 2 classes	0.082	12837	1.0

Class 6 accounts for a fairly large proportion of persons. This class is difficult to associate with a strategy or rule. Our best guess is that it is class of children who are somewhere between rule 2 and the addition rule. The second unrestricted class (class 7) is a global pattern, only grouping children that have in common that they do not consider the answer 'balance' an option (the last five items are almost never correctly answered). This class was also mentioned by Jansen and van der Maas (1997).

In the current version of the balance scale task, the items are chosen on the basis of being of a certain type (e.g., conflict balance item). The magnitude of the physical quantities is not varied systematically. We suggest that in further research one chooses the items of the same type more carefully, such that the role of the physical quantities can be asserted. For example, choose addition items within the conflict distance item such that the items vary from a big difference between addition torque to a small difference in a controlled way.

4.6 Conclusion

This chapter illustrated that theories can be included in latent class models using inequality constraints among the class specific probabilities. An example from the domain of developmental psychology was used to illustrate the resulting CLCA. If, in a certain research domain, one or more theories exist, CLCA has advantages over ELCA. First of all it provides a straightforward way to select the best of a number of competing theories. Secondly, as illustrated using the balance scale data, it allows theory refinement using the current state of knowledge as the point of departure.

Siegler and Chen (2002) mention some disadvantages of the LCA. One is the arbitrary alpha level of 5% and the unclear interpretation of it. This

is acknowledged in Bayesian statistics for quite some time, and a solution has been found in the form of posterior model probability (Sellke, Bayarri, & Berger, 2001). We use this solution, because instead of a probability of incorrectly rejecting the null hypothesis, the posterior model probability gives the probability of the data given the model among other models.

Currently user friendly software containing an implementation of the proposed approach is being developed. Readers interested in this software can send an e-mail to the first author at *o.laudy@fss.uu.nl*. The e-mail should include a description of the research at hand, the data, and the theories involved.

References

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance: Classes, strategies or rules for the balance scale task. *Cognitive Development, 16*, 717-735.

Clogg, C. C. (1981). New developments in latent structure analysis. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 215-246). Beverly Hills, CA: Sage.

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

Everitt, B. S. (1988). A Monte Carlo investigation of the likelihood ratio test for number of classes in latent class analysis. *Multivariate Behavioral Research, 23,* 531-538.

Gelfand, A. E., Smith, A. F. M., & Lee, T. (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. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica, 6*, 733-807.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology, 79*, 1179-1295.

Haberman, S. J. (1988). *Analysis of quantitative data, vol 2. New developments.* New York: Academic Press.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171-180.

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

Hoijtink, H. (2001). Confirmatory latent class analysis: model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research, 16*, 717-735.

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical test of rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321-357.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non nested latent class models. *Journal of Educational and Behavioral Statistics, 22*, 249-264.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

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

Normandeau, S., Larivee, S., Roulin, J., & Longeot, F. (1989). The balance-scale dilemma: Either the subject or the experimenter muddles through. *Journal of Genetic Psychology 150*, 237-250.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*, 1151-1172.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician 55*, 62-71.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development. 46*(2, Serial No. 189).

Siegler, R. S., & Chen, Z. (2002). Development of rules and strategies: Balancing the old and the new. *Journal of Experimental Child Psychology, 81*, 446-457.

Vermunt, J. K. (1997). *Log-linear models for event histories.* Thousand Oakes, CA: Sage.

Wilkenings, F., & Anderson, N. H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin, 92*, 215-237.