Multilevel Mixture Models

Tihomir Asparouhov
Muthen & Muthen

Bengt Muthen
UCLA

Version 3
November 27, 2006

[1] This research was supported by SBIR grant R43 AA014564-01 from NIAAA to Muthen & Muthen. The research of the second author was also supported under grants R01 DA11796 from NIDA and K02 AA 00230 from NIAAA.

The authors thank E. Erosheva for helpful comments on the earlier draft.

1 Introduction

Multilevel statistical models allow researchers to evaluate the effects of individuals' shared environment on an individual's outcome of interest. Finite mixture models allow the researchers to question the homogeneity of the population and to classify individuals into smaller more homogeneous latent subpopulations. Structural equation models allow the researchers to explore relationships between observed variables and latent constructs. As researchers get more and more experience with these techniques they will inevitably want to use them within a unified framework that will enable them to combine all these ideas into a comprehensive statistical model that addresses all features present in the data. In this article we will describe a general statistical model that incorporates multilevel models, finite mixture models and structural equation models into a very general and flexible modeling framework. The basis of this methodology was first implemented in Mplus Version 3 (Muthen & Muthen, 2004), while the complete modeling framework described in this article is available in Mplus 4.2 (Muthen & Muthen, 2006).

The topic of Multilevel Mixture Models is relatively new, although a number of articles have discussed similar frameworks and applications. Among these are Asparouhov (2006), Bijmolt et. al. (2004), Vermunt (2003) and Vermunt and Magidson (2005).

The goal of this article is to describe a two-level modeling framework that includes multiple latent variables. Each latent class variable can be either a within level variable, a between level variable or a within-between level variable. The more general and flexible a statistical model is, the bigger the effort on the part of the researcher to interpret the model and the results in a practically meaningful way. In this article we will illustrate the general model with some specific simple examples and will describe the advantages of these models over conventional modeling techniques.

In Section 2 we describe the basic two-level mixture framework. In Section 3 we introduce the multiple class variables framework and describe a two-level latent transition analysis (LTA) model. In Section 4 we use the two-level LTA model to analyze students behavior data and illustrate the modeling capabilities of this framework. In Section 5 we describe a two-level model with a between level latent class variable and compare this model to the model described in Section 2. We show that a between level latent class variable is a special case of the within-between latent class variable used in

Section 2 and 3. In Section 6 we describe a two-level model that incorporates both a within and a between latent class variables. In Section 7 we describe the grade of membership model (GoM) and show its advantages over latent class analysis (LCA) models. We also show how this model can be incorporated into the modeling framework described in Section 2. We illustrate 2- and 3- class GoM model with an application of criminal offense data. The GoM modeling idea can be combined with most finite mixture models. In Section 8 we apply the GoM modeling idea to the factor mixture model (FMA). Thus we incorporate the GoM, the IRT and the LCA model into one general model. We illustrate the capabilities of this FMA-GoM model with a practical application using ADHD diagnostics data. In Section 9 we show that the 3PL guessing IRT model is a special case of the FMA-GoM model. Section 10 discusses the technical aspects of the estimation of the two-level mixture models.

2 The Basic Two-level Mixture Model

Let y_{pij} be the $p-$th observed dependent variable for individual i in cluster j. In this section we will only consider two types of variables, categorical and normally distributed continuous variables. However it is possible to incorporate other types of distributions and link function as in the generalized linear models of McCullagh and Nelder (1989). Suppose that C_{ij} is a latent categorical variable for individual i in cluster j which takes values $1, ..., L$.

To construct a structural model for the categorical variables we proceed as in Muthen (1984) by defining an underlying normally distributed latent variable y^*_{pij} such that for a set of parameters τ_{ck}

$$[y_{pij} = k|C_{ij} = c] \Leftrightarrow \tau_{ck} < y^*_{pij} < \tau_{ck+1}. \tag{1}$$

A linear regression for y^*_{pij} is thus equivalent to a Probit regression for y_{pij}. Alternatively, y^*_{pij} can have a logistic distribution. Linear regression for y^*_{pij} will then translate to a logistic regression for y_{pij}. For continuous variables we define $y^*_{pij} = y_{pij}$.

Let y^*_{ij} be the vector of all dependent variables and let x_{ij} be the vector of all covariates. The structural part of the model is defined by

$$[y^*_{ij}|C_{ij} = c] = \nu_{cj} + \Lambda_{cj}\eta_{ij} + \varepsilon_{ij} \tag{2}$$

$$[\eta_{ij}|C_{ij} = c] = \mu_{cj} + B_{cj}\eta_{ij} + \Gamma_{cj}x_{ij} + \xi_{ij} \tag{3}$$

$$P(C_{ij} = c) = \frac{\exp(\alpha_{cj} + \beta_{cj} x_{ij})}{\sum_c \exp(\alpha_{cj} + \beta_{cj} x_{ij})}. \tag{4}$$

where η_{ij} are normally distributed latent variables, ε_{ij} and ξ_{ij} are zero mean normally distributed residuals. Some parameters have to be restricted for identification purpose. For example, the variance of ε_{pij} should be 1 for categorical variables y_{pij}. Also $\alpha_{Lj} = \beta_{Lj} = 0$.

The multilevel part of the model is introduced as follows. Each of the intercept, slope or loading parameters in equations (2-4) can be either a fixed coefficient or a cluster random effect, i.e., a coefficient that varies across clusters. Let η_j be the vector of all such random effects and let x_j be the vector of all cluster level covariates. The between level model is then described by the following equation

$$\eta_j = \mu + B\eta_j + \Gamma x_j + \xi_j \tag{5}$$

where ξ_j is a normally distributed residual.

The above four equations comprise the definition of a simple multilevel structural mixture model. There are many extensions of this model that are available in the Mplus framework. For example observed dependent variables can be incorporated on the between level. Other extensions arise from the fact that a regression equation can be constructed between any two variables in the model. Such equations can be fixed or random effect regressions. Another interesting extension is to have all intercept and slopes parameters in equation (5) vary across the latent class. This essentially amounts to interaction between the C variable and the random effect variables.

The model described in this section can also be extended to include multiple latent class variables. This extension is described in the following section.

3 Multiple Latent Class Variables

In this section we describe the basic framework for a multilevel mixture model with multiple latent categorical variables $C_1, C_2, ...$ etc. For simplicity we will focus on the model with two latent categorical variables C_1 and C_2 however the framework easily extends to more than two class variables. One application of the multiple latent class variable framework is the latent transition analysis (LTA) model. The LTA model is used in longitudinal settings and C_t represents the latent class variable at time t. As in the previous section let y^*_{tij} be the vector of all dependent variables observed at time t and x_{tij} be the vector of all covariates at time t. The structural part of the model is given by

$$[y^*_{tij}|C_{tij} = c] = \nu_{tcj} + \Lambda_{tcj}\eta_{tij} + \varepsilon_{tij} \tag{6}$$

$$[\eta_{tij}|C_{tij} = c] = \mu_{tcj} + B_{tcj}\eta_{tij} + \Gamma_{tcj}x_{tij} + \xi_{tij} \tag{7}$$

where η_{tij} are normally distributed latent variables and ε_{tij} and ξ_{tij} are normal residuals. The multinomial logistic regression for the class variable C_1 at the first time point is given by

$$P(C_{1ij} = c) = \frac{\exp(\alpha_{1cj} + \beta_{1cj}x_{1ij})}{\sum_c \exp(\alpha_{1cj} + \beta_{1cj}x_{1ij})}. \tag{8}$$

The multinomial logistic regression for the second class variable C_2 includes C_1 as a covariate

$$P(C_{2ij} = d|C_{1ij} = c) = \frac{\exp(\alpha_{2dj} + \gamma_{dcj} + \beta_{2dj}x_{2ij})}{\sum_d \exp(\alpha_{2dj} + \gamma_{dcj} + \beta_{2dj}x_{2ij})}. \tag{9}$$

where γ_{dcj} shows the effect of C_1 on C_2. Equations (8) and (9) form a set of recursive system of logit models, see Agresti (1996), and can be used to explore the dependence of C_2 on C_1. When there are more than two latent categorical variables C_1,...,C_T the LTA models the dependence of C_t on the previous class variables C_1,...,C_{t-1}. A first order Markov chain model is a special case of the LTA model which assumes that C_t depends only on C_{t-1} but not on earlier class variables.

As in the previous section, each of the intercept, slope and loading parameters in equations (6-9) can be either a fixed coefficient or a random effect. If η_j are all random effects and x_j are all cluster level covariates, equation (5) again describes the cluster level structural model.

The multilevel framework described above allows us to study the effect of C_1 on C_2 on the individual level through equation (9) but also on the cluster level by estimating the intercepts α_{2dj} and α_{1cj} as random effects and estimating a regression equation

$$\alpha_{2dj} = \mu + \beta\alpha_{1cj} + \varepsilon_{dcj}. \tag{10}$$

We illustrate this modeling technique with a practical example in the next section.

4 Two-level LTA Example

To illustrate the two-level LTA model we use data from the Baltimore study of aggressive and disruptive behavior in the classroom, see Muthen et. al. (2002). The data to be analyzed consists of 10 Likert scale items known as the TOCA instrument. These items are teacher-rated student's behavior on the scale of 1 to 6. The items are strongly skewed and to simplify the illustration we convert all items to binary scale. All values larger than 1 are recoded as 2. The statistical model that we present here is not intended to draw any substantive conclusions. We only illustrate the statistical methodology that could be useful in such applications. The model that we described however was suggested by Nicholas Ialongo as an appropriate approach in these settings. We analyze the first grade data collected in the fall and in the spring. The 10 fall measurements, U_{1p}, $p = 1, ..., 10$, are used to estimate a latent class model with two classes. Denote this class variable with C_1. Similarly the 10 spring measurements, U_{2p}, $p = 1, ..., 10$, are used to estimate a latent class model with two classes. Denote this class variable with C_2. Consequently we combine the two models into a single model and estimate a transitional model from C_1 to C_2. Of particular interest is the effect of C_1 on C_2 which can be estimated in a logistic regression where C_2 is the dependent variable and C_1 is the predictor variable as in equation (9). The structural part in the fall data is fairly similar to the structural part in the spring data so we estimate the joint model with a time invariant latent class model. In both the fall and the spring the first class contains the more disruptive students and the second class contains the less disruptive students. The model is described by the following equations

$$P(U_{tp} = 2|C_1 = c) = \pi_{cp} \tag{11}$$

$$P(C_1 = 1) = \frac{exp(\alpha_1)}{exp(\alpha_1) + 1} \tag{12}$$

$$P(C_2 = 1|C_1) = \frac{exp(\alpha_2 + \gamma I(C_1))}{exp(\alpha_2 + \gamma I(C_1)) + 1} \tag{13}$$

where $I(C_1)$ is an indicator variable for C_1, $I(C_1) = 1$ if $C_1 = 1$ and $I(C_1) = 0$ if $C_1 = 2$. Large values of γ will indicate strong relationship between C_1 and C_2. It is also possible to calculate the R^2 contribution of C_1 as a predictor of C_2

$$R^2 = \frac{\gamma^2 P(C_1 = 1)(1 - P(C_1 = 1))}{\gamma^2 P(C_1 = 1)(1 - P(C_1 = 1)) + \pi^2/3}. \tag{14}$$

Table 1: Probability Profiles for the Two Classes.

class	$C_t = 1$	$C_t = 2$
parameter	π_{1p}	π_{2p}
Stubborn	0.92	0.36
Break Rules	0.96	0.29
Harm Others	0.73	0.03
Break Things	0.59	0.03
Yells at Others	0.82	0.18
Take Others' Property	0.78	0.07
Fights	0.73	0.08
Lies	0.81	0.10
Tease Classmates	0.90	0.24
Trouble Accepting Authority	0.78	0.12

The term $\pi^2/3$ represents the variance of the error term with the logistic distribution.

The model described so far however will not allow us to evaluate the classroom effects on individual behavior. Previous analysis on the Baltimore data have shown that the average level of aggressive/disruptive behavior in the classroom strongly influences individual aggressive behavior development. To incorporate the classroom effects we estimate α_1 and α_2 as normally distributed classroom level random effects. These random effects will allow us to model the differences between the classrooms. For example, large α_1 values correspond to classrooms with a large number of disruptive students. In addition we can estimate a regression equation between the two random effects

$$\alpha_{2j} = \mu + \beta\alpha_{1j} + \varepsilon_j \tag{15}$$

where μ and β are fixed coefficients and ε_j is a mean zero normally distributed residual. Thus the model will allow us to evaluate not only the individual effect of C_1 on C_2 in equation (13) but also the direct effects of α_1 on C_1 in equation (12), α_2 on C_2 in equation (13) and α_1 on α_2 in equation (15). In addition the total effect of α_1 on C_2 can be computed which consists of two indirect effects, via C_1 and via α_2.

The results of this two-level LTA are presented in Table 1 and Figure 1.

Figure 1: Two-level Latent Transition Model

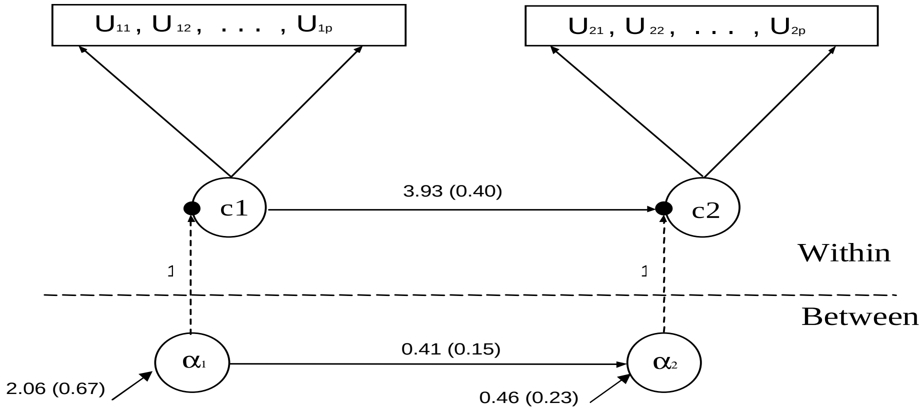


Table 1 contains the probability profiles for the two classes and all ten TOCA instruments. It is clear that the two classes are very well separated. For the fall data class 1 contains 46% of the students while for the spring data class 1 contains 52% of the students. The probability of switching from class 1 to class 2 is only 7% while the probability of switching from class 2 to class 1 is 18%.

Figure 1 shows the structure of the model and the estimated regression paths between C_i and α_i. As in equation (14) we can compute the proportion of explained variance by this model. For example the logistic regression equation from α_1 to C_1 shows that α_1 explains 39% of the variance of C_1. For C_2 this model explains 65% of the variance, 35% is explained by the classroom effect α_1, 30% is explained by the residual individual effect of C_1 (the part of C_1 that is unexplained by α_1), and 5% is explained by the residual classroom effect α_2 (the part of α_2 that is unexplained by α_1). Also we can see that alone C_1 explains 41% of the variance of C_2, while the addition of the classroom effect α_2 explains now only 24% of the variance, which is a significant reduction from the fall classroom influence of 39%. This seems to indicate that much of the classroom influence has occurred in the fall.

5 Between Level Class Variables

Models where the latent class variable is not an individual level variable but is a cluster level variable are also of interest. Such models will allow us to explore population heterogeneity that is caused by cluster level variables. For example, when heterogeneity in students' performance is caused by heterogeneity among teachers the latent class variable in the model should be a cluster level variable. Small modification in the model described in Section 2 are needed to accommodate between level class variables. The first modification is that

$$C_{ij} = C_j \tag{16}$$

which essentially is a stochastic type equality constraint that guarantees equality between the class variables within a cluster of observations. The second modification is that equation (4) should be replaced by

$$P(C_j = c) = \frac{\exp(\alpha_c + \beta_c x_j)}{\sum_c \exp(\alpha_c + \beta_c x_j)}, \tag{17}$$

because only between level covariates can be used as class predictors. Note also that the intercepts and slopes in equation (17) are not random as we now have only one such equation per cluster which makes it very hard to model these parameters as random effects. A between level class variable allows a more flexible structural model on the between level. In the between level structural model all parameters, including the residual variance covariance matrix can be class specific

$$\eta_j = \mu_c + B_c \eta_j + \Gamma_c x_j + \xi_j. \tag{18}$$

In multilevel mixture models estimating a between level class variable is actually easier than estimating a two level model with a within level class variable. For example the forward-backward algorithm (see, Vermunt, 2003) is not needed when the class variable is on the between level and one can use a simple EM estimation approach as in Muthen and Shedden (1999). However, it is not clear in general if between level heterogeneity is feasible to estimate in many practical applications with relatively small sample size on the between level. Between level sample size of 100 clusters or less is a rather common situation in multilevel data sets. The key question in modeling between level class variables is whether the within level observed variables can be used directly to identify the classes. If that is not the case, the within level

Table 2: Model specification and MSE of α_1

Model	C	v	MSE
Model 1	within	0	0.31
Model 2	between	0	0.10
Model 3	between	0.1	0.26
Model 4	between	0.2	0.44

observed data would be used simply to measure the between level random effects which will consequently identify the classes. In that case we will have a rather limited sample size to identify the classes, namely the between level sample size. An alternative way to pose this question is how to construct models with between level latent class variables where the within level data can contribute directly to the class formation and identification, which will produce more reliable models with more accurate parameter estimates.

The answers to the above questions will be illustrated with the following simulation study. We generate and estimate four different two class mixture models and evaluate the stability of the estimation by the mean squared error (MSE) of parameters α_1 in equation (4). The smaller the MSE of α_1 the easier it is to recover the heterogeneity in the population. We use a simple two class mixture model for a two-level random effect regression

$$Y_{ij} = \mu_{cj} + \beta_{cj} X_{ij} + \varepsilon_{ij} \tag{19}$$

where μ_{cj} and β_{cj} are between level random effects with variance v and covariance 0. We vary the parameter v across the models. The means of μ_{cj} and β_{cj} are 1 and 0.2 in class 1 and 0 and 0.8 in class 2. The residual variable ε_{ij} is a zero mean normally distributed variable with variance $\theta = 1$. The covariate X_{ij} is also a standard normal random variable. We vary the status of the C variable across the models. In Model 1 the variable is a within variable and in Model 2-4 it is a between level variable. Table 2 summarizes the specification of the models and the MSE of the log odds ratio parameter α_1 in equation (17). All models are generated and estimated with the correct specification. The sample size is 500 for all models, there are 50 clusters of size 10 each. The two classes are of equal size. We generate 100 samples for each model.

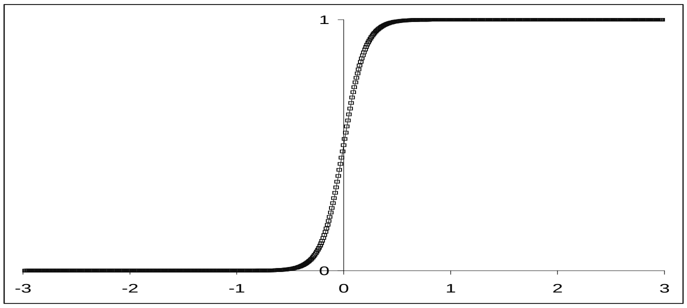
Table 3: Parameter Estimates (Standard Errors) and Log-Likelihood Value

Parameter	True Value	Model 5	Model 6
μ_1	1	0.90(0.14)	0.90(0.13)
μ_2	0	0.13(0.13)	0.12(0.12)
β_1	0.2	0.25(0.07)	0.25(0.06)
β_2	0.8	0.81(0.07)	0.81(0.06)
θ	1	1.01(0.04)	1.01(0.04)
v	0.2	0.22(0.08)	0.22(0.08)
LL		-1517.6	-1517.6

The results of this simulation study show clearly that there are two competing forces moving in opposite direction when it comes to identifying between level class variables. The difference between Model 1 and Model 2 is only in the stochastic constraint (16). The fact that the class variables are constrained to be identical across cluster contributes greatly to easing the class identification process. The difference between Models 2-4 is in the variance of the between level random effects. As the variance increases, the advantage of the stochastic constraint (16) is lost due to the fact that the random effect means in the two classes are fewer and fewer standard deviation units apart, which makes the two classes overlap. The conclusion of this simulation study is that to produce reliable models with between level class variables the models should include sufficient number of parameters that differ across class and utilize the within level observed data directly. Such parameters are class varying fixed effects or random effects with large variation across class but small variation across clusters. For example a sound strategy would be to include significant between level random effects but to eliminate from the model between level random effects that are not significant and replace them with fixed effects that can vary across classes.

The next modeling issue we want to address is the situation when the researcher does not a priori know whether or not the latent class variable is a cluster level variable or individual level variable. This is a substantively important question as it provides an insight into the causation of the heterogeneity. We illustrate this issue with the following simulation study. We generate a sample according to a modification of Model 4 where the random

Figure 2: Conditional Latent Class Probability, Var(α_1)=100



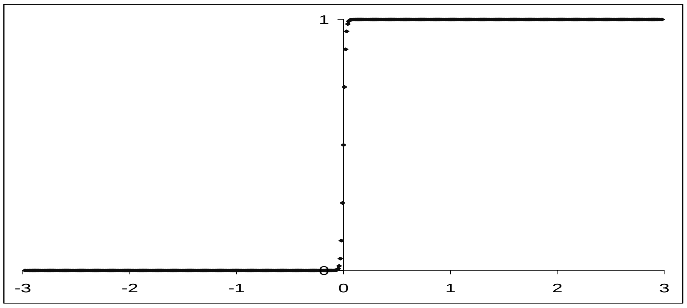
slope variance is fixed to 0 while the random intercept variance is 0.2. We generate a sample of 100 clusters of size 10 for a total of 1000 observations. Let's call this model Model 5. We analyze the data according to this true model assuming that the class variable is a between level variable. We also analyze this data according to Model 6 where the latent class variable C is assumed to be a within level variable with a between level random intercept α_1. Model 6 would be the model of choice when the researcher is uncertain about the status of the latent class variable. In Model 6 we assume that the class variable is a within level variable, but with a between level component, so essentially it is a variable that has both individual and cluster effect, i.e., within and between variable. The results of the analysis are presented in Table 3. The two estimated models have almost identical parameter estimates. For all parameters the true value is within the confidence limits. The variance of α_1 in Model 6 was estimated as 609 which can be used to compute the ICC value for C

$$ICC = \frac{Var(\alpha_1)}{Var(\alpha_1) + \pi^2/3} = 0.995. \tag{20}$$

This simulation shows that even if the latent class variable is a between level variable the data can be analyzed as if the variable is a between-within variable as in Model 6.

Formal tests can be constructed to test whether the class variable is only a between level variable or it is a between-within level variable. Using the

Figure 3: Conditional Latent Class Probability, Var(α_1)=10000



ICC value one can test whether this value is significantly different from 1 by constructing a confidence interval using the delta method or by using the bootstrap resampling method. Note however that because the ICC parameter is approaching its boundary value of 1, the delta method is not as reliable and should only be used as an approximation. This is because maximum likelihood estimation of bounded parameters does not follow the usual asymptotic theory of unbounded parameters. For practical purposes however an ICC value above 0.8 would be a good indication to pursue models where the class variable is only on the between level. In addition, an LRT test can be used to test for significant differences between the two models. That is because the two models are actually nested and Model 6 has one more parameter than Model 5. This additional parameter is the variance of α_1. Again however because of the boundary proximity the LRT test will not have the usual chi-square distribution when the ICC approaches 1.

Finally let's focus on the question why Model 5 is nested within Model 6. When the variance of α_1 in Model 6 approaches infinity the model becomes equivalent to Model 5 because the variables C_{ij} are so highly correlated that

$$P(C_{i_1j} \neq C_{i_2j}) \approx 0, \tag{21}$$

i.e. the stochastic constraint (16) is mandated by the model. Figure 2 and Figure 3 also illustrate this point. In Figure 2 we plot the conditional class probability $P(C_{ij} = 1|\alpha_1)$ on a standard scale of α_1, when the variance of α_1 is 100 and the mean is 0. Figure 3 shows this probability when the

Table 4: Specification for Latent Class Variable Status

Status	$Var(\alpha_{cj})$
Within	0
Between	huge
Within-Between	positive

variance of α_1 is 10000. It is clear from these plots that for most values of α_1 the conditional probability is either 0 or 1. This is especially so when the variance of α_1 is 10000. When this probability is 0 or 1 then C_{ij} can not vary across individuals in the cluster as it is completely determined by the value of the random effect α_1 in the $j-$th cluster. When the variance of α_1 is large the influence of the random intercept α_1 on C_{ij} is large which makes the class variables within a cluster so highly correlated that no variation of the class variables within the cluster is possible.

Table 4 summarizes the modeling possibilities for the latent class variables and the corresponding interpretation for the intercept variance parameter. Note however that when estimating a between level class variable model that there is a numerical advantage for directly specifying the class variable as a between level variable as in Model 5, rather than as a within-between variable with large variance as in Model 6. Both approaches are possible in Mplus 4.2. The advantage of the Model 5 approach is that it does not use numerical integration for the random effect α_{cj}. Model estimation with numerical integration will typically be more computationally demanding.

6 Within and Between Class Variables

All three types of latent class variables given in Table 4 can be used simultaneously in a model. The variables can be measured and predicted by different observed variables, as in the two-level LTA model described in Section 3, or they can be measured and predicted by the same observed variables. Within level latent class variables can be measured and predicted only by within level observed variables. Between level latent class variables can be measured by within and between observed variables but can be predicted only by between observed variables. Within-between latent class variables can be measured and predicted by within level observed variables, while the random effects α_{cj} can be measured and predicted by between observed variables.

In this section we illustrate these modeling combinations by describing a model which includes a within and a between latent class variable. Let C_{ij} be a latent class variable for individual i in cluster j which takes values $1, ..., L$. Let D_j be a between level latent class variable for cluster j which takes values $1, ..., M$. The within level model is

$$[y^*_{ij}|C_{ij} = c, D_j = d] = \nu_{cdj} + \Lambda_{cdj}\eta_{ij} + \varepsilon_{ij} \tag{22}$$

$$[\eta_{ij}|C_{ij} = c, D_j = d] = \mu_{cdj} + B_{cdj}\eta_{ij} + \Gamma_{cdj}x_{ij} + \xi_{ij} \tag{23}$$

The multinomial logistic regression for the within class variable C_{ij} includes the between class variable D_j as a covariate

$$P(C_{ij} = c|D_j = d) = \frac{\exp(\alpha_{cj} + \gamma_{cdj} + \beta_{cj}x_{ij})}{\sum_c \exp(\alpha_{cj} + \gamma_{cdj} + \beta_{cj}x_{ij})} \tag{24}$$

where γ_{cdj} shows the effect of D on C. For identification purposes $\alpha_{Lj} = \beta_{Lj} = \gamma_{Ldj} = \gamma_{cMj} = 0$.

Each of the intercept, slope or loading parameters in equations (22-24) can be either a fixed coefficient or a cluster random effect. If η_j is the vector of all random effects the between level model is then described by

$$[\eta_j|D_j = d] = \mu_d + B_d\eta_j + \Gamma_d x_j + \xi_j. \tag{25}$$

The model for the between level class variable D_j is also a multinomial logit regression

$$P(D_j = d) = \frac{\exp(\alpha_d + \beta_d x_j)}{\sum_d \exp(\alpha_d + \beta_d x_j)} \tag{26}$$

where x_j are the between level covariates.

7 The Grade of Membership Model

In this section we show how to utilize the two-level mixture framework described in Section 2 to estimate the Grade of Membership model (GoM). GoM is not a multilevel model, it is a single level model. The model is an extension of the latent class analysis (LCA) model that allows not only separation of individuals into classes that show similar outcomes but also allows for individuals to be modeled as partially members of several different classes. Individuals with such partial class membership will be those that on some measurements behave like the individual in one class while on some other measurements behave like individuals in another class. Partial class membership is a substantively useful concept for modeling individuals that are in a transitional state. For example when modeling different levels of disability in the elderly, an individual can be classified as healthy or as disabled but can also be in a state of deteriorating health, in which case the individual can be classified as partially healthy and partially disabled and the level of membership in each of the two classes can be specific to that particular individual. The GoM modeling idea also allows us to determine whether individuals transition from one class to another even when we have a cross sectional sample rather than longitudinal. Observing individuals with partial membership at one point in time is an indication that individual transition from one class to another.

In this article we follow the Erosheva (2002) exposition of the GoM model. For simplicity we assume that all observed variables are binary and that there are only two classes in the model, however the model description below generally applies to any type of observed variables and any number of classes. Let Y_{ij} be the $i-$th measurement for individual j. In the LCA model a class variable C_j is defined for each individual and the distribution of Y_{ij} is given by

$$P(Y_{ij} = 1|C_j = c) = \Phi(\tau_{ic}) \tag{27}$$

where Φ is the standard normal distribution function. In the GoM model a latent class variable C_{ij} is defined specifically for the $i-$th measurement of individual j. The distribution of Y_{ij} is given by

$$P(Y_{ij} = 1|C_{ij} = c) = \Phi(\tau_{ic}). \tag{28}$$

The distribution of C_{ij} is given by

$$P(C_{ij} = 1) = f_j \tag{29}$$

where f_j is a subject specific random effect of some kind. Typically f_j is given a Dirichlet prior which facilitates Bayesian estimated methods, see Erosheva (2003), however other functional forms can be used as well. Here we will adopt the logistic regression equation that we have used in this article up to now. Let

$$P(C_{ij} = 1) = \frac{exp(\alpha_{1j})}{1 + exp(\alpha_{1j})} \tag{30}$$

where α_{1j} is a normally distributed random variable. Equations (28) and (30) describe a simple GoM model. It is easy to see that this model is a special case of the two-level mixture model described in Section 2 where now the individual j takes the role of a cluster j and the multivariate vector of all measurements Y_{ij} is treated as a univariate observations clustered in the individual j. Many statistical packages, including Mplus, implement data transformation routine that converts the data from the original "wide-multivariate" format to "long-multilevel" format. Such transformation is applied here as well. It is easy to see now that the GoM model is equivalent to a univariate two-level mixture model. There is one complication in the GoM model that is not directly available in the two-level mixture model. This complication is the fact that the parameter τ_{ic} depends on i, i.e., it varies across observations in the cluster. This complication however can easily be resolved by incorporating dummy variable for all measurements. Suppose that there are L measurements and N individuals in the sample. Define the dummy variables X_{qij} for $i = 1, ..., N$, $j = 1, ..., L$ and $q = 1, ..., L$

$$X_{qij} = \begin{cases} 1 & \text{if } q = i \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

Each observation now consist of one dependent variable Y_{ij} and L independent variables X_{qij} for $q = 1, ..., L$. The equation (28) can now be written as

$$P(Y_{ij} = 1 | C_{ij} = c) = \Phi\left(\sum_{q=1}^{L} \tau_{qc} X_{qij}\right). \tag{32}$$

Note now that the parameters in equation (32) are independent of i and thus the GoM model is part of the two-level framework described in Section 2.

The above formulation of the GoM model has the advantage that it can easily accommodate predictors X_j for the class allocation by adding a regression equation such as

$$\alpha_{1j} = \alpha_1 + \gamma X_j + \varepsilon_j \tag{33}$$

Equivalently the predictors can be added directly in the multinomial logistic regression (30). It is also possible to add item specific covariate. This is done by adding a new covariate in equation (32) which is the product of the dummy variable corresponding to that item and the covariate.

The above formulation of the GoM model also allows us to easily see that the LCA model is nested within the GoM model. In fact, the two class GoM model that we described above has just one more parameter than the 2 class

LCA model. This parameter is the variance of the random effect variable α_{1j}. Note that the LCA is different from the GoM model only by the fact that it imposes the stochastic restriction (16), which basically means that on all measurements for individual j can be in one and the same class. As we explained in the previous section the stochastic restriction (16) is equivalent to fixing the variance of α_{1j} to infinity, or to a numerically equivalent large value. For GoM models with $K > 2$ classes equation (30) is replaced by

$$P(C_{ij} = c) = \frac{exp(\alpha_{cj})}{1 + \sum_{c=1}^{K-1} exp(\alpha_{cj})} \tag{34}$$

where now there are $K - 1$ random effects α_{cj} for $c = 1, ..., K - 1$. The LCA class model with K classes is again nested within the GoM model with K classes, which has $K(K - 1)/2$ more parameter, namely the variances and covariance parameters of the random effects α_{cj}. When the variances are fixed to large values and the covariances to 0 the GoM model becomes equivalent to the LCA model.

In the estimation of the GoM model the random effects α_{cj} are numerically integrated which makes the estimation more challenging when there are large number of classes. To avoid this problem a more restricted model has been proposed by Hedeker (2003) and Vermunt (2003) which essentially adds a one factor analysis model without residuals on the random effects α_{cj}

$$\alpha_{cj} = \alpha_c + \lambda_c \eta_j \tag{35}$$

where η_j is a standard normal random effect and α_c and λ_c are fixed parameters. This model requires only one dimension of numerical integration and has $K - 1$ more parameters, namely the loadings parameters λ_c. Note however that the LCA model is not nested into this more restricted GoM model. When $K = 2$ the restricted GoM (35) is equivalent to the unrestricted GoM.

It is possible to include other types of dependent variables in the above formulation of the GoM model. If the variables are continuous for example the dummy variables structure described above will produce item specific class varying means. For polytomous variables taking d categories the Probit equation (34) should be replaced with a multinomial logistic regression so that each dummy variable contributes $d - 1$ parameters rather than 1 and all item probabilities are unconstrained and class specific.

We now illustrate the GoM model and compare it to the LCA model. We use the Antisocial Behavior (ASB) data taken from the National Longitudinal

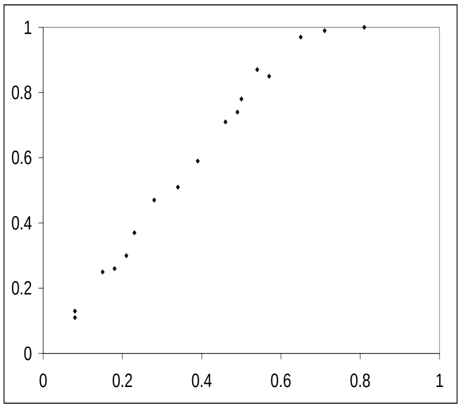
Table 5: Log-Likelihood Comparison for the LCA and GoM Models.

Model	2 class	3 class
LCA	-42625.7	-41713.1
GoM	-42159.1	-41554.6

Survey of Youth (NLSY) that is sponsored by the Bureau of Labor Statistics. These data are made available to the public by Ohio State University. Data for the analysis include 17 antisocial behavior items that were collected in 1980 when respondents were between the ages of 16 and 23. The ASB items assessed the frequency of various behaviors during the past year. A sample of 7,326 respondents has complete data on the antisocial behavior items. The 17 items include 8 property offense items, 5 personal offense items and 4 drug offense items. The items were dichotomized 0/1 with 0 representing offense not occurred in the past year.

Table 5 shows the log-likelihood values obtained for a 2 and 3 class LCA and GoM models. For the 3 class GoM model we used the restricted model (35). For both the two and the three class models the GoM model improved the log-likelihood value substantially. The GoM estimation showed more starting values dependence then the LCA model. To obtain these results we used 30 randomized starting value sets and completed the top 5 after preliminary optimization. A good strategy for selecting starting values for the GoM model estimation is to use the LCA parameter estimates. The GoM model estimation is also more computationally demanding because it needs numerical integration for the random effects α_{cj} and because the data is expanded to include the dummy variables. In the two class model both the LCA and the GoM model essentially split the population into a more offense prone class and less offense prone class. Figure 4 shows the probability profiles for the offense prone class for the two models. Each of the plotted points represents an item, the X coordinate is the LCA probability of occurrence and the Y coordinate is the GoM probability of occurrence. The probability profiles are different. The GoM probabilities in this class are all higher than the LCA probabilities. However the correlation between the two sets of probabilities is 99%. The difference in the probability profiles is due to the fact that the two models identify the classes differently. In the LCA model the classes are identified by the average probability pattern in

Figure 4: 17 Items Probability Profiles for the Offense Prone Class for GoM v.s. LCA models



the class while in the GoM model the classes are identified by the extreme probability pattern that members are drawn to as they become closer and closer to being fully in that class.

8 Combining LCA, IRT and GoM Models

In this section we describe a model that incorporates the modeling capabilities of 3 types of models, namely the LCA, IRT and the GoM model. The combination of LCA and IRT models is sometimes refereed to as factor mixture analysis (FMA) model or alternatively as mixture IRT model. This model was used for example in Qu et. al. (1996) to model residual correlations within a class. In Muthen (2005) and Muthen and Asparouhov (2005) FMA models were also explored as substantively important generalization of the IRT and LCA models.

Following is a brief formulation of the FMA model. As in the previous section for simplicity we assume that Y_{ij} is the $i-$th binary observed variables for individual j. Let C_j be the latent class variable for individual j. The

Table 6: Log-Likelihood Comparison for the LCA and GoM Models.

Model	Log-Likelihood	Number of Parameters
LCA	-3650.0	37
FMA	-3502.4	56
FMA-GoM	-3501.7	57

model is then described by the following two equations

$$P(Y_{ij} = 1 | C_j = c) = \Phi(\tau_{ic} + \lambda_{ic}\eta_j) \tag{36}$$

$$P(C_j = c) = \frac{exp(\alpha_c)}{1 + \sum_{c=1}^{K-1} exp(\alpha_c)} \tag{37}$$

where η_j is a standard normal random effect and τ_{ic}, λ_{ic} and α_c are fixed parameters. To generalize this model to include partial class membership into an FMA-GoM model we proceed as in the previous section. Let C_{ij} be an item specific latent class variable. The FMA-GoM model is described by the following equations

$$P(Y_{ij} = 1 | C_{ij} = c) = \Phi(\tau_{ic} + \lambda_{ic}\eta_j) \tag{38}$$

$$P(C_{ij} = c) = \frac{exp(\alpha_{cj})}{1 + \sum_{c=1}^{K-1} exp(\alpha_{cj})} \tag{39}$$

As in the previous section equation (38) is equivalent to

$$P(Y_{ij} = 1 | C_{ij} = c) = \Phi\left(\sum_{q=1}^{L} \beta_{qc} X_{qij}\right). \tag{40}$$

$$\beta_{qc} = \tau_{qc} + \lambda_{qc}\eta_j \tag{41}$$

where α_{cj} are normally distributed random variables. This model is again a special case of the framework described in Section 2. All random effects α_{cj} and η_j are now numerically integrated. Just as the LCA model is nested in the GoM model the FMA model (36-37) is nested in the FMA-GoM model (39-41). For example for a two class model the FMA-GoM model has just one more parameter.

For illustration we estimate a 2 class FMA-GoM model for a UCLA clinical sample of 425 males ages 5-18, all with ADHD diagnosis. The data consists of nine inattentiveness items and nine hyperactivity items all dichotomously scored. For simplicity we estimate the restricted FMA model where the factor loadings λ_{ic} are class invariant, $\lambda_{ic} = \lambda_i$. Table 6 shows the log-likelihood values and the number of parameters for the 2 class LCA, FMA and FMA-GoM models. In this example, the FMA improved the likelihood dramatically over the LCA model however the GoM-FMA improved the likelihood only marginally. The ICC of the α_{1j} was estimated to be 86%. Thus in this example the concept of partial class membership is not supported by the data. The substantive conclusion appears to be that individuals are never in a transitional phase and are preset to be in one of the two classes.

One of the original applications of the FMA/Mixture IRT model is the ability to separate individuals into classes that respond similarly to the various items. For example, individuals solve mental rotation problems using one of several solution strategies. The Mixture IRT model allows us to separate the population into classes that appear to be using the same solving strategy. Adding the GoM modeling idea to the Mixture IRT model will allow us to also model individuals that may use one strategy on one items but another strategy on a different item.

9 The Three Parameter Guessing IRT Model

In this section we show that the three parameter (3PL) guessing IRT model is a special case of the GoM-FMA model described in the previous section. The 3PL model is described by the following equation

$$P(Y_{ij} = 1) = g_i + (1 - g_i)\Psi(a_i(\eta_j - b_i)) \tag{42}$$

where Ψ is the normal or the logistic distribution function. Now let's consider the following GoM-FMA model with 2 classes

$$P(Y_{ij} = 1|C_{ij} = c) = \Psi(\tau_{ic} + \lambda_{ic}\eta_j) \tag{43}$$

$$P(C_{ij} = 1) = \frac{exp(\alpha_{1j} + \sum_q \gamma_q X_{qij})}{1 + exp(\alpha_{1j} + \sum_q \gamma_q X_{qij}))} \tag{44}$$

The difference between model (38-39) and model (43-44) is that in the multinomial logistic regression on C_{ij} we have included covariates. Let's now constraint several of the parameters in the above model

$$\alpha_{1j} = 0$$

$$\lambda_{i1} = 0$$

$$\tau_{i1} = 15.$$

The constant 15 above is chosen to be sufficiently high so that $\Psi(15) \approx 1$. This approximation holds for both the normal and the logistic distribution. Given these constraints we can simplify the above model

$$P(Y_{ij} = 1|C_{ij} = 1) = 1 \tag{45}$$

$$P(Y_{ij} = 1|C_{ij} = 2) = \Psi(\tau_{i2} + \lambda_{i2}\eta_j) \tag{46}$$

$$P(C_{ij} = 1) = \frac{exp(\gamma_i)}{1 + exp(\gamma_i)} \tag{47}$$

It is now easy to see that model (42) is just a reparameterization of model (45-47). The parameters in model (42) are obtained from the parameters of model (45-47) via the following equations

$$g_i = \frac{exp(\gamma_i)}{1 + exp(\gamma_i)} \tag{48}$$

$$a_i = \lambda_{i2} \tag{49}$$

$$b_i = -\frac{\tau_{i2}}{\lambda_{i2}}. \tag{50}$$

10 Technical Aspects of The Estimation

All models presented in this article were estimated with Mplus Version 4.2. Mplus uses maximum-likelihood estimation with robust standard error estimation (see White (1980)).

The estimation of Multilevel Mixture Models presents a number of challenges. The maximum likelihood estimation of mixture models in general is susceptible to local maximum solutions. To avoid this problem Mplus uses an algorithm that randomizes the starting values for the optimization routine. An initial sets of random starting values are first selected. Partial optimization is performed for all starting value sets which is followed by complete optimization for the best few starting value sets. It is not clear how many starting value sets should be used in general. Different models and data may require different starting value sets. Most results in this article were obtained by selecting 20 initial sets and completing the best 5. One useful criterion that the starting value perturbation has been thorough is that the best log-likelihood value is reached at least twice, however even if this criteria is satisfied, it is no guarantee that the number of starting value sets is sufficient. A sound strategy to minimize the impact of the starting values of the optimization routine is to build Multilevel Mixture Models gradually starting with simpler models that have few random effects and classes. Consequently one can use the parameter estimates from the simpler models for starting values for the more advanced models.

Another estimation challenge is the fact that most Multilevel Mixture Models require numerical integration techniques for some of the normally distributed latent variables in the model. Adaptive numerical integration can be performed in Mplus as well as non-adaptive. In general adaptive numerical integration tends to be more accurate but it is also more unstable and frequently can fail the optimization process. In such situation Mplus will abandon adaptive integration and will use non-adaptive integration. Gauss-Hermite, the trapezoid and the Monte-Carlo integration methods are implemented in Mplus. It is well know that Gauss-Hermite is very dependent on the adaptiveness of the integration method and without it can produce very inaccurate results. This is not the case for the trapezoid integration method which performs quite well even without adaptive integration. Given the instability of adaptive integration we prefer the trapezoid integration method over the Gauss-Hermite integration method, which is also the Mplus default and the method we used for the results presented here. The Monte-Carlo integration method is appropriate when the number of integration dimensions is high (5 or more). It is usually the least accurate integration method. Model parameterization is also vary important and rather broad but also unexplored component in the estimation of Multilevel Mixture Models. Certain parameterizations will facilitate faster convergence of the optimization, see for example the PM-EX algorithm of Liu et. at. (1997). Another parameterization that generally improves convergence speed is the Cholesky parameterization, see Hedeker and Gibbons (1996). This parameterization is available for all Multilevel Mixture Models in Mplus. Another attractive feature of this parameterization is that when it is used with non-adaptive quadrature it can guarantee monotonically increasing likelihood in the EM-algorithm which makes the convergence process very stable even with small number of integration points and large number of integration dimensions. The parameterization in the model can also affect the number of dimensions of the numerical integration and therefore affect dramatically the computational speed. Ultimately choosing the most optimal parameterization for a model is more difficult than choosing other technical options. Mplus is very flexible and can be used with most parameterizations, especially since Mplus can implement a separate auxiliary parameterization model in addition to the statistical model. There are other technical options related to numerical integration, see Muthen and Muthen (1998-2006). A sound strategy when selecting these technical options would be that they should not affect the estimation results and if they do such effects should conform with the published literature on this topic.

11 Conclusion

In this article we described a modeling framework that incorporates three popular modeling techniques, multilevel modeling, structural equation modeling and finite mixture modeling. This modeling framework has the potential of uncovering previously unexplored aspects of the data. Two-level analysis with multiple latent categorical variables was illustrated with a two-level latent transition analysis. We also described how heterogeneity can be modeled as a within, between, or a within-between phenomenon. We illustrated how the GoM modeling idea can be incorporated within a single level mixture model to allow partial class membership. The GoM models are estimated within the two-level mixture modeling framework.

The Mplus user's guide, Muthen and Muthen (2006), has a number of other practical multilevel mixture examples as well as details on the Mplus model specifications.

Another important application of this framework is the non-parametric hierarchical regression models. These models provide an alternative to the popular hierarchical regression models with normally distributed random effects by assuming a more realistic non-parameteric distribution instead. A detailed discussion on this topic is available in Muthen and Asparouhov (2006).

Two-level mixture models vary greatly in their complexity. In this article we illustrated most of the basic modeling principles. Researchers familiar with multilevel models and mixture models will find it easy to combine these modeling ideas. Sound modeling strategies should be used with these complex new models. Gradual model building, comparison with single class models and single level models should always be performed. The flexible modeling framework we described in this article will offers researchers many competing modeling strategies. Rigorous statistical techniques should be used to choose among these alternatives. In addition, researchers should promote models that have solid connections with substantive theory.

12 References

Agresti, A. (1996) An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc. New York, New York, USA

Asparouhov, T. (2006). General Multilevel Modeling with Sampling Weights. Communications in Statistics: Theory and Methods, Volume 35, Number 3, pp. 439-460(22).

Bijmolt, T.H., Paas, L.J., Vermunt , J.K. (2004) Country and Consumer Segmentation: Multi-level Latent Class Analysis of Financial Product Ownership. International Journal of Research in Marketing, 21, 323-340.

Erosheva, E. (2002). Partial Membership Models With Application to Disability Survey Data. In H. Bozdogan, ed., Proceedings of Conference on the New Frontiers of Statistical Data Mining, CRC Press, 117-134.

Erosheva, E. (2003) Bayesian Estimation of the Grade of Membership Model, Bayesian Statistics 7, Oxford University Press, 501-510.

Hedeker, D. (2003) A mixed-effects multinomial logistic regression model. Stat Med. 22(9), 1433-46.

Hedeker, D. & Gibbons, R. D. (1996). MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. Computer Methods and Programs in Biomedicine, 49, 229-252.

Liu, C.H., Rubin, D.B., and Wu, Y.N. (1998), Parameter Expansion to

Accelerate EM - the PX-EM algorithm. Biometrika 85(4):755-770.

McCullagh P. & Nelder, J. A. (1989) Generalized Linear Models. London. Chapman & Hall.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Muthen, B. (2006). Should substance use disorders be considered as categorical or dimensional? Addiction, 101 (Suppl. 1), 616.

Muthen, B., Brown, C.H., Masyn, K., Jo, B., Khoo, S.T., Yang, C.C., Wang, C.P., Kellam, S., Carlin, J., & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. Biostatistics, 3, 459-475.

Muthen, B. & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics, 55, 463-469.

Muthen, B. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. Addictive Behaviors, 31, 1050-1066.

Muthen, B. & Asparouhov, T. (2006). Non-parametric hierarchical regressions. In preparation.

Muthen, L.K. & Muthen, B.O. (1998-2006). Mplus User's Guide. Fourth Edition. Los Angeles, CA: Muthen & Muthen

Qu Y, Tan M and Kutner MH. (1996) Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. Biometrics, 52:797-810.

Vermunt, J. (2003) Multilevel Latent Class Models, Sociological Methodology, 33, 213-239.

Vermunt, J. & Magidson J. (2005) Hierarchical mixture models for nested data structures. C. Weihs & W. Gaul, Classification: The Ubiquitous Challenge, 176-183. Heidelberg: Springer

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica, 41,* 733-750.