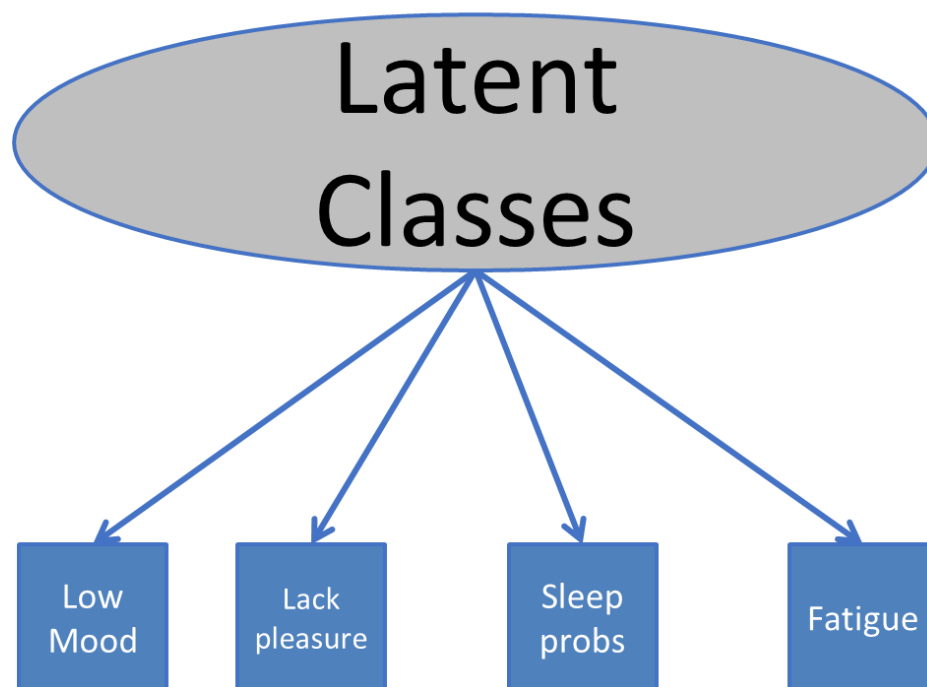


## Introduction to Latent Class Analysis (LCA) with Covariates and Distal Outcomes

I have highlighted that LCA is, firstly, a measurement model. Researchers use LCA to make sense of inter-individual differences and to identify categories or groups (i.e. classes) of individuals that differ in their propensity to display patterns of behaviour. Most researchers want to use LCA to identify categories of individuals, and then investigate what explains these individual differences. Often, they also want to investigate what are the consequences of individual differences in behaviour patterns. For example, if we identify different latent classes of individuals with mental health issues, do individuals in these categories vary in their responses to Cognitive Behavioural Therapy?

A problem that has marred the use of LCA in practice lies in the process of estimating the measurement model *concurrently* with covariates or distal outcomes. To illustrate, let's consider the example in Figure 1 where we use observed symptoms of depression to identify different latent classes:



*Figure 1: Schematic example of latent class estimation based on four indicators.*

Once we have a satisfactory latent class model that explains heterogeneity in patterns of symptoms, let's say we want to test if the estimated latent classes significantly predict the age of retirement, a variable we collect some years after we had tested our latent class

model of depression. The most obvious thing would be to estimate the latent classes based on the four indicators (observed symptoms) and, **at the same time**, regress distal outcome “Age at time of retirement” on the latent classes we estimate, as in Figure 2.

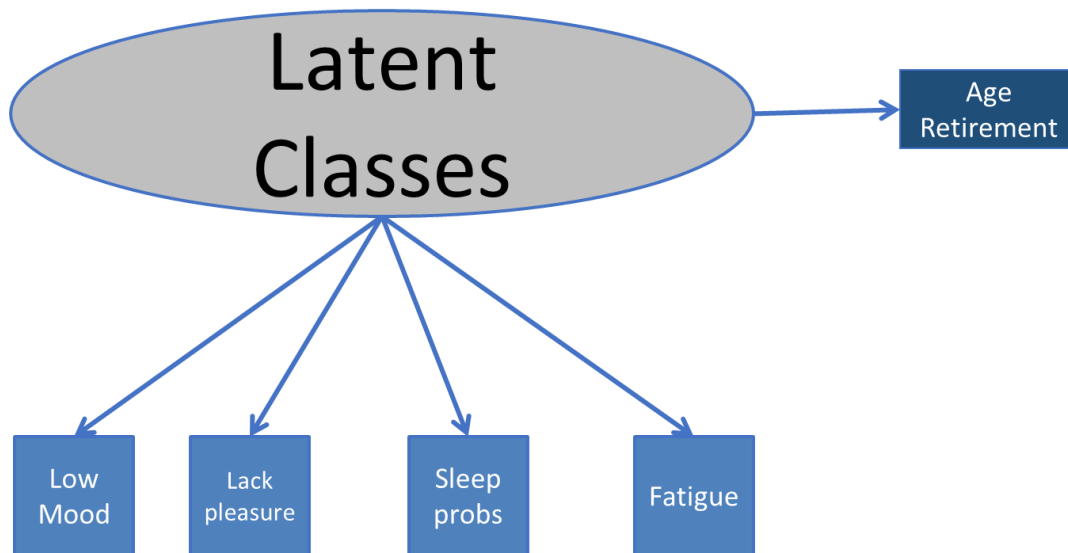


Figure 2: Schematic example of latent class estimated with four indicators and a distal outcome

Figure 2 may alert you about the main problem in this model: There is nothing that distinguishes the regression of symptoms of depression on the latent classes from the regression of the distal outcome on the latent classes. Therefore, if we run the latent class model in Figure 2 **the latent classes will represent individual variation (heterogeneity) in the symptoms and in the distal outcome.**

This causes **practical problems**: while we had a satisfactory latent class model of depression when we only included the observed symptoms (as in Figure 1), our latent class measurement model will be very different when we introduce the distal outcome, as in Figure 2. Maybe the optimal latent class model in Figure 2 will even have a different number of classes compared to the measurement model in Figure 1. We will have to re-analyse and interpret the new model in Figure 2, since this model represents heterogeneity in the indicators *and* the distal outcome.

The problems are **not just practical**, as we also have the problem of interpreting a model that represents heterogeneity in variables collected on different occasions, where the distal outcome represents events that may take place years after we had collected information on the symptoms.

The same problems arise when we consider covariates that can predict latent class affiliation (e.g. Gender, and Socio-Economic Status). If we estimated the latent class model with the observed indicators while, at the same time, including covariates, the latent class model will represent the heterogeneity in the indicators (symptoms) and the covariates. Apart from being impractical, this approach is not really answering questions we are posing about mechanisms leading to differences in behaviour patterns.

**Different solutions have been tried to solve this problem.** The more naïve solution would be to estimate the latent class measurement model (as in Figure 1) and then consider to which latent class participants are most likely to belong. The latent classes assignment is then used as a nominal variable in further analyses: We can investigate the association between covariates and participants' class affiliation, or that between latent class affiliation and distal outcomes. For example, do individuals in different latent classes of depression retire at significantly different ages?

The main problem with this naïve approach is that it does not take into account the measurement error in latent class membership. The latent class models are probabilistic, and participants' assignment to estimated latent classes (their latent class membership) is uncertain. If we fail to account for this uncertainty and use latent class membership as a variable in a model, we will obtain biased results. See

<http://statmodel.com/download/relatinglca.pdf> for a more in-depth discussion.

There are other solutions (e.g. use probability weights for the estimated latent class affiliations), but the most satisfactory ones are **Multiple Pseudo-Class Draws**, and the **Three-Step Approach**. The following sections introduce these two approaches, which will be the focus of the two exercises proposed later.

### Multiple Pseudo-Class Draws

This approach proposes to control for uncertainty in latent class membership by using a method akin to multiple imputation of missing data. In fact, after estimating a latent class model, we can consider each participants' posterior latent class probabilities and use these to create multiple datasets (e.g.  $n = 100$ ) where each participant is randomly assigned to latent classes based on these posterior probabilities. Therefore, in these multiple datasets, the random draws will provide a set of plausible values of latent class membership, but at the same time representing the uncertainty about this membership.

Once these datasets have been created, the latent class draws can be used as a variable in regression analyses by combining these analyses using Rubin's procedure and the same rules derived for multiple imputation of missing data.

Mplus facilitates the application of this approach through syntax. In fact, predictors of latent class membership can be specified in the **VARIABLE:** command as “Auxiliary” variables. For example, if we wanted to include gender, or better, a dummy-variable indicating gender male, and dummy-coded variables representing Socio-Economic Status quintiles, we would indicate as “**Auxiliary**” variables in this way:

```
VARIABLES:
NAMES: id mood anhedonia sleep fatigue male ses1 ses2 ses3 ses4 ses5 ageretir;
USEVAR = mood anhedonia sleep fatigue;
CATEGORICAL = mood anhedonia sleep fatigue;
MISSING = all (-999);
CLASSES= depress(2);
AUXILIARY= male (R) ses2-ses5 (R);
```

The command **AUXILIARY=** together with the **(R)** following the variable names instructs Mplus to consider these variables as predictors of latent classes in multinomial logistic regressions, where the categorical latent classes are estimated using posterior probability-based multiple imputations (pseudo-class draws).

Note the use of *dummy-coded variables*: the SES variable has 5 levels (i.e. 5 quintiles), and each of these levels is represented by dummy variables where individuals in quintile 1 received score=1 in variable **ses1** and those in other quintiles receive score=0, and so on, until we have 5 dummy variables for each quintile: **ses1, ses2, ses3, ses4, ses5**. By omitting one of these variables (**ses1** in the example above) we are instructing the software to consider the omitted variable as a reference category for comparisons. The multinomial logistic regression will therefore represent the changes in the probability of being in different classes for individuals in SES quintile 2, 3, etc. when compared to individuals in SES quintile 1.

In order to test the association between latent classes and a distal outcome, we can use the **AUXILIARY=** option in **VARIABLE:** with a different notation:

```
VARIABLES:
NAMES: id mood anhedonia sleep fatigue male ses1 ses2 ses3 ses4 ses5 ageretir;
USEVAR = mood anhedonia sleep fatigue;
CATEGORICAL = mood anhedonia sleep fatigue;
MISSING = all (-999);
CLASSES= depress(2);
AUXILIARY= ageretir(E);
```

The last line in the box above instructs Mplus to test the null hypothesis of equal means in variable **ageretir** (Age at time of retirement) across the latent classes estimated using posterior probability-based multiple imputations (pseudo-class draws).

Note that it is not possible to specify auxiliary some variables as predictors (**R**) and other as distal outcomes (**E**) at the same time in the **AUXILIARY=** option.

### The Three-Step Approach: Introduction

This approach has been more recently developed. The solution to the problem of including covariates and distal outcomes lies in conducting the measurement model and the modelling of structural relationships (e.g. regressing latent classes on covariates) in separate steps (respectively the first and the third steps of this procedure). An intermediate step links the other two steps by estimating measurement error in class assignment, thus allowing to control for this error when imposing structural relationships between other variables and the latent classes estimated.

I will illustrate these steps with a practical example.

- **Step 1: Estimate the Optimal Model and Assign Individuals to the Most Likely Class (Modal Class)**

Let's assume we have estimated two latent classes based on the frequency of Depression symptoms. The output of the model will provide posterior probabilities of being in each of these two classes, with the "most likely" latent class membership for each individual, see Figure 3.

ID	Low Mood	Anhedonia	Sleep probs.	Fatigue	p Class1	p Class2	Most likely class
101	1	1	2	1	.043	.957	2
102	3	3	2	3	.969	.031	1
103	1	2	1	1	.099	.901	2

104	2	1	3	2	.424	.576	2
...							

Figure 3: Fictional example of data representing frequency of symptoms (higher value=more frequent), probability of membership in two latent classes, and the most likely class (Latent Class Modal Assignment).

The most likely class to which each individual is assigned will be used in Step 3 as a nominal variable to estimate class membership while controlling for uncertainty in this membership, as I will illustrate in Step 3. Before that, I will explain the necessary steps to obtain estimates of uncertainty in latent class estimation.

- **Step 2: Estimate measurement error (i.e. uncertainty in class allocation)**

As highlighted in other occasions, these posterior probabilities indicate the level of uncertainty in class membership. For example, while membership into Class 1 appears more certain for ID=102, membership into Class 2 for ID=104 appears quite uncertain.

We can use these probabilities to calculate the average probability of being in each class if the most likely class is 1 or 2. Considering the example in Figure 3, **the average probability of being in latent Class 2 if the most likely class=2** will be given by:

$$\frac{0.957 + 0.901 + 0.576}{3}$$

That is, the probability of being in latent Class 2 for IDs 101, 103, and 104, who are most likely in latent Class 2.

In the same way, we can calculate all the others average probabilities of being in class 1 or 2 if the most likely class is 1 or 2. These average probabilities can then be reported in Table like the one in Figure 4.

	Class 1	Class 2	N
Class 1	0.924	0.076	3,472
Class 2	0.054	0.946	5,449

Figure 4: Average Latent Class Probabilities for Most Likely Latent Class Membership (Rows) by Latent Class (Columns)

For example, the first cell in the table in Figure 4 represents the average probability of being in latent Class 1 if the most likely latent Class = 1 (0.924). The *N*s in the last columns represent the number of participants who, in this fictional example, have been assigned to latent Class 1 and latent Class 2, respectively.

Taking the table in Figure 4 as a reference, we can then calculate the **classification probabilities** for the most likely latent class membership by latent class. For example, the classification probability when the most likely class membership is Class 1 and individuals are classified in latent class 1 will be equal to:

$$\frac{(0.924 * 3,472)}{(0.924 * 3,472) + (0.054 * 5,449)} = 0.916$$

Namely, this classification probability is equal to the product of the average probability of being in Class 1 when the most likely class=1 by the number of individuals whose most likely class=1, divided by the sum of the latter product and the product of the average probability of being in Class 1 when the most likely class=2 by the number of individuals whose most likely class=2.

In the same way, we can calculate the other classification probabilities, which we can then report in another table, see Figure 5:

	Class 1	Class 2
Class 1	0.916	0.084
Class 2	0.049	0.951

*Figure 5: Classification Probabilities for the Most Likely Latent Class Membership (Rows) by Latent Class (Columns).*

Now we can use these classification probabilities to calculate the logit ratios of being in Class 1 rather than Class 2 when the most likely class=1:

$$\ln\left(\frac{0.916}{0.084}\right) = 2.389$$

Similarly, we can calculate the logit odds of being in Class 1 rather than in Class 2 when the most likely class=2:

$$\ln\left(\frac{0.049}{0.951}\right) = -2.972$$

- **Step 3: Impose structural relationships between classes and covariates/distal outcomes, while controlling for measurement error in class assignment**

In this final step we use the information from Step 1 (i.e. the most likely class membership of each participant) and from Step 2 (i.e. the measurement error expressed by the logits for classification probabilities) to create a latent class model that is defined by these estimated values. In other words, the latent class model is fixed to these values that reflect the uncertainty in latent class membership, and we can therefore add covariates and distal outcomes without re-estimating the latent class model. In Figure 6 the 3<sup>rd</sup> step in this approach is represented schematically.

Figure 6 highlights that the association between the most likely class and the latent class is fixed at the measurement error parameters estimated in Step 2: therefore the latent class model is given and will not change.



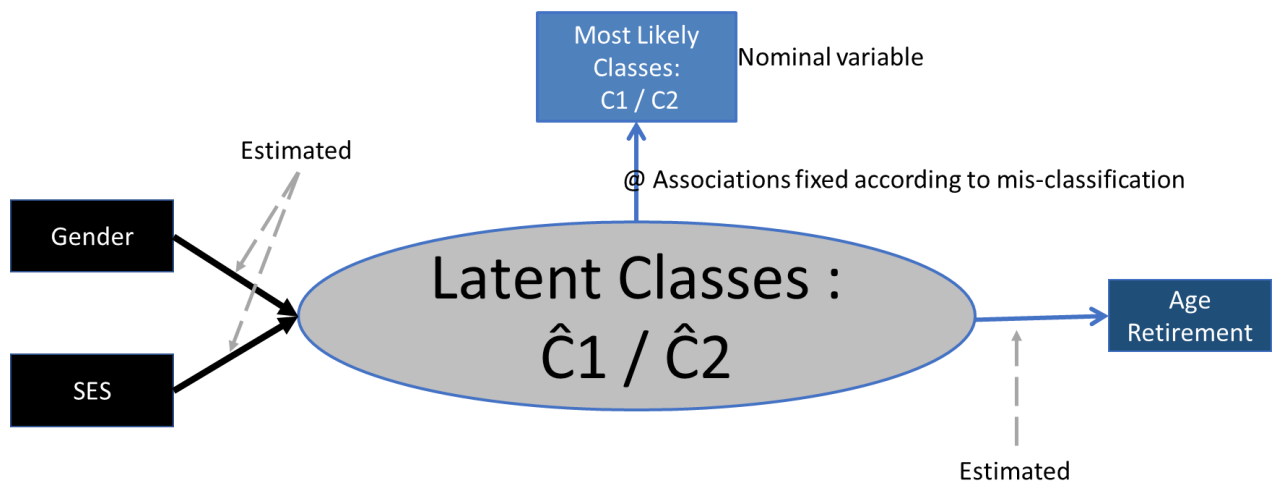


Figure 6: Schematic representation of Step 3 in the Three-Step Approach.

### The Three-Step Approach in Mplus

The Three-Step Approach is facilitated in Mplus by the fact that the logit odds that are used to fix the measurement parameters in the 3<sup>rd</sup> step are readily available in the Mplus Output when running latent class models.

- **Step 1**

The first step is to estimate the latent class model. If, for example, after initial analyses the optimal model for the data appears to be a model with 2 latent classes, estimate this model ensuring that a data file is saved that includes the posterior latent class probabilities and the most likely class membership for each participant.

To this aim, add **SAVEDATA:** command in the INPUT file. For example:

```
SAVEDATA:
FILE= twoclasses.dat;
SAVE=cprob;
MISSFLAG=-999;
```

The options above indicate the name of the datafile that will be created when Mplus runs the model ("twoclasses.dat"). Note that you can also specify the path where you want to save this file, e.g.: **FILE= "C:\DESKTOP\ twoclasses.dat"**; You can also save the datafile in other text-based formats (e.g. .txt).

The option **SAVE=cprob**; ensures that the datafile created will include the posterior probabilities of latent class membership, as well as the most likely class of each participant (as long as the participant has valid data for at least one of the indicators).

The option **MISSFLAG= -999**; instructs Mplus to assign value -999 to cells with missing data.

To ensure this datafile can also be match-merged with other datafiles for checks and other uses, make sure you also save the participants IDs in the datafile created by Mplus. To this end, include the ID variable in the **VARIABLE:** command using the option **IDVAR=** , as in the example below:

```
VARIABLES:
NAMES= id mood anhedonia sleep fatigue male ses1 ses2 ses3 ses4 ses5 ageretir;
USEVAR = mood anhedonia sleep fatigue;
CATEGORICAL = mood anhedonia sleep fatigue;
MISSING = all (-999);
CLASSES= depress(2);
IDVAR=id;
```

Since the datafile that Mplus will produce after this model estimation will include the most likely class membership, which will be used in the 3<sup>rd</sup> step of the analysis, it would be useful to also ensure that covariates and distal outcomes are saved in the datafile. We can do this by adding option **AUXILIARY=** and the name of the variables we want to transfer in the datafile that Mplus will create:

```
VARIABLES:
NAMES= id mood anhedonia sleep fatigue male ses1 ses2 ses3 ses4 ses5 ageretir;
USEVAR = mood anhedonia sleep fatigue;
CATEGORICAL = mood anhedonia sleep fatigue;
MISSING = all (-999);
CLASSES= depress(2);
IDVAR=id;
AUXILIARY= male ses1 ses2 ses3 ses4 ses5 ageretir;
```

The last line in the box above ensures that variables listed after **AUXILIARY=** will not be included in model estimation, but will be saved in the datafile we will create using command **SAVEDATA:**

After estimating the model, the Mplus OUTPUT will provide information about the datafile it created:

```

SAVEDATA INFORMATION

Save file
  twoclasses.dat

Order and format of variables

  MOOD          F10.3
  ANHEDONI      F10.3
  SLEEP         F10.3
  FATIGUE       F10.3
  ID            F8.0
  MALE          F10.3
  SES1          F10.3
  SES2          F10.3
  SES3          F10.3
  SES4          F10.3
  SES5          F10.3
  AGERETIR      F10.3
  CPROB1        F10.3
  CPROB2        F10.3
  DEPRES        F10.3

Save file format
  4F10.3 F8.0 10F10.3

Save file record length  10000

```

The variable **CPROB1** and **CPROB2** are the probabilities of membership in latent class 1 and in latent class 2 respectively, and the variable **DEPRESS** represents the most likely class membership for each participant. Note that “**depress**” is the name I gave to the latent class variable after the **VARIABLE:** command: you can give your latent class variable any name (within Mplus rules, e.g. names should not exceed 8 characters).

## • Step 2

In this step we estimate measurement errors in latent class membership for the model we estimated in Step 1. Mplus facilitates this task by providing in the OUTPUT file tables with the average latent class probabilities, classification probabilities and, crucially, the logits for the classification probabilities.

In the example of the 2-class model estimated in Step 1, we obtain a table such as this in the OUTPUT file:

Logits for the Classification Probabilities for the Most Likely Latent Class Membership (Row)  
by Latent Class (Column)

	1	2
1	2.295	0.000
2	-2.775	0.000

We can use these logits as measurement errors in latent class affiliation in Step 3.

- **Step 3**

In this step we will use the datafile we obtained in Step 1:

#### SAVEDATA INFORMATION

Save file

twoclasses.dat

Order and format of variables

MOOD	F10.3
ANHEDONI	F10.3
SLEEP	F10.3
FATIGUE	F10.3
ID	F8.0
MALE	F10.3
SES1	F10.3
SES2	F10.3
SES3	F10.3
SES4	F10.3
SES5	F10.3
AGERETIR	F10.3
CPROB1	F10.3
CPROB2	F10.3
DEPRES	F10.3

Save file format

4F10.3 F8.0 10F10.3

Save file record length 10000

The variable file name and variable names are those that Mplus indicated, so we will write a similar INPUT file:

#### DATA:

FILE= twoclasses.dat ;

#### VARIABLES:

```

NAMES= mood anhedonia sleep fatigue id male ses1 ses2 ses3 ses4 ses5 ageretir cprob2
cprob2 depress;

USEVAR = depress male ses2 ses3 ses4 ses5;
NOMINAL = depress;

MISSING = all (-999);

CLASSES= newcl(2);

```

Note that the order of the variables must follow exactly the order in which Mplus put these variables in the datafile.

In the **VARIABLE:** command, we will define the variable **depress** as a nominal variable. This is the variable that represents the most likely class membership for each participant. This variable is then used to estimate latent class membership in a new latent class variable with 2 classes, **newcl**, specified in **CLASSES= newcl(2);**

To ensure the association between the most likely class (variable **depress**) and the **newcl** variable is fixed according to the measurement error estimated in Step 2, we will fix the association between the indicator **depress** and **newcl** in the **MODEL:** command in this way:

```

MODEL:
%OVERALL%
newcl ON male ses2 ses3 ses4 ses5;

%newcl#1%
[depress#1 @ 2.295];
%newcl#2%
[depress#1 @ -2.775];

```

Remember that the **%OVERALL%** statement in **MODEL:** specifies the part of the model that concerns all latent classes. In the box above, we are instructing Mplus to estimate the multinomial regression of latent classes **newcl** on covariates **male** and SES (through the use of dummy variables **ses2**, etc.).

The **%newcl#1%** statement concerns just class 1 of the latent variable **newcl**. The statement **[depress#1 @ 2.295];** is fixing the measurement relationship between the nominal most likely class variable **depress** and latent class **newcl** to the level of uncertainty determined in

Step 2. This is effectively fixing the estimation of latent class to the measurement error determined in Step 2, therefore avoiding a new estimation of the latent class measurement model.

Because of that, when running Step 3, the **STARTS=** option in command **ANALYSIS:** should be set to 0. This avoids re-estimating the measurement model, since the model has been fixed to the level of uncertainty determined in Step 2. Thus, the **ANALYSIS:** command should state:

```
ANALYSIS:
TYPE=MIXTURE;
STARTS=0;
```

Putting all this together, Mplus will run multinomial regression models where latent class affiliation into Class 1 or Class 2 is regressed on the covariates, and the latent class affiliation is represented with the uncertainty.

In Step 3 it is also possible to estimate the association between latent classes and distal outcomes such as **ageretir** (Age at time of retirement). Since this variable is continuous, we can estimate the average value of this variable across the two latent classes estimated:

```
DATA:
FILE= twoclasses.dat ;

VARIABLES:
NAMES= mood anhedonia sleep fatigue id male ses1 ses2 ses3 ses4 ses5 ageretir cprob2
cprob2 depress;
USEVAR = depress male ses2 ses3 ses4 ses5 ageretir;
NOMINAL = depress;
MISSING = all (-999);
CLASSES= newcl(2);

ANALYSIS:
TYPE=MIXTURE;
STARTS=0;

MODEL:
%OVERALL%
newcl ON male ses2 ses3 ses4 ses5;
```

```
%newcl#1%  
[depress#1 @ 2.295];  
[ageretir] (p1);  
  
%newcl#2%  
[depress#1 @ -2.775];  
[ageretir] (p2);  
  
MODEL TEST:  
p1=p2;
```

The statements **[ageretir]** in **%newcl#1%** and **%newcl#2%** ask Mplus to estimate the average value of **ageretir** for latent class 1 and latent class 2 respectively. By adding a name (p1) and (p2) for these two estimated means, we can use the **MODEL TEST:** command to invoke a Wald test testing the null hypothesis that the mean of **ageretir** for latent class 1 (which we labelled p1) is equal to the mean of **ageretir** for latent class 2. If the  $p$  value of the test is  $<.05$ , we can reject the null hypothesis and accept  $p1 \neq p2$ . In a similar way, we can also free the variances of the distal outcome to differ across classes, and test hypotheses concerning them.