LATENT CLASS ANALYSIS

I. INTRODUCTION

- In 1950 Paul Lazarsfeld developed *latent structure analysis* to describe the use of mathematical models for characterizing latent variables in the analysis of attitudinal measures from survey research

- four (4) latent structure methods based on the type of latent and observed (*manifest*) variables

LATENT VARIABLES	OBSERVED INDICATOR	
	Discrete	Continuous
Discrete	Latent Class Analysis	Latent Profile Analysis
Continuous	Latent Trait Analysis	Factor Analysis

- In LCA, the latent variable is static (unchanging) and divides the population into a set of mutually exclusive and exhaustive homogeneous groups, referred to as *latent classes*. This set of latent classes account for the distribution of cases that occur within a cross-tabulation of observed discrete variables.

- LCA views the population as comprising of subpopulations

- It enables the researcher to determine or measure, as to which subpopulation an individual belongs.

- *exploratory latent class analysis* : uses the latent class model to explore the latent structures among a set of observed variables

- *confirmatory latent class analysis* : tests hypotheses about the latent structures among a set of observed variables

A. DEFINITION OF TERMS

<u>Observed Variables</u> – discrete (or categorical) variables that are indirect indicators of a person's (object's) class membership. These are also known as "manifest" or "response" variables.

<u>Latent Variable</u> – an antecedent variable that causes interrelationships among observed variables. "Latent" refers to the fact that a person's (or object's) class membership is not directly observed but rather is inferred from observed variables.

Latent Classes – categories of a latent variable, each one of which contains individuals who are similar to each other and different from individuals in other categories. These are the homogeneous groups of people (or objects) from a heterogeneous population.

B. APPLICATION OF LATENT CLASS ANALYSIS

well-suited to marketing research, survey research, sociology, psychology and education

uses patterns of responses (or attitude structures) to observed categorical variables to identify the number of underlying classes, classify each individual into one class, or "cluster" individuals rather than response variables

also used in health studies / applications to identify disease subtypes or diagnostic subcategories given the presence / absence of several symptoms

Example

This example is an unrestricted latent class problem about opinions on male and female role patterns. The data is based on a survey of 60 respondents (30 male and 30 female) at least 16 years old (at least a college freshman). Time constraints did not allow for a larger sample size. To illustrate the application of LCA, the data were bootstrapped to yield responses for 400 respondents (200 Male and 200 Female). LCA is used to examine whether the items differentiate the opinions and points of view of the population.

Respondents were asked to indicate agreement/disagreement to the following statements:

1. Women's liberation sets women against men.

2. It's better for a wife not to have a job because that always poses problems in the household, especially if there are children.

3. The most natural situation occurs when the man is the breadwinner and the woman runs the household and takes care of the children.

4. It isn't really as important for a girl to get a good education as it is for a boy.

5. A woman is better suited to raise small children than a man.

The responses on these five items were rated on a 4-point scale. For purposes of illustration, these possible responses were dichotomized as follows:

Category 1: strongly agree or agree

Category 2: strongly disagree or disagree

Descriptive Statistics

Table 1

Mean Age of the Respondents		
	Original Sample	As Bootstrapped
Male	29.27	29.36
Female	31.43	29.85
Both	30.35	29.61
N	60	400

Table 2

Percentage of Single and Married Respondents (Original Sample)			
	Single	Married	Frequency
Male	35.00%	15.00%	30
Female	30.00%	20.00%	30
Both	65.00%	35.00%	60
Frequency	39	21	

Percentage of Single and Married Respondents (As Bootstrapped)			
	Single	Married	Frequency
Male	35.25%	14.75%	200
Female	33.25%	16.75%	200
Both	68.50%	31.50%	400
Frequency	274	126	

Table 3

Percentage of Students and Working Respondents		
	Original Sample	As Bootstrapped
Student	23.33%	24.00%
Working	76.67%	76.00%
N	60	400

Table 4

Mean and Standard Deviation of the Items for the Original and as Bootstrapped Samples					
Items	Original Sample		95% CI for Mean of Original Sample	As Bootstrapped	
	Mean	Std. Dev.		Mean	Std. Dev.
Item 1	1.7000	0.4583	(1.5840, 1.8160)	1.6600	0.4743
Item 2	1.8833	0.3210	(1.8021, 1.9646)	1.8875	0.3164
Item 3	1.5167	0.4997	(1.3902, 1.6431)	1.5350	0.4994
Item 4	1.9000	0.3000	(1.8241, 1.9759)	1.8850	0.3194
Item 5	1.3500	0.4770	(1.2293, 1.4707)	1.3150	0.4651

As shown in Table 4, the mean responses for all items in the sample of 400 respondents fall inside the 95% confidence intervals for the corresponding means based on the original sample. It is thus safe to use the 400-respondent sample.

Table 5

Percentage of Category 2 (Disagree) Responses (Original Sample)			
	Male	Female	Total
Item 1	63.33%	76.67%	70.00%
Item 2	83.33%	93.33%	88.33%
Item 3	40.00%	63.33%	51.67%
Item 4	86.67%	93.33%	90.00%
Item 5	36.67%	33.33%	35.00%
N	30	30	60

Table 6

Percentage of Category 2 (Disagree) Responses (As Bootstrapped Sample)			
	Male	Female	Total
Item 1	60.00%	72.00%	66.00%
Item 2	82.50%	95.00%	88.75%
Item 3	41.50%	65.50%	53.50%
Item 4	84.50%	92.50%	88.50%
Item 5	33.00%	30.00%	31.50%
N	200	200	400

As shown in Tables 5 and 6, the percentages of category 2 (disagree) responses for all items in the sample of 400 respondents are close to the corresponding percentages for the original sample. We may thus proceed with the analysis using the 400-respondent sample.

The percentages of respondents who did not agree with the items are presented in Table 6. For the male respondents, it can be seen that approximately two-thirds of the male respondents believe that a woman is better suited to raise small children than a man (item 5). Male responses in Items 1, 2 and 4 also reveal that more men subscribe to a pro-women's liberation or pro-gender equality point of view. In Item 3, however, nearly half of the male respondents do not agree that the most natural situation occurs when the man is the breadwinner and the woman runs the household and takes care of the children.

For the female respondents, majority of them disagree to the items except Item 5. The women subscribe to a pro-women's liberation or pro-gender equality point of view but still believe that a woman is better suited to raise small children than a man.

The patterns of responses for male and female respondents appear to be similar, differing only in magnitudes. Most of them do not agree to Items 1, 2 and 4. On the other hand, roughly half of the respondents agree to Item 3, while most of them agree to Item 5. Furthermore, it appears that women are generally more pro-women's lib than are men.

The frequency distribution of the $2^5 = 32$ response patterns is presented below:

Response Patterns With Frequencies

Item 12345	Frequency	Item 12345	Frequency
11111	0	21111	0
11112	0	21112	0
11121	22	21121	10
11122	0	21122	10
11211	0	21211	0
11212	0	21212	0
11221	0	21221	3
11222	0	21222	0
12111	7	22111	6
12112	5	22112	8
12121	6	22121	66
12122	30	22122	16
12211	10	22211	10
12212	0	22212	0
12221	47	22221	87
12222	9	22222	48

Total Number of Possible Response Patterns = 32
Actual Number of Response Patterns with Frequencies = 18

C. COMPARISON OF LATENT CLASS ANALYSIS TO OTHER STATISTICAL METHODS

LCA is strongly related to Cluster Analysis since it is used to discover groups or types of case based on observed data, and possibly, to also assign cases to groups.

LCA is also often called a categorical-data analogue to Factor Analysis. LCA is concerned with the structure of cases (person-centered) while FA is concerned with the structure of variables (variable-centered).

Similarities between LCA and FA :

1) Both are useful for data reduction
2) Latent classes, like factors, are unobserved constructs, inferred from the observed data; and
3) Determining the number of latent classes is analogous in certain respects to the determination of number of factors.

Espeland and Handelman (1988) approached LCA as a mixture of log-linear models. The standard latent class model has been popular among social scientists as a natural extension of the log-linear model in order to take measurement error into account. Haberman (1979) showed that the LCA model can also be specified as a log-linear model for a table with missing cell entries or, more precisely, as a model for the expanded table including the latent variable as an additional dimension.

D. COMMON SOFTWARE FOR LATENT CLASS ANALYSIS

WinLTA –can be used to fit latent class and latent transition models to data.

Models tested using this program may include both static and dynamic latent variables.

Uses the Expectation-Maximization (EM) algorithm to perform parameter estimation

At each iteration, the program computes the convergence index, the Mean Absolute Deviation (MAD). The WinLTA program output provides the Likelihood ratio chi-square, usually denoted by G^2. Also, in WinLTA, the user is given the option of specifying restrictions on any parameter.

PanMark – provides automatic testing of multiple start values, which is important for larger models with, say, more than four latent classes

can automatically check model identifiability and provide asymptotic standard errors for model parameter estimates

also handles longitudinal latent class models

Bootstrapping methods are implemented to help determine the optimal number of latent classes

LEM – can estimate a variety of models in addition to the standard latent class model handles discrete (and continuous) latent trait models

is especially good for estimating local dependence latent class models

can check model identifiability and provide asymptotic standard errors

Latent Gold – a new LCA program

includes optional Bayesian constants to help avoid boundary-value parameter estimates (estimated response probabilities of 0 or 1), diagnostics for detecting conditional dependence

Model identifiability checks and parameter standard errors are provided.

MLLSA – an older software, but still quite serviceable for many applications, and may be obtained for free.
does not provide asymptotic standard errors, but does check model identifiability. Unlike other programs, MLLSA stores all possible unique rating patterns in memory.

MPlus – program for problems with combinations of continuous and categorical variables, for relaxing conditional independence assumptions, and for handling a wider class of mixture problems than ordinary latent class analysis.

II. METHODOLOGY

A. ASSUMPTIONS

1. The population consists of a set of mutually exclusive and exhaustive homogeneous subpopulations, which together make up a latent classification. The subpopulations are homogeneous in a sense that the probability for giving some particular response on a particular item given that one belongs to a latent class is the same for all individuals belonging to that specific latent class.

2. *local independence*: the association between the manifest variables is assumed to depend on the relationship between the manifest and latent variables. All associations among the manifest variables can be explained by the dependence of these manifest variables upon the latent variables. In other words, when the latent variable is held constant, the manifest variables should be statistically independent (i.e., the manifest variables are independent within a class).

B. PROCEDURE

LCA is suitable for binary, nominal, ordered-category and Likert-scale data. It cannot be used, however, with purely ordinal (rank-order) data. But with binary or nominal data, LCA is straightforward, while with ordered-category and Likert-scale data, one may wish to apply certain constraints to response probability parameters.

LCA supposes a simple parametric model and uses observed data to estimate parameter values for the model. The basic concepts and steps are presented as follows:

1) THE LATENT CLASS MODEL

There are two (2) types of probability in the latent class analysis model. The first type of probability indicates the likelihood of a response by respondents in each of the classes. This represents the probability of a particular response to a manifest variable, conditioned on latent class membership. This is used to interpret a latent class in the same manner that factor loadings are used to interpret a latent factor.

The second type indicates the relative frequency or the proportion of the population of interest that are expected to be members of a particular latent class.

To formally present the LCA model, we denote the number of latent classes by *C* and the probability that a subject drawn at random from the population belongs to latent class *c* is written as γ_c .

Since the latent classification is mutually exclusive and exhaustive, $\sum_{c=1}^{C} \gamma_c = 1$. In connection with the first assumption discussed earlier, the subpopulations are homogeneous in the sense that the probability for giving some particular response on a particular item given that one belongs to latent class *c* is the same for all individuals belonging to that specific latent class.

Assume that the subjects are measured on three discrete indicators or items (the extension to fewer than or more than three indicators is direct). Suppose Item 1 has *i* response categories; Item 2 has *j* response categories; and Item 3 has *k* response categories. Let *Y* = {*i*, *j*, *k*} represent a "response pattern", a vector of possible responses to the three items. Using the assumption of local independence, the probability that a randomly chosen subject will come up with response pattern *Y* is given by:

$$P(Y)=\sum_{c=1}^{C}\gamma_c\,\rho_{i|c}\,\rho_{j|c}\,\rho_{k|c}$$

Only two types of parameters appear in the LCA model: the latent probabilities, γ_c which represents the proportion of the population in latent class *c* and the conditional response probabilities,

$$\rho_{i|c},\rho_{j|c} \text{ and } \rho_{k|c}$$

Particularly, $\rho_{i|c}$ represents the probability of response *i* to Item 1, conditional on membership in latent class *c*, $\rho_{j|c}$ represents the probability of response *j* to Item 2, conditional on membership in latent class *c*, $\rho_{k|c}$ represents the probability of response *k* to Item 3, conditional on membership in latent class *c* .

It is useful to distinguish between *exploratory* and *confirmatory latent class analysis*. Exploratory LCA is typically used whenever researchers attempt to identify a set of latent classes from a set of observed measures and no attempt is being made to test a hypothesis regarding the characteristics of the conditional or latent probabilities. The latent class model is said to be *unrestricted* in exploratory latent class analysis, since researchers do not impose a priori constraints on either type of the model's parameters. In confirmatory LCA, researchers, however, wish to test specific hypothesis regarding the values of either the conditional or latent class probabilities. This model is said to be restricted since researchers impose a priori size restrictions on either type of the parameters, or both, depending on the specifics of the hypothesis to be tested.

Confirmatory LCA provides researchers with a powerful method for testing hypotheses regarding the nature of the latent variable. The latent class and conditional probabilities can be restricted in a variety of ways to test different types of hypotheses.

Conditional Probability Restrictions

Specific value constraints on conditional probabilities restrict the maximum likelihood procedure to fitting a latent class model in which one or more of the conditional probabilities has been set to a specified value. Perhaps the most useful type of specific value constraint, however, is the restriction of conditional probabilities to be either 0 or 1.

Another kind is the equality restriction. Unlike specific value restrictions, equality restrictions do not require a priori specification of a value for the conditional probabilities. Instead, two or more of the conditional probability estimates are restricted to be equal to one another, such as when the conditional probability of an observed measure is restricted to be identical for two or more classes. This restriction tests the hypothesis that the observed variable does not discriminate between (or among) specified classes.

<u>Latent Class Probability Restrictions</u>

In the same manner, specific value restrictions on the latent class probabilities specify a priori values for one or more of the latent classes, and equality restrictions on the latent class probabilities restrict two or more of the classes to be of equal size.

2) PARAMETER ESTIMATION

The parameters of the LCA model are typically estimated by means of a Maximization Likelihood (ML) approach. The ML estimates are those most likely to account for the observed results. The log-likelihood function that is maximized is given by

$$\ln \ell = \sum_{i=1}^{I} n_i \ln P(Y = y_i)$$

where we denote by I the total number of cell entries in the frequency table, n_i the observed frequency in cell i and $P(Y=y_i)$ the probability of having the response pattern of cell i

Notice that only non-zero observed cell entries contribute to the log-likelihood function.

The most popular methods for solving the ML estimation problem are the Expectation-Maximization (EM) and Newton-Raphson (NR) algorithms. EM is a very stable iterative method for ML estimation with incomplete data. NR is a faster procedure that, however, needs good starting values to converge. The latter method makes use of the matrix of second-order derivatives of the log-likelihood function, which is also needed for obtaining standard errors of the model parameters.

Parameter estimation in the WinLTA software is performed by means of the EM algorithm. At each iteration of the EM algorithm, the program computes a value known as MAD (Mean Absolute Deviation), which is the mean of the absolute values of the differences between the current value of each parameter being estimated and its value at the previous iteration. MAD is an index of how much progress is being made at each iteration. In the first several iterations, MAD is usually relatively large, because each iteration is making large changes in the parameter estimates. MAD usually becomes steadily smaller with each successive iteration.

3) MODEL SELECTION

LCA models differ mainly in the number of latent classes. Usually, models with more parameters (e.g. more latent classes) provide a better “fit”, and more parsimonious models tend to have a somewhat poorer fit. However, the task is to find the most parsimonious model, the one with the fewest parameters that can describe the associations among a set of observed categorical variable in a theoretically meaningful fashion.

The goodness-of-fit of an estimated LCA model is usually tested by the Pearson or the Likelihood-Ratio Chi-Square (G^2), with the latter defined as

$$G^2 = 2\sum_{i=1}^{I} n_i \ln\left(\frac{n_i}{N * P(Y = y_i)}\right)$$

where N denotes the total sample size, I the total number of cell entries in the frequency table, n_i the observed frequency in cell i and $P(Y=y_i)$ the probability of having the response pattern of cell i

Asymptotically, G^2 is distributed as a chi-square with degrees of freedom (*df*) equals the number of cells in the frequency tables minus one, then minus the number of independent parameters.

When the data are sparse (i.e., there are few or no observations in many cells), the chi-square distribution is usually not a good approximation for the distribution of G^2. Unfortunately, a better approximation for the distribution of G^2 is not known at this time.

It is not valid to compare models with C and $C+1$ classes by subtracting their G^2 and *df* values because this conditional test does not have an asymptotic chi-squared distribution. This means that alternative methods are required for comparing models with different numbers of classes. One popular method is the use of information criteria such as Bayesian Information Criteria (BIC) and Akaike Information Criteria (AIC), both of which corrects G^2 for number of estimated parameters. Generally, models with a lower AIC and BIC are preferred.

Residuals may help the researcher understand why a model does not fit well. A large residual for a particular response pattern indicates that the model does not predict the response pattern well and usually points to adjustments that can be made in the model to improve fit.

4) INTERPRETATION OF RESULTS

Conditional probabilities are used to characterize latent classes in much the same way that factor loadings are used to characterize factors in Factor Analysis. A conditional probability parameter, then, might be seen as the probability that a member of Latent Class 1 answers “yes” to Question 1.

Also, a graphical representation of the best-fitting model in terms of conditional probabilities may be constructed.

C. LIMITATIONS AND ISSUES

Like other contingency table procedures, LCA is not suitable for very small samples. Although there is no particular recommendation on the sample size requirement, since it would depend on the complexity of the model fit, the size of the contingency table and the strength of rho parameters. In general, *it is recommended that LCA be used with great caution, and that it be not used at all, with sample sizes less than about 300.*

One of the problems in the estimation of LCA models is that model parameters may be non-identified, even if the number of degrees of freedom is larger or equal to zero. Non-identification means that different sets of parameter values yield the same maximum of the log-likelihood function or, worded differently, that there is no unique set of parameter estimates. The formal identification check is via the information matrix, which should be positive definite. Another option is to estimate the model of interest with different sets of starting values.

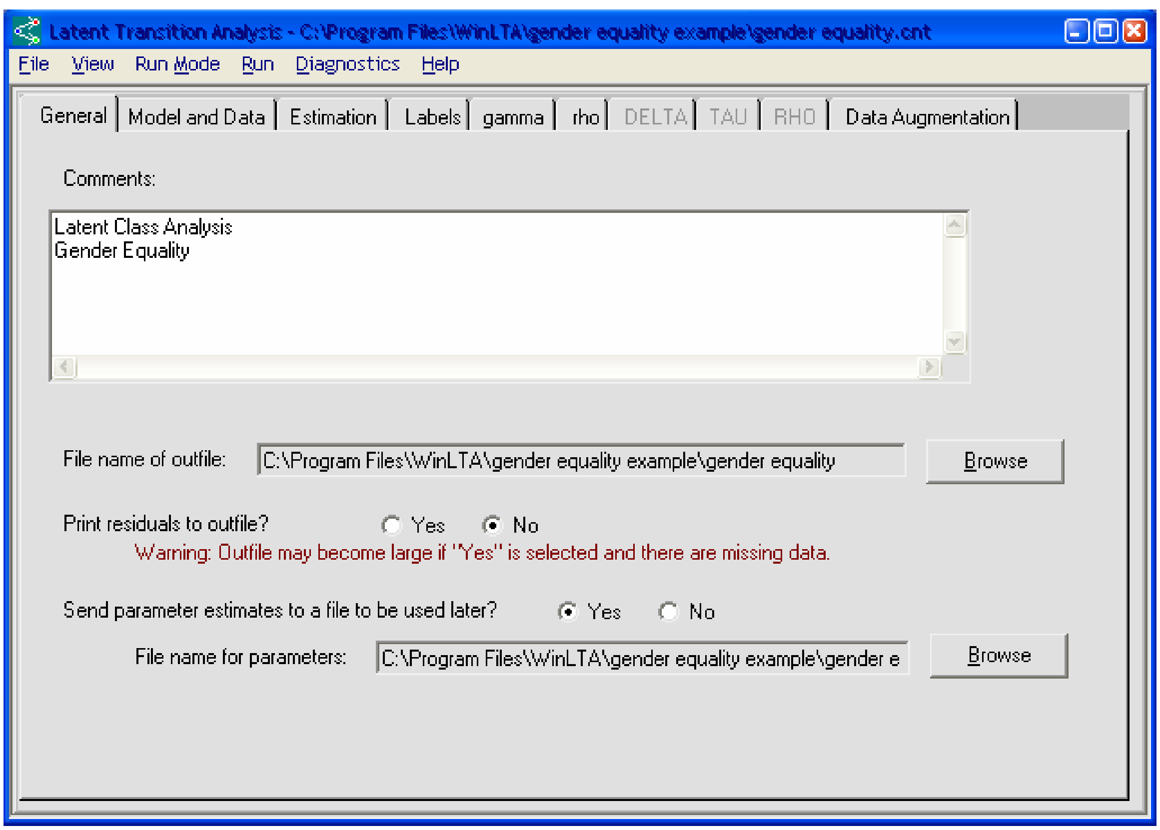
Another problem in LCA modeling is the presence of local maxima. The log-likelihood function of a LCA model is not always concave, which means that hill-climbing algorithms may converge to a different maximum depending on the starting values. Usually, we are looking for the global maximum. The best way to proceed is, therefore, to estimate the model with different sets of random starting values. Typically, several sets converge to the same highest log-likelihood value, which can then be assumed to be the ML solution. Some software packages have automated the use of several sets of random starting values in order to reduce the probability of getting a local solution.

Also, a problem associated with the occurrence of boundary solutions, which are probabilities equal to zero (or one) or log-linear parameters equal to minus (or plus) infinity. These may cause numerical problems in the estimation algorithms, occurrence of local solutions, and complications in the computation of standard errors and number of degrees of freedom of the goodness-of-fit tests. Boundary solutions can be prevented by imposing constraints or by taking into account other kinds of prior information on the model parameters.

III. SOFTWARES FOR LCA

A. THE WINLTA INTERFACE

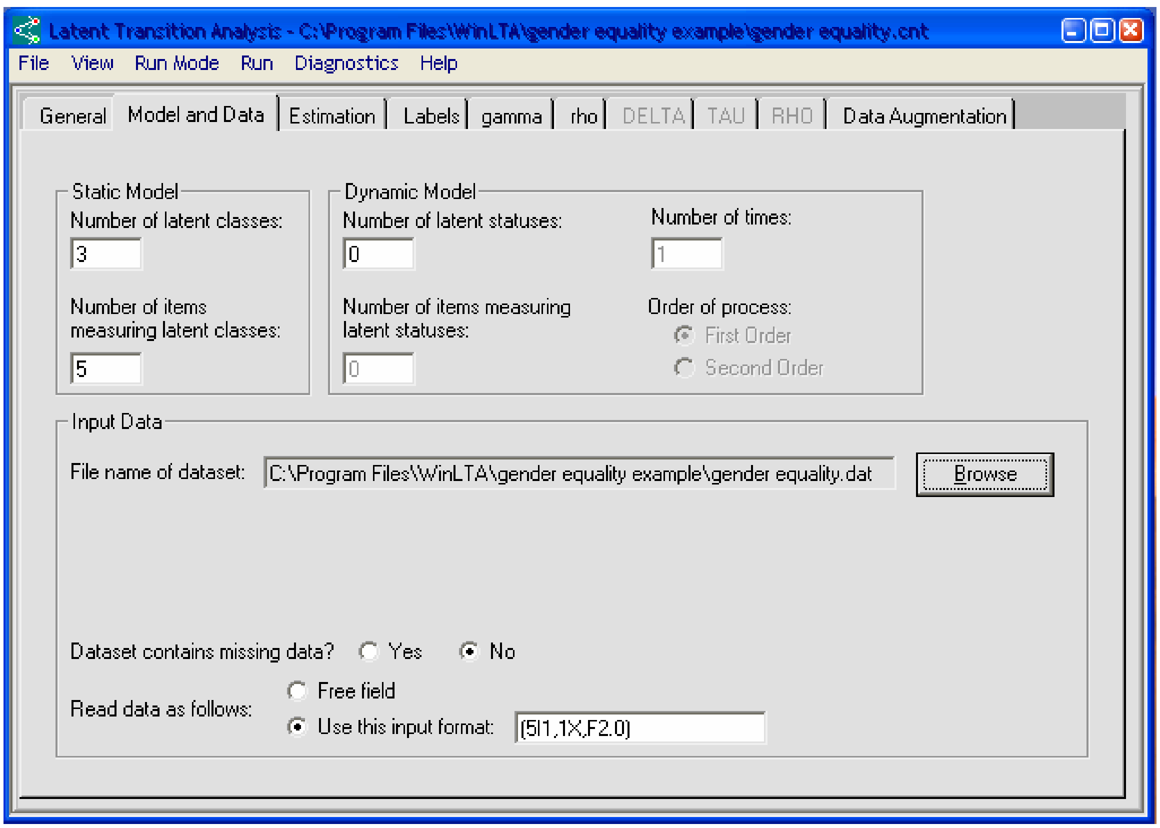
General Tab Window



The field descriptions for the General Tab Window are:

- **Comments:** Enter a title or comments to appear at the top of the output file.
- **File name of outfile:** Specify the file in which the output will be saved by pressing the browse button.
- **Print residuals to outfile?:** Choose whether the residuals will be included in the output file.
- **Send parameter estimates to a file to be used later?:** Choosing "yes" saves the parameter estimates to a separate file specified in the "File name for parameters" field. Saving this file allows you to perform cross validation or to continue a run that fails to converge.

Model and Data Tab Window



The field descriptions for the Model and Data Tab Window are:

- **Static Model:** This section contains information about the static latent variable.
- **Number of latent classes:** In this example, 3 latent classes are chosen.

- **Number of items measuring latent classes:** In this example, there are 5 items.
- **Dynamic Model:** This section contains information about the dynamic latent variable.
- **Number of latent statuses:** Because this example is a latent class problem, this number is 0. For latent class problems, all fields related to the dynamic part of the model are inaccessible.
- **Number of items measuring latent statuses:** 0
- **Number of times:** 1
- **Order of process:** First order.
- **Input Data:**
- **File name of dataset:** Specify the file containing the data by pressing the browse button, which will bring up a dialog box in which you can search for a file (not shown). This dialog box automatically lists .dat as the extension, but you can change this to view all files (*.*), if necessary. The data file must be in response pattern format.
- **Dataset contains missing data?:** If the dataset contains missing data, select yes; otherwise, select no. Remember that missing data must always be coded as 0 in the dataset.
- **Read data as follows:** This option allows you to specify free or fixed formatting in your data file. If each field in the data file (i.e., variables and the frequency) is separated by one or more spaces, then it is not necessary to specify an input format. In this case, choose the free field option. If each variable is not separated by a space, you will need to specify an input format. Use a Fortran input format for this purpose.

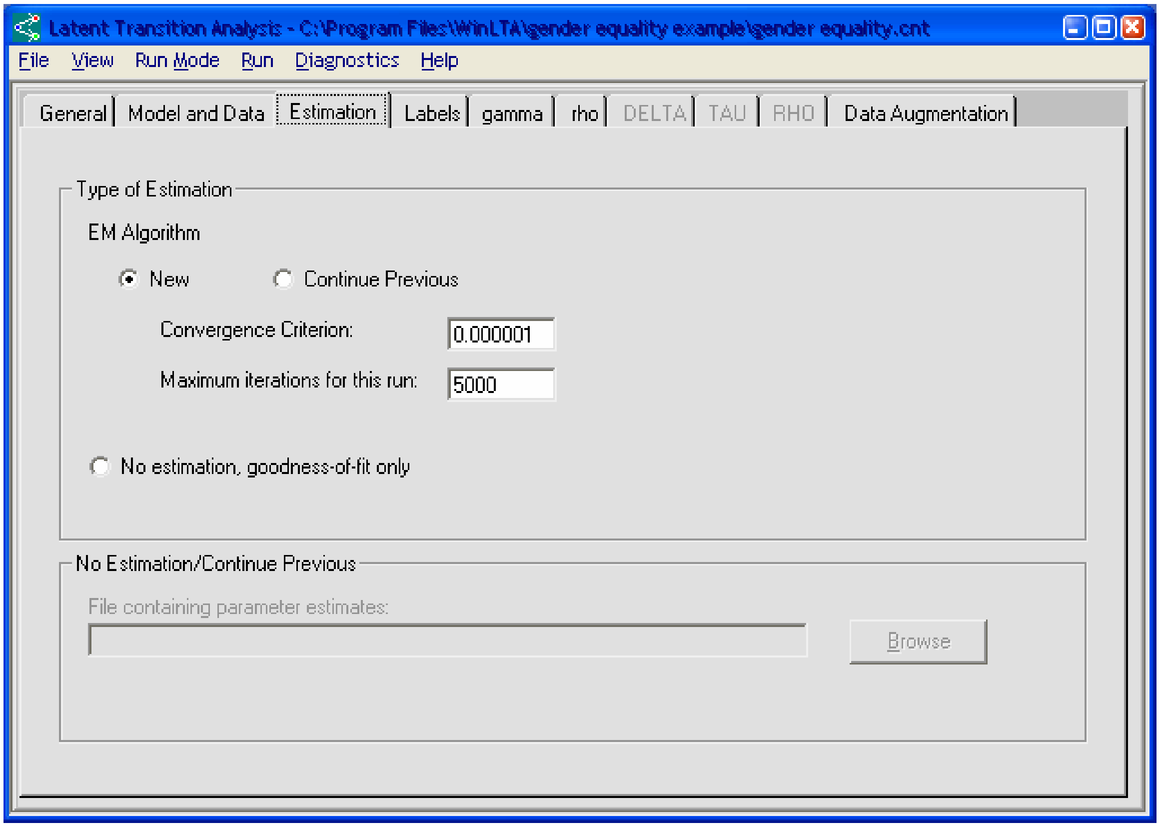
    For example, a dataset is in the following format:

    11121 22
    12111 7
    12112 5
    12121 6
    12122 30
    …

Because there are no spaces between every variable, we must specify an input format. The Fortran input format for this example is (5I1,1X,F2.0). 5I1 indicates that there are five columns with one integer in each column. 1X indicates one column that is blank, or that acts as a space. F2.0 provides information about the column that contains the number of subjects having each response pattern. This is a real number that is 2 digits long and has 0 digits after the decimal place. If you have 1000 or more subjects in a particular response pattern, the format would need to be F4.0.

Note: the part of the format statement corresponding to the variables must be integer type and the frequency count must be a real type. More information about input formats as well as another example can be found in the Help file in WinLTA.

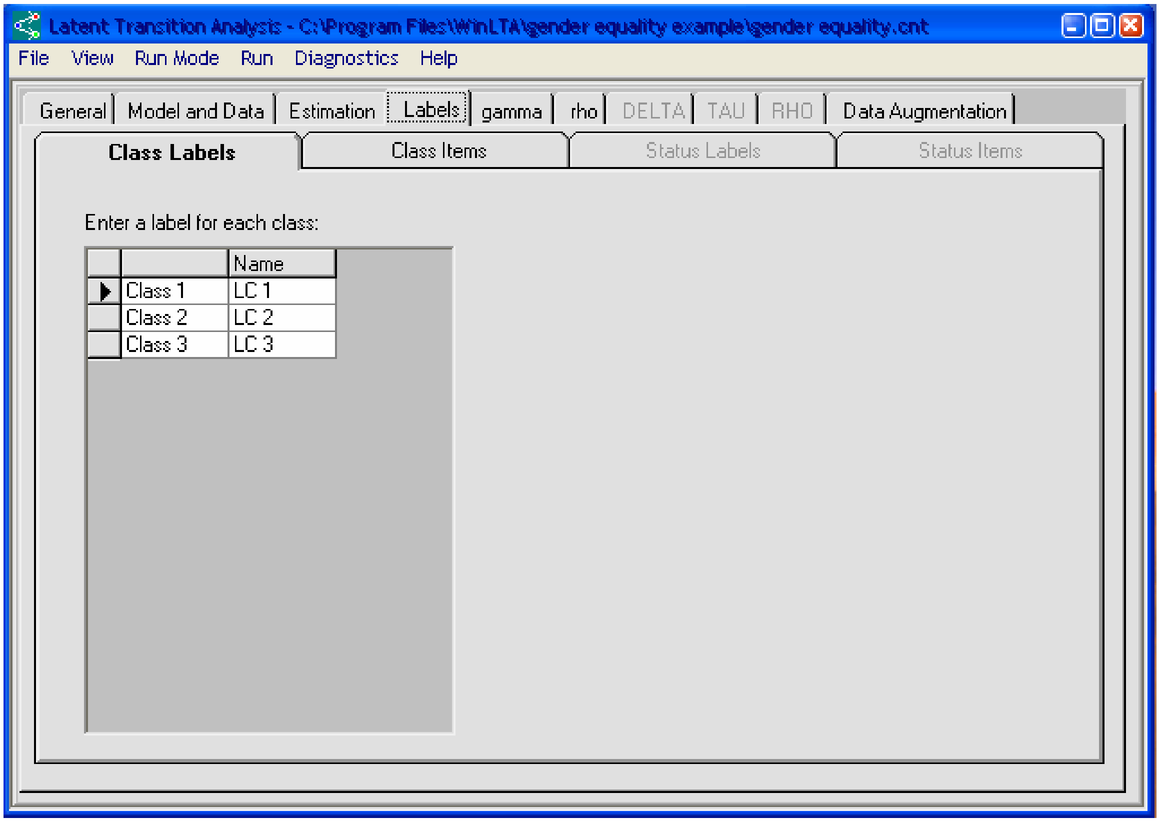
Estimation Tab Window



The field descriptions for the Estimation Tab Window are:

- **Type of estimation:**
- **EM Algorithm:**
- **New/Continue Previous:** For new runs, select New. Select Continue Previous only when you want to continue a run that previously failed to converge.
- **Convergence Criterion:** When the Mean Absolute Deviation (MAD) is less than or equal to the convergence criterion, the program has converged and stops estimation. It is set to the default value of 10-6.
- **Maximum iterations for this run:** If 5000 iterations of the EM algorithm are performed, the program will stop. The default value for this field is 5000.
- **No estimation, goodness-of-fit only:** Selecting this option causes the program to calculate the likelihood ratio statistic only; no estimation is performed.
- **File containing parameter estimates:** If you choose to continue a previous run or if you select no estimation, the name of a file containing parameter estimates must be provided. The Browse button will allow you to search for the file if you do not know the name.

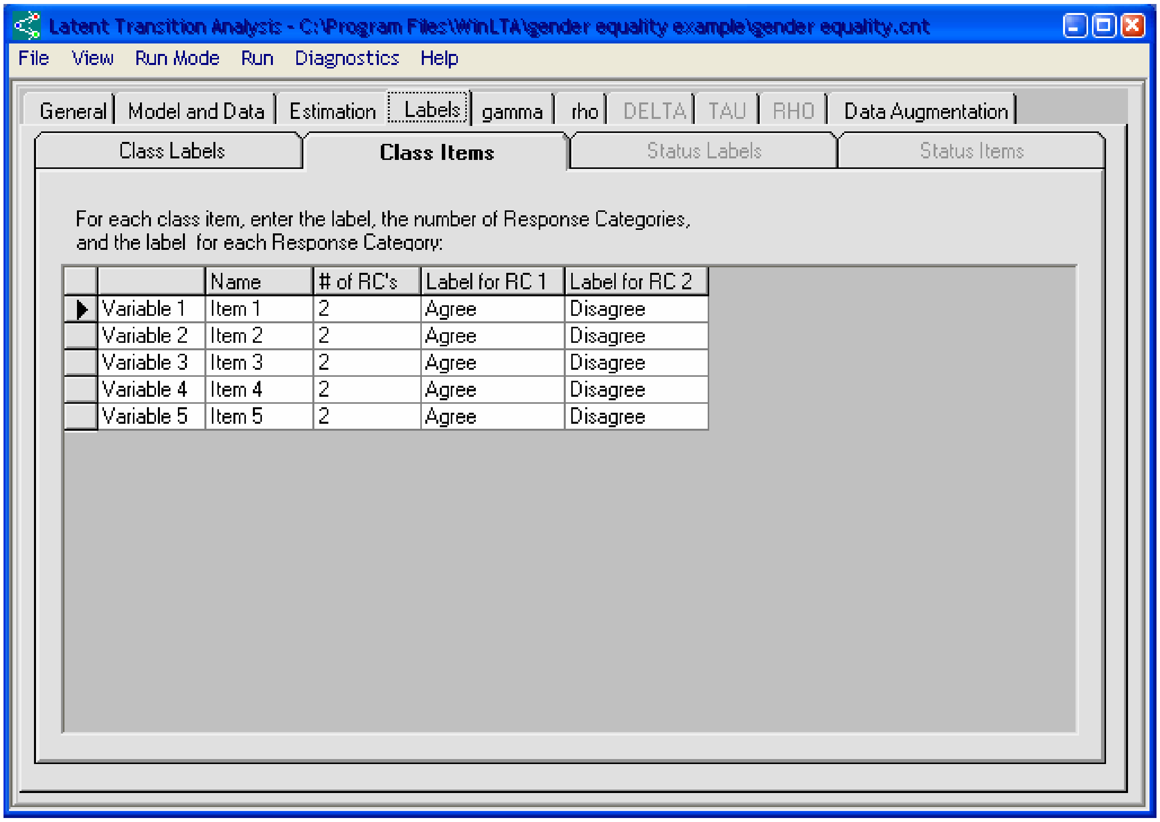
Labels - Class Labels Tab Window



The field descriptions for the Labels - Class Labels Tab Window are:

- **Class labels:** Here you may enter labels for the latent classes. There is a limit of 8 characters per label.

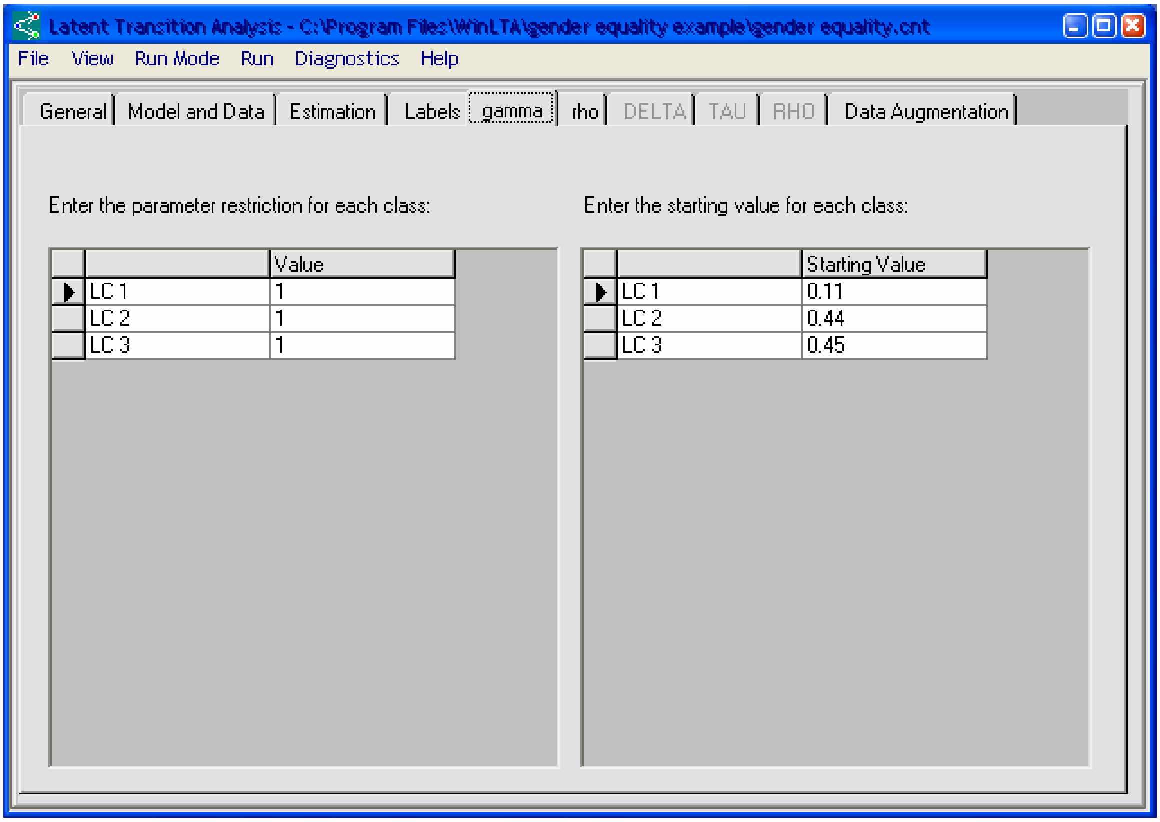
Labels - Class Items Tab Window



The field descriptions for the Labels - Class Items Tab Window are:

- **Class items:** In the Model and Data tab, the number of items measuring the latent classes was entered. For example, there are 5 items. Accordingly, these items are listed as the **5 variables** in the Class Items tab. For each of these items, the following is entered: 1) a **label**, 2) the **number of response categories**, and 3) a **label for each of the response categories**. For example, the first variable has 2 response categories. Response category #1 is named “Agree” and response category #2 is named “Disagree”.

Gamma Tab Window

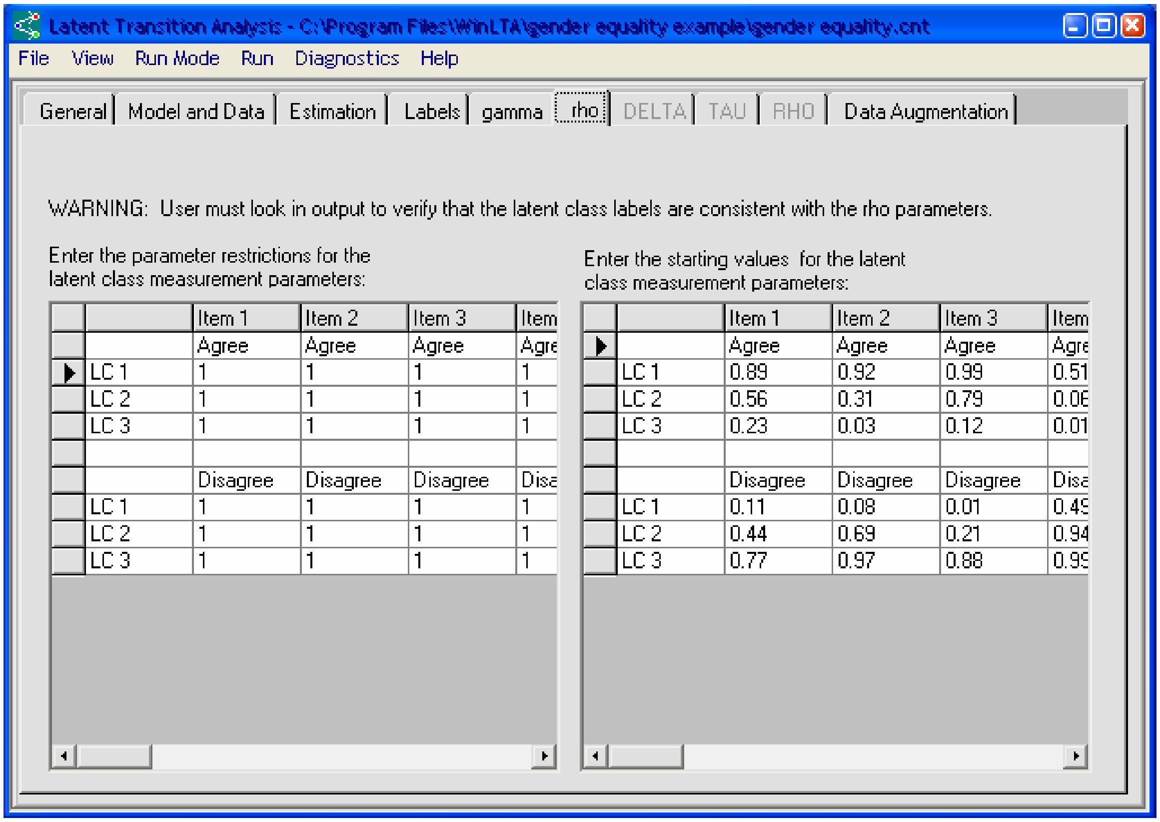


The gamma parameters are the unconditional probabilities of membership in each latent class. Estimating the gamma parameters yields the expected proportion of respondents in each latent class.

The field descriptions for the Gamma Tab Window are:

- **Enter the parameter restrictions for each class:** A parameter can be constrained, estimated freely, or not estimated at all. A "1" denotes that the parameter is to be estimated freely. A "0" denotes that the parameter is not to be estimated (i.e., it will be fixed at the starting value). Integers greater than 1 denote parameters that are constrained to be equal. In this example, all gammas are estimated freely.
- **Enter the starting value for each class:** In this column, the starting values for each of the latent classes are provided. If any equality constraints have been placed on the parameters (i.e., if you entered a 2 or greater in the constraints column), then the starting values should also be equal. For example, if two parameters are constrained to be equal, their starting values should also be equal. Note: the sum of the gamma starting values must be one.

Rho Tab Window



Finally, we need to fill in the rho tab. The little rhos are the probability of responding yes (or no) to an item conditional on latent class membership. For example, the people with a pro-women's lib point of view will have a low probability of agreeing to the 1st item.

Since it is of interest to estimate the rho parameters freely, 1's in the appropriate cells must be entered.

The field descriptions for the Rho Tab Window are:

- **Enter the parameter restrictions for latent class measurement parameters:** In the lefthand side of the Rho Tab Window, enter the parameter restrictions for all response categories.
- **Enter the starting values for latent class measurement parameters:** In the righthand side of the Rho Tab Window, enter the starting values for all response categories.

Saving Your Work and Running WinLTA

- **Saving the control file:** To save the control file, in WinLTA, click on File, and then Save. If this is the first time you have saved this file, a Save As dialog box will appear. In this box you can choose the location of the file and you will be required to enter a filename. By default, the file will be saved with the file extension .cnt. You are encouraged to save your work often during the process of creating a control file.

- **To run WinLTA:** Once the control file is complete, click on Run, and then Run EM. The EM part of WinLTA will begin running automatically. If you have not saved the current version of the control file, you will see a dialog box that asks you if you would like to save before proceeding. Choosing Yes will save the file and run EM automatically. Choosing No will run EM without saving the file. Choosing Cancel will allow you to return to the control file without running EM.

- Once EM begins running, a separate dialog box entitled “LTA EM Run” will appear on your screen and WinLTA will automatically be minimized. This dialog box tells you the status of the EM run as well as the date and time the run began and finished. The box also has two buttons: Abort and Get Info. Pressing the Abort button during the run will cancel the run. Pressing Get Info gives you the iteration number and the MAD at the time shown, although the information is automatically updated every 5 seconds. Once the run is finished (either because the run has converged, you pressed Abort, or the maximum number of iterations has been reached), the Abort button changes to Close. When you press the Close button, WinLTA will automatically be restored as the active window. You can still access WinLTA on your taskbar without pressing the Close button, but Close must be pressed before another EM run can be started.

- **Viewing the WinLTA outfile:** Once the EM run is completed, you can view the output file by clicking on View, and Current LTA Outfile (if you would like to view an older saved outfile, click on Choose LTA Outfile). This will bring up the output file that corresponds with the most recent EM run.

B. THE MICROSOFT EXCEL LATENT CLASS ANALYSIS ADD-IN

How to install the Add-In

1) Unzip the *Latent Class Analysis.zip* file into a directory in your hard disk. This directory should contain the following files:

   LATENT CLASS ANALYSIS.XLA - Excel Add-In file
   LATENT CLASS ANALYSIS.DLL - dynamic link library used
   EXAMPLE.XLS - a sample spreadsheet

2) Move the *LATENT CLASS ANALYSIS.DLL* files to the *Windows/System* directory.
3) Open the Microsoft Excel spreadsheet containing the data to be analyzed, which should be at the same directory mentioned in step 1.

4) Select *Tools* in the menu, and drag the *Add-Ins* option.

5) Click on *Browse*, locate the *LATENT CLASS ANALYSIS.XLA* file and click the OK button. This will automatically install the Add-In.

How to use the Add-In

- Your data should be in Excel spreadsheet, with **one row per respondent**, and **one column per variable**.
- The **first row must contain some sort of description or label for each variable**.
- The **first column must contain some ID or name for each respondent**.
- The answers to each question in the questionnaire must be coded from 1 to (at most) 9.
- **If a variable is nominally scaled, each response category must be used by at least one respondent**. Otherwise, you should recode the variable.
- **Zeros are read as missing data**.
- This demo is **limited to a sample of 9999 respondents and 40 variables**. Speed will reduce substantially as the number of respondents increase.
- To run the analysis, **highlight the range containing the data, making sure to include the first column (respondent ID) and first row (variable labels)**. Then select **Latent Class Analysis** from the **Tools** menu.

IV. BACK TO ILLUSTRATIVE EXAMPLE

The Questionnaire

Name (optional): ______________________ Age:___
Gender:______
Occupation: __________________________ Civil Status: ___________
Do you have children? Yes ___ No ___

Statements	Strongly Agree	Agree	Disagree	Strongly Disagree
1. Women's liberation sets women against men				
2. It's better for a wife not to have a job because that always poses problems in the household, especially if there are children				
3. The most natural situation occurs when the man is the breadwinner and the woman runs the household and takes care of the children				
4. It isn't really as important for a girl to get a good education as it is for a boy				
5. A woman is better suited to raise small children than man.				

Thank you very much! ☺

A. MODEL SELECTION

The Microsoft Excel Latent Class Analysis Add-In Output

➢ Two Classes

	A	B	C	D	E	F
1	Marketing Research Tools		Results from the Latent Class Analysis			
2	Demo under development by	LogLikelihood=	-1035.0	Overlap=	35.0%	
3	Wagner A. Kamakura (kamakura@duke.edu)	CAIC=	2135.9			
4			Response Probabilities within Classes			
5		Class Sizes=>	31.2%	68.8%		
6	Item	Response	Class 1	Class 2	Class 3	Class 4
7	Statement 1 (1=Agree, 2=Disagree)	1	46.3%	28.4%		
8		2	53.7%	71.6%		
9						
10	Statement 2 (1=Agree, 2=Disagree)	1	33.0%	1.4%		
11		2	67.0%	98.6%		
12						
13	Statement 3 (1=Agree, 2=Disagree)	1	99.9%	22.3%		
14		2	0.1%	77.7%		
15						
16	Statement 4 (1=Agree, 2=Disagree)	1	14.4%	10.2%		
17		2	85.6%	89.8%		
18						
19	Statement 5 (1=Agree, 2=Disagree)	1	57.9%	73.3%		
20		2	42.1%	26.7%		

➢ Three Classes

	A	B	C	D	E	F
1	Marketing Research Tools		Results from the Latent Class Analysis			
2	Demo under development by	LogLikelihood=	-1014.7	Overlap=	8.7%	
3	Wagner A. Kamakura (kamakura@duke.edu)	CAIC=	2131.3			
4			Response Probabilities within Classes			
5		Class Sizes=>	10.2%	8.4%	81.4%	
6	Item	Response	Class 1	Class 2	Class 3	Class 4
7	Statement 1 (1=Agree, 2=Disagree)	1	53.3%	81.1%	26.7%	
8		2	46.7%	18.9%	73.3%	
9						
10	Statement 2 (1=Agree, 2=Disagree)	1	100.0%	0.0%	1.3%	
11		2	0.0%	100.0%	98.7%	
12						
13	Statement 3 (1=Agree, 2=Disagree)	1	100.0%	100.0%	34.3%	
14		2	0.0%	0.0%	65.7%	
15						
16	Statement 4 (1=Agree, 2=Disagree)	1	0.0%	21.6%	11.9%	
17		2	100.0%	78.4%	88.1%	
18						
19	Statement 5 (1=Agree, 2=Disagree)	1	76.7%	0.2%	74.5%	
20		2	23.3%	99.8%	25.5%	

➢ Four Classes

	A	B	C	D	E	F
1	Marketing Research Tools		Results from the Latent Class Analysis			
2	Demo under development by	LogLikelihood=	-1001.3	Overlap=	8.1%	
3	Wagner A. Kamakura (kamakura@duke.edu)	CAIC=	2140.4			
4			Response Probabilities within Classes			
5		Class Sizes=>	14.9%	10.1%	17.1%	58.0%
6	Item	Response	Class 1	Class 2	Class 3	Class 4
7	Statement 1 (1=Agree, 2=Disagree)	1	80.5%	54.7%	96.7%	0.0%
8		2	19.5%	45.3%	3.3%	100.0%
9						
10	Statement 2 (1=Agree, 2=Disagree)	1	0.0%	100.0%	0.0%	2.1%
11		2	100.0%	0.0%	100.0%	97.9%
12						
13	Statement 3 (1=Agree, 2=Disagree)	1	100.0%	100.0%	0.0%	37.2%
14		2	0.0%	0.0%	100.0%	62.8%
15						
16	Statement 4 (1=Agree, 2=Disagree)	1	29.5%	0.0%	15.2%	7.8%
17		2	70.5%	100.0%	84.8%	92.2%
18						
19	Statement 5 (1=Agree, 2=Disagree)	1	27.8%	77.6%	86.3%	72.1%
20		2	72.2%	22.4%	13.7%	27.9%

Items Found in the MS Excel LCA Add-In Output

- **Overlap** - Degree of overlap in class membership. The lower the overlap, the surer you are about who belongs to which class

- **CAIC** (Corrected Akaike Information Criteria) - Fit criterion to decide which solution is best. The best solution is the one with the lowest CAIC

- **Class Size** - these are the gamma parameter estimates

- **Response Probabilities within Classes** - these are the little rho parameter estimates

The CAICs computed by the Microsoft Excel add-in program, Latent Class Analysis, are used to select the best-fitting model.

Model	AIC
2 classes	2135.9
3 classes	**2131.3**
4 classes	2140.4

Since models with a lower AIC are preferred, an LCA model with 3 classes was used in proceeding with the estimation of parameters using the WinLTA software.

B. THE WINLTA OUTPUT

1

```
PROGRAM VERSION:                          3.1.0 REL
PROGRAM STARTED:       Sun Feb 26 04:33:21 2006

* Latent Class Analysis
* Gender Equality

****************************************************************
```

2

```
INFORMATION ABOUT THIS JOB:
RUN TYPE: PARAMETER ESTIMATION BY EM

CONTROL DATA READ FROM FILE:
C:\Program Files\WinLTA\gender equality example\gender
equality.cnt
DATA ANALYZED IN THIS RUN READ FROM FILE:
C:\Program Files\WinLTA\gender equality example\gender
equality.dat
OUTPUT SAVED IN FILE:
C:\Program Files\WinLTA\gender equality example\gender equality
PARAMETER ESTIMATES SAVED IN FILE:
C:\Program Files\WinLTA\gender equality example\gender equality
parameters

 STATIC LATENT VARIABLE                        YES
 NUMBER OF LATENT CLASSES                        3
 NUMBER OF MANIFEST ITEMS                        5
 DYNAMIC LATENT VARIABLE                        NO
 NUMBER OF SUBJECTS                            400
 NUMBER OF UNIQUE RESPONSE PATTERNS             18
 MAXIMUM NUBMER OF ITERATIONS                 5000
 CONVERGENCE CRITERION          .00000100000000
 MISSING DATA IN RESPONSE PATTERNS              NO
 PRINT RESIDUALS                               YES

****************************************************************
```

3

```
THE FOLLOWING PARAMETER RESTRICTIONS HAVE BEEN SPECIFIED
WHERE 0=FIXED TO START VALUE
       1=FREE
       2 OR GREATER MEANS CONSTRAINED EQUAL TO ANY OTHER
         PARAMETER WITH THE SAME DESIGNATION

****************************************************************
```

4

```
LITTLE RHO PARAMETERS
LITTLE RHOS ARE PROBABILITIES OF RESPONSES
TO ITEMS MEASURING THE STATIC LATENT VARIABLE
```

```
CONDITIONAL ON LATENT CLASS MEMBERSHIP

                   RESPONSE CATEGORY 1
         | I A | I A | I A | I A | I A |
         | t g | t g | t g | t g | t g |
         | e r | e r | e r | e r | e r |
         | m e | m e | m e | m e | m e |
         |   e |   e |   e |   e |   e |
         | 1   | 2   | 3   | 4   | 5   |
 LC 1      1     1     1     1     1
 LC 2      1     1     1     1     1
 LC 3      1     1     1     1     1

                  RESPONSE CATEGORY 2
         | I D | I D | I D | I D | I D |
         | t i | t i | t i | t i | t i |
         | e s | e s | e s | e s | e s |
         | m a | m a | m a | m a | m a |
         |   g |   g |   g |   g |   g |
         | 1 r | 2 r | 3 r | 4 r | 5 r |
         |   e |   e |   e |   e |   e |
         |   e |   e |   e |   e |   e |
 LC 1      1     1     1     1     1
 LC 2      1     1     1     1     1
 LC 3      1     1     1     1     1

****************************************************************
```

5

```
GAMMA PARAMETER RESTRICTIONS
GAMMAS ARE UNCONDITIONAL PROBABILITIES OF MEMBERSHIP
IN EACH LATENT CLASS OF THE STATIC LATENT VARIABLE

 LC 1        1
 LC 2        1
 LC 3        1

****************************************************************
```

6

```
          START VALUES

LITTLE RHO PARAMETERS
LITTLE RHOS ARE PROBABILITIES OF RESPONSES
TO ITEMS MEASURING THE STATIC LATENT VARIABLE
CONDITIONAL ON LATENT CLASS MEMBERSHIP
```

7

```
                RESPONSE CATEGORY 1

            | I A | I A | I A | I A | I A |
            | t g | t g | t g | t g | t g |
            | e r | e r | e r | e r | e r |
            | m e | m e | m e | m e | m e |
            |   e |   e |   e |   e |   e |
            | 1   | 2   | 3   | 4   | 5   |
 LC 1        0.890 0.920 0.990 0.510 0.990
 LC 2        0.560 0.310 0.790 0.060 0.740
 LC 3        0.230 0.030 0.120 0.010 0.240
```

8

```
                 RESPONSE CATEGORY 2

            | I D | I D | I D | I D | I D |
            | t i | t i | t i | t i | t i |
            | e s | e s | e s | e s | e s |
            | m a | m a | m a | m a | m a |
            |   g |   g |   g |   g |   g |
            | 1 r | 2 r | 3 r | 4 r | 5 r |
            |   e |   e |   e |   e |   e |
            |   e |   e |   e |   e |   e |
 LC 1        0.110 0.080 0.010 0.490 0.010
 LC 2        0.440 0.690 0.210 0.940 0.260
 LC 3        0.770 0.970 0.880 0.990 0.760

****************************************************************
```

9

```
GAMMA PARAMETERS
GAMMAS ARE UNCONDITIONAL PROBABILITIES OF MEMBERSHIP
IN EACH LATENT CLASS OF THE STATIC LATENT VARIABLE

 LC 1      0.110
 LC 2      0.440
 LC 3      0.450

****************************************************************
```

10

```
 ITERATION HISTORY
 STARTING G-SQUARED=   411.207

 ITER-   MAD           ITER-   MAD           ITER-   MAD
 ATION                 ATION                 ATION
    1  .0924163693        2  .0312044413        3  .0170738515
    4  .0099477047        5  .0075775440        6  .0063670694
    7  .0053380002        8  .0043271610        9  .0033869819
   10  .0026016930       11  .0019629562       12  .0016255540
   13  .0014225762       14  .0012880404       15  .0012126485
```

```
  334  .0000044982    335  .0000042209    336  .0000039607
  337  .0000037164    338  .0000034871    339  .0000032720
  340  .0000030700    341  .0000028805    342  .0000027027
  343  .0000025358    344  .0000023792    345  .0000022322
  346  .0000020943    347  .0000019649    348  .0000018435
  349  .0000017296    350  .0000016227    351  .0000015224
  352  .0000014283    353  .0000013400    354  .0000012571
  355  .0000011794    356  .0000011065    357  .0000010380
  358  .0000009738

***************************************************************
```

11

```
MODEL FIT

G-Squared Test of Model Fit:                    85.038
Degrees of Freedom:                                 14

***************************************************************

**WARNING**: BE SURE TO INTERPRET THE LATENT CLASSES CAREFULLY
             BASED ON THE ESTIMATED RHO PARAMETERS REPORTED
             BELOW.
             YOU MAY WISH TO CHANGE THE LABELS YOU PREVIOUSLY
             ASSIGNED TO THE LATENT CLASSES IN ORDER TO MAKE
             THEM CONSISTENT WITH YOUR INTERPRETATION.

***************************************************************
LITTLE RHO PARAMETERS
LITTLE RHOS ARE PROBABILITIES OF RESPONSES
TO ITEMS MEASURING THE STATIC LATENT VARIABLE
CONDITIONAL ON LATENT CLASS MEMBERSHIP
```

12

```
              RESPONSE CATEGORY 1

         | I A | I A | I A | I A | I A |
         | t g | t g | t g | t g | t g |
         | e r | e r | e r | e r | e r |
         | m e | m e | m e | m e | m e |
         |   e |   e |   e |   e |   e |
         | 1   | 2   | 3   | 4   | 5   |
 LC 1     0.702 1.000 1.000 0.000 1.000
 LC 2     0.549 0.196 1.000 0.188 0.000
 LC 3     0.274 0.015 0.335 0.115 0.756
```

13

```
              RESPONSE CATEGORY 2

          | I D | I D | I D | I D | I D |
          | t i | t i | t i | t i | t i |
          | e s | e s | e s | e s | e s |
          | m a | m a | m a | m a | m a |
          |   g |   g |   g |   g |   g |
          | 1 r | 2 r | 3 r | 4 r | 5 r |
          |   e |   e |   e |   e |   e |
          |   e |   e |   e |   e |   e |
 LC 1      0.298 0.000 0.000 1.000 0.000
 LC 2      0.451 0.804 0.000 0.812 1.000
 LC 3      0.726 0.985 0.665 0.885 0.244
```

****

14

GAMMA PARAMETERS
GAMMAS ARE UNCONDITIONAL PROBABILITIES OF MEMBERSHIP IN EACH
LATENT CLASS OF THE STATIC LATENT VARIABLE

LC 1	0.077
LC 2	0.119
LC 3	0.804

****

15

EXPECTED CELL FREQUENCIES AND RESIDUALS

	OBS	EXP	RESID	PEARSON	
12111	7.0000	2.5343	4.4657	2.8052	*
22111	6.0000	6.6996	-0.6996	-0.2703	
12211	10.0000	5.0397	4.9603	2.2095	*
22211	10.0000	13.3248	-3.3248	-0.9108	
11121	22.0000	22.0000	0.0000	0.0000	
21121	10.0000	10.0000	0.0000	0.0000	
12121	6.0000	19.4548	-13.4548	-3.0505	*
22121	66.0000	51.4338	14.5662	2.0311	*
21221	3.0000	1.5298	1.4702	1.1887	
12221	47.0000	38.6920	8.3080	1.3356	
22221	87.0000	102.2998	-15.2998	-1.5127	
12112	5.0000	4.7586	0.2414	0.1107	
22112	8.0000	5.4076	2.5924	1.1148	
21122	10.0000	3.6646	6.3354	3.3095	*
12122	30.0000	23.2676	6.7324	1.3957	
22122	16.0000	30.5968	-14.5968	-2.6389	*
12222	9.0000	12.5020	-3.5020	-0.9904	
22222	48.0000	33.0548	14.9452	2.5995	*

****

```
PROGRAM FINISHED: Sun Feb 26 04:33:21 2006
ELAPSED TIME:    0 HOURS,    0 MINUTES,    0 SECONDS.
```

Explanation of the WinLTA Output File for the Example

The first sections of the output contain a listing of the parameter restrictions, starting values, and other information entered in the program control file. The next section is the iteration history, which is followed by the parameter estimates. In this example, residuals were requested, so these are the final section of the output.

Program Control File Information

1 The title lines and comments entered in the General tab will be printed first.

2 Basic information from the Model and Data tab and the Estimation tab is echoed back in the first section of the output file. The filenames for the control file, the data, and the output are shown first. The next lines include the number of latent classes, number of latent statuses, number of items (for statuses and classes), number of participants, number of observed response patterns, maximum number of iterations allowed, convergence criterion, and whether there are missing data in the response patterns. Finally, there is a line stating whether or not residuals will be printed.

3 This section contains a listing of the user-specified parameter restrictions.

4 The little rho parameter, sectioned by response categories, are estimated freely (restriction = 1 for free). In the example provided, two response categories are represented (agree = 1, disagree = 2).

5 In this section, the user-specified parameter restrictions for the gamma parameters are printed. In this case, since gamma parameters are estimated freely (restriction = 1 for free).

6 This section contains a listing of the user-specified starting values. For this example, we used the parameter estimates of the Dutch study (See Ton Heinen, 1996) as starting values.

7 In this section, the user-specified starting values for the little rho parameters in response category 1 are printed.

8 In this section, the user-specified starting values for the little rho parameters in response category 2 are printed.

9 In this section, the user-specified starting values for the gamma parameters are printed.

10 After each iteration, the iteration number and the Mean Absolute Deviation (MAD) are printed. The maximum iterations for the run as specified in the Estimation tab determine the maximum number of iterations allowed. MAD is the mean absolute difference between the parameter estimates resulting from the current iteration and the parameter estimates from the previous iteration.

11 The final G^2 reflects the fit of the model, with degrees of freedom equal to the number of possible response patterns minus the number of parameters estimated minus one. For this example, we have 5 items with 2 response categories for each, thus we have 32 possible response patterns; 15 little rho parameters and 2 gamma parameters to be estimated. Thus the degrees of freedom is 32 – 17 – 1 = 14. The value of the goodness-of-fit statistic is compared to a chi-squared distribution. In this case, the observed test statistic, G^2 = 85.038, is much greater than the α = .05 critical value of the chi-squared distribution, χ^2(14) = 23.68. Therefore, we conclude that the model does not fit the data very well. This could be due to having only 18 unique response patterns out of the possible 32. One approach to obtaining a better fitting model would be to try a completely different set of parameter restrictions.

The estimates for the parameters are printed in the next sections of output.

12 The little rho parameters are the probabilities of a particular item response conditional on latent class membership. The probabilities are grouped by response category. This section is for the first response category. For example, the probability of agreeing to the 1st item given membership in the 1st latent class (LC1) is 0.702, and so on.

13 The estimates for the little rho parameters for the second response category are given in this section. In this example, a response of "2" designates "disagree". For example, the probability of not agreeing to the 1st item given membership in the 1st latent class (LC1) is 0.298, and so on.

In factor analysis, the patterns of factor loadings characterize the factors. In latent class analysis, the conditional probabilities characterize the latent classes. For the first latent class, the probabilities of agreeing to items 1, 2, 3 and 5 are very high. Therefore, this latent class is characterized as the ones in the latent class of people who have more traditional positions concerning male and female role patterns but still value the importance of education for both male and female. For the second latent class, the probabilities of agreeing to the 3rd and 5th items are 1.0 and 0.0, respectively. High probabilities of not agreeing in items 2 and 4 are also notable. This latent class may seem to take on a strong position towards gender equality through education and career but still believes firmly that the most natural situation occurs when the man is the breadwinner and the woman runs the household and takes care of the children. This position may be caused by the respondents' own personal experiences. For the last latent class, the probabilities of not agreeing in items 1, 2, 3 and 4 and the probability of agreeing in item 5 are very high. This latent class takes a pro-women's lib point of view or push for gender equality but still acknowledge the role of a woman in child rearing.

14 The estimates for the gamma parameters are the unconditional probabilities of latent class membership. These probabilities mean that according to this model, 7.7% of the respondents belongs to a class of people who have more traditional positions concerning male and female role patterns but still value the importance of education for both male and female, 11.9% of the respondents have partly traditional and liberated view on equality of male and female role patterns, and 80.4% or majority of the respondents is the class of people who take a pro-women's lib point of view or push for gender equality but still acknowledge the role of a woman in child rearing. This position could also be prevalent to most of the respondents because of the influence of Filipino culture that women are more involved in nurturing the children.

15 *RESID* is the raw residual associated with each response pattern. It is the difference between the observed and expected cell frequencies. The *PEARSON* or the standardized residual is obtained by dividing the raw residual by the square root of the expected frequency. Thus, for the response pattern of disagreeing to all of the items, the expected frequency is 33, or the probability of obtaining such response pattern from a randomly chosen subject is 33.0548 / 400 = 0.0826.

R E F E R E N C E S

WEBSITE

http://astor.som.jhmi.edu/~esg/statsforps1_2004.html
http://www.statmodel.com/
www.gla.ac.uk/External/RSS/RSSComp/vermunt.pdf
www.ipn.uni_kiel.de/aktuell/buecher/rostbuch/c19.pdf
www.spitswww.uut.nl/~vermunt
http://utcmhsrc.csw.utk.edu/caseyproject/presentations
http://ourworld.compuserve.com/homepages/jsuebersax/faq.htm
www.ats.ucla.edu/stat/mplus/seminars/lca/lca/ppt
John Uebersax (2000). *A Practical Guide to Local Dependence in Latent Class Models*. http://ourworld.compuserve.com/homepages/jsuebersax/condep.htm

BOOKS

McCutcheon, Alan. L. (1987) *Latent Class Analysis*. Newbury Park, Ca: Sage Publications

Heinen, Ton (1996) *Latent Class and Discrete Latent Trait Models, Similarities and Differences*. Sage Publications