PENNSTATE

The Methodology Center

# PROC LCA & PROC LTA Users' Guide Version 1.3.2

Stephanie T. Lanza
John J. Dziak
Liying Huang
Aaron Wagner
Linda M. Collins



Please send questions and comments to *MChelpdesk@psu.edu*.
The development of PROC LCA and PROC LTA was supported by the National Institute on Drug Abuse Grant P50-DA10075 to The Center for Prevention and Treatment Methodology. We thank Bethany Bray and Amanda Applegate for their helpful comments and suggestions. We also thank Joseph L. Schafer and David Lemmon who contributed heavily to the development of this document and software.

The suggested citation for this users' guide is
Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A. T., & Collins, L. M. (2015). *Proc LCA & Proc LTA users' guide* (Version 1.3.2). University Park: The Methodology Center, Penn State. Available from methodology.psu.edu.

# Table of Contents

# 1 Overview of Procedures and Recent Improvements

PROC LCA and PROC LTA are two related applications that are distributed together. They were developed for SAS® for Windows. They are designed for the SAS® software package for Windows (version 9.1 or higher).[1]

**The first procedure, PROC LCA, is a SAS procedure for latent class analysis (LCA).** This procedure can be used to estimate latent classes that are measured by categorical indicators. Key features of PROC LCA include

- Multiple-groups LCA
- Option to impose measurement invariance across groups
- LCA with covariates (prediction of latent class membership)
- Binary and multinomial logistic regression options for predicting latent class membership
- The ability to take into account sampling weights and clusters.

**The second procedure, PROC LTA, is a SAS procedure for latent transition analysis (LTA)**, in which the latent variable is dynamic and indicators are measured in a longitudinal panel design. The term "latent status" is used to refer to latent classes that are measured longitudinally. Key features of PROC LTA include

- Multiple-groups LTA
- Two or more measurement occasions (times)
- Change over time reflected in transition probabilities
- Option to impose measurement invariance across groups and/or times
- LTA with covariates (prediction of latent status membership and transitions)
- Separate sets of covariates may be specified for Time 1 and for each transition (Time 1 to Time 2, Time 2 to Time 3, etc.)
- Binary and multinomial logistic regression options for predicting latent status membership at Time 1 and modeling transition probabilities.

**Both PROC LCA and PROC LTA have the following key features:**

- Option for automatic starting values
- Option for applying data-derived prior in order to stabilize logistic regression (BETA PRIOR)
- Posterior probabilities can be saved to a SAS data file
- Parameter estimates can be saved to a SAS data file
- Input data can be in aggregated (response-pattern data) form or one record per case

This guide assumes the user has a working knowledge of LCA and LTA. An introduction to these models can be found in Lanza, Bray, and Collins (2013) and Collins and Lanza (2010). A detailed empirical demonstration of PROC LCA appears in Lanza, Collins, Lemmon, and Schafer (2007), and an empirical demonstration of PROC LTA appears in Lanza and Collins (2008).

**Important changes from version 1.3.1**

[1] SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

- Addition of OUTCOVB output option for PROC LCA. This is for use with the LCA Distal Outcome macro and does not affect other uses of the procedure.

**Important changes from version 1.3.0**

- Bug fixes

**Important changes from version 1.2.7**

- NSTARTS is now available for use in models with covariates.
- The "best" column has been added to the OUTPOST file. This column indicates which latent class is the best match for each individual based on posterior probabilities (i.e., maximum-probability assignment).
- The new SEED_DRAWS statement allows users to generate 20 random simulations for each individual's potential class membership based on posterior probabilities (i.e., pseudo-class draws) and save them to the OUTPOST file.

**Important changes from version 1.2.6:**

- Minor bug fixes, to address a memory handling limitation which occurred for certain models.

**Important changes from version 1.2.5:**

- LCA models can now incorporate some complex survey sample features (clustering and weights) taking a pseudo-maximum-likelihood approach (Vermunt & Magidson 2005a, 2005b).
- Covariates can now be included in models even when the data are in aggregated form.

**Important changes from version 1.2.4:**

- A bug involving the computation of the CAIC fit statistic was corrected.

**Important changes from version 1.2.3:**

- When a $\rho$ prior is applied to a model with covariates, it is now also applied automatically to the null model used to test the significance of each covariate.
- The output format has been improved to clarify the optimal solution when the NSTARTS command is used.
- Messaging in the output window has been improved when priors are applied to $\beta$, $\gamma$, or $\rho$ parameters.

**Important changes from version 1.1.5:**

- Multiple random starts are now available via the NSTARTS command, to help assess and avoid suboptimal estimates caused by local maxima of the likelihood function. The starting seeds and associated likelihoods can be exported to a dataset using the OUTSEEDS option, to help in assessing whether the maximum of the likelihood has been identified.
- Standard errors for parameter estimates are now provided where possible given the parameter estimates. They can be saved to a dataset using the OUTSTDERR option.
- In place of the older STABILIZE command, the new version of PROC LCA now offers an expanded set of stabilizing options based on data-derived priors: BETA PRIOR=, GAMMA PRIOR=, and RHO PRIOR=. The user can stabilize one or more kinds of

parameters, and also choose the strength of the prior. These options are explained further in the LCA Syntax section.

- For users with multiple-core personal computers, faster computation using parallel cores is now available via a CORES statement.

# 2 The LCA Mathematical Model

Up to three sets of parameters are provided in PROC LCA output:

- Gamma ($\gamma$) parameters: latent class membership probabilities
- Rho ($\rho$) parameters: item-response probabilities conditional on latent class membership
- Beta ($\beta$) parameters: logistic regression coefficients for covariates, predicting class membership

The $\rho$ parameters express the correspondence between the observed items and the latent classes, and form the basis for interpretation of the latent classes. When no covariates are included, only $\rho$ and $\gamma$ parameters are estimated. When covariates are included, only $\rho$ and $\beta$ parameters are estimated; in this case, the $\gamma$ parameters are calculated as functions of $\beta$ parameters and the covariates, and are provided in PROC LCA output. If a grouping variable is included, all sets of parameters ($\gamma$, $\rho$, $\beta$) can be conditioned on group.

Suppose we estimate a latent class model with $n_c$ classes from a set of $M$ categorical items and include a covariate denoted $X$ which may be either continuous or dichotomous (zero/one coded). Let the vector $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iM})$ represent individual $i$'s responses to the $M$ items, where the possible values of $Y_{im}$ are 1, …, $r_m$. Let $L_i = 1,2,\ldots,n_c$ be the latent class membership of individual $i$, and let $I(y = k)$ be the indicator function; that is, a function which equals 1 if $y$ equals $k$, and 0 otherwise. Suppose we let the last class be the reference class. Let $X_i$ represent the value of the covariate for individual $i$; the covariate may be related to the probability of membership in each latent class, $\gamma$, but is assumed to be otherwise unrelated to $\mathbf{Y}_i$. Then the contribution by individual $i$ to the likelihood is

$$P\left(\mathbf{Y}_i = \mathbf{y} \mid X_i = x\right) = \sum_{l=1}^{n_c} \gamma_l(x) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_m = k)} \qquad (1)$$

The $\beta$ parameters are the coefficients in logistic regressions using the covariate $X$ to model the class membership parameters $\gamma$. The $\gamma$ parameters can be expressed as

$$\gamma_l(x) = P(L_i = l \mid X_i = x) = \frac{\exp\left(\beta_{0l} + x\beta_{1l}\right)}{\sum_{j=1}^{n_c} \exp\left(\beta_{0j} + x\beta_{1j}\right)} = \frac{\exp\left(\beta_{0l} + x\beta_{1l}\right)}{1 + \sum_{j=1}^{n_c - 1} \exp\left(\beta_{0j} + x\beta_{1j}\right)} \qquad (2)$$

for $l = 1,\ldots,n_c$. Note that the latter two terms on the right are equal because we assume that the last (i.e., the $n_c$ th) class is used as the reference class. The reference class has its $\beta$s constrained to zero because the relative probabilities of being in the other classes are being compared to the probability of this reference class. It is necessary to set the $\beta$s for some class to zero for the sake of model identifiability, because of the natural constraint that the probabilities for all classes must sum to one for each individual, but it need not be the last class. The choice of reference class does

not affect the final fitted probability estimates for any individual or class.

This model allows us to estimate the log odds that individual $i$ falls in latent class $l$ relative to the reference class. For example, if class 2 is the reference class, then the log odds of membership in class 1 relative to class 2 for an individual with value $x$ on the covariate is

$$\log\left(\frac{\gamma_1(x)}{\gamma_2(x)}\right) = \beta_{01} + \beta_{11}x \quad (3)$$

.

Exponentiated $\beta$ parameters are odds ratios, reflecting the increase in odds of class membership (relative to reference class $n_c$) corresponding to a one-unit increase in the covariate. Note that multiple covariates can be included simultaneously, just as in logistic regression. For models involving three or more latent classes, PROC LCA also includes an option to conduct binary logistic regression, as opposed to baseline-category multinomial logistic regression, when predicting latent class membership. A comparison class is specified by the user, and all other latent classes are combined into one reference group. Covariates are then used to predict membership in the specified class relative to the others. This option provides a more parsimonious prediction model and may be useful in some cases where the multinomial logistic regression model is not estimable due to sparseness.

# 3 The LTA Mathematical Model

The following sets of parameters are estimated in PROC LTA:

- Delta ($\delta$) parameters: latent status membership probabilities at Time 1
- Tau ($\tau$) parameters: probabilities of transitions between latent statuses over time
- Rho ($\rho$) parameters: item-response probabilities conditional on latent status membership and time

The $\rho$ parameters express the correspondence between the observed items and the latent statuses, and form the basis for interpretation of the latent statuses. When one or more covariates are included, two additional sets of $\beta$ parameters may be estimated:

- A set of $\beta$ parameters which are logistic regression coefficients for covariates predicting latent status membership at Time 1
- A further set of $\beta$ parameters which are logistic regression coefficients for covariates predicting transitions over time.

When covariates are included, only $\rho$ and $\beta$ parameters need actually be estimated; in this case, the $\delta$ and $\tau$ parameters are calculated as functions of $\beta$ parameters and the covariates, and are provided in PROC LTA output. If a grouping variable such as gender is included, all sets of parameters ($\delta$, $\tau$, $\rho$, $\beta$) can be conditioned on group.

Suppose a latent transition model with $n_s$ latent statuses is to be estimated based on a dataset including $M$ categorical response items measured at each of $T$ times for a total of $MT$ items; a covariate $X$; and a grouping variable $G$. Let

$$\mathbf{Y}_i = (Y_{i11}, Y_{i12}, \ldots, Y_{i1M}, Y_{i21}, Y_{i22}, \ldots, Y_{i2M}, \ldots, Y_{iT1}, Y_{iT2}, \ldots, Y_{iTM})$$

represent the vector of individual $i$'s responses for all times $t$=1,…,$T$, and items $m$=1, …., $M$, where an individual response $Y_{itm}$ may take on the values 1, 2, …, $r_m$. Let $S_{1i}$=1, 2, …, $n_s$ be individual $i$'s latent status membership at Time 1, $S_{2i}$=1, 2, …, $n_s$ be individual $i$'s latent status membership at Time 2, and so on. Let $I(y = k)$ be the indicator function which equals 1 if $y$ equals $k$ and 0 otherwise. Let $G_i$ represent individual $i$'s group membership. Finally, let $X_i$ be the value of the covariate $X$ for individual $i$; the value of $X$ may be related to the probabilities of membership in the latent statuses (the $\delta$s) as well as the transition probabilities (the $\tau$s). The latent transition model can then be expressed as

$$\begin{aligned} &P(\mathbf{Y}_i = \mathbf{y} \mid X_i = x, G_i = g) \\ &= \sum_{s_1=1}^{n_s} \cdots \sum_{s_T=1}^{n_s} \delta_{s_1|g}(x)\tau_{s_2|s_1,g}(x)\cdots\tau_{s_T|s_{T-1},g}(x) \prod_{m=1}^{M}\prod_{k=1}^{r_m}\prod_{t=1}^{T} \rho_{mk|s_t,g}^{I(y_m=k)}. \end{aligned} \tag{4}$$

The probability of belonging to latent status $s$ at Time 1 is given by the $\delta$ parameter $\delta_{s|g}(x) = P(S_{1i} = s \mid X_i = x, G_i = g)$. As with the $\gamma$s in LCA, the $\delta$s are related to the covariates via a standard baseline-category multinomial logistic model (see, e.g., Agresti, 2002). For example, with one covariate $X$, the parameters are expressed as a function of the $\beta$ parameters (i.e., the multinomial logistic regression coefficient estimates) and $X$:

$$\delta_{s|g}(x) = P(S_{1i} = s \mid X_i = x, G_i = g) = \frac{\exp\left(\beta_{0s|g} + x\beta_{1s|g}\right)}{1 + \sum_{j=1}^{n_s - 1} \exp\left(\beta_{0s|j} + x\beta_{1s|j}\right)} \quad (5)$$

for $s = 1, 2, ..., n_s - 1$. The $n_s^{\text{th}}$ latent status serves as the reference class in this logistic regression, which estimates the log-odds that an individual falls in latent status $s$ relative to reference status $n_s$. For example, if latent status 2 is the reference status, then the log-odds of membership in latent status 1 relative to latent status 2 for an individual in group 1 with value $x$ on the covariate X is

$$\log\left(\frac{\delta_{1|1}(x)}{\delta_{2|1}(x)}\right) = \beta_{01|1} + \beta_{11|1}x. \quad (6)$$

Exponentiated $\beta$ parameters are odds ratios. For example, $\exp\left(\beta_{11|1}\right)$ is an odds ratio reflecting the increase in odds of membership in latent status 1 (relative to the reference latent status, $n_s$) corresponding to a one-unit increase in the covariate, among individuals in group 1.

Similarly, $\tau_{s_2|s_1,g}(x) = P(S_{2i} = s_2 \mid S_{1i} = s_1, X_i = x, G_i = g)$ is determined by a baseline-category multinomial logistic model estimating the probability of individual $i$'s move to latent status $s_2$ conditional upon current membership in status $s_1$. For example, the probability of individual $i$ transitioning from latent status $s_1$ at Time 1 to latent status $s_2$ at Time 2 given membership in group $g$ and covariate value $x$ is

$$\tau_{s_2|s_1,g}(x) = \frac{\exp\left(\beta_{0s_2|s_1,g} + x\beta_{1s_2|s_1,g}\right)}{1 + \sum_{j=1}^{n_s - 1} \exp\left(\beta_{0s_2|s_1,j} + x\beta_{1s_2|s_1,j}\right)} \quad (7)$$

for $s_2 = 1, \ldots, n_s$. (Here latent status $n_s$ is serving as the reference status.) Note that more than one covariate can be included, and different covariates can be specified for $\delta$ and for the $\tau$ matrices.

For models involving three or more latent statuses, PROC LTA also includes an option to conduct binary logistic regression, as opposed to baseline-category multinomial logistic regression, when predicting latent status membership and transitions over time. For each regression, a comparison status is specified by the user, and all other latent statuses are combined into one reference group. Covariates are then used to predict membership in the specified status relative to any other status. This option provides a more parsimonious prediction model and may be used in some cases where the multinomial logistic regression model is not estimable due to sparseness of data, which is most likely to occur in the prediction of transition probabilities.

Estimation, missing data on the latent class/status indicators, and determining convergence are handled in the same way in PROC LCA and PROC LTA.

# 4 Technical Details

## 4.1 Estimation

In PROC LCA and PROC LTA, parameters are estimated by maximum likelihood using the EM algorithm, with Newton-Raphson incorporated into the estimation of regression coefficients for covariates. The convergence index used is the maximum absolute deviation (MAD). The MAD associated with a particular iteration of the estimation procedure is computed by calculating the absolute value of the difference between the current iteration parameter estimates and those corresponding to the previous iteration; the value assigned to MAD for that iteration is the largest number in this array. Ordinarily the value of MAD becomes smaller with each iteration of the estimation procedure, although there are conditions under which this may not hold. The estimation procedure iterates until either a previously specified criterion value of MAD (the convergence criterion) or a previously specified maximum number of iterations is reached.

## 4.2 Missing Data

Missing data on the latent class and latent status indicators are permitted in these procedures. Missing values should be represented as SAS system missing (".") as usual in SAS. When there are missing data the models expressed in Equations 1 and 4 are modified so that the product over $m$=1,…,$M$ is replaced by a product over the items observed for that individual.

Data are assumed to be missing at random (MAR). A test of the null hypothesis that data are missing completely at random (MCAR) also appears in the output. Missing data on covariates, groups, clusters, or weights (if these features are included in the model) are not allowed. That is, any record with missing data on covariates, groups, clusters or weights variables specified in the model is eliminated from the analysis.

## 4.3 Standard Errors (PROC LCA only)

Asymptotic standard errors for LCA parameter estimates are provided when available. For models without weights or clustering, standard errors are found by inverting the Hessian matrix of the log likelihood (see the "standard" option in Latent GOLD; Vermunt & Magidson, 2005a, pp. 98-100, for technical details). For models with weights or clustering, a "robust" or "sandwich" standard error based on Taylor linearization is used (see the "robust" option in Latent GOLD).

## 4.4 Clusters and Weights (PROC LCA only)

In many contexts in the social sciences, data arise from a sampling scheme more complicated than a simple random sample. Very often, participants are selected with unequal probabilities, so that in order to accurately describe population proportions, observations need to be given different weights. Also, instead of being independent, participants are often nested within clusters ("primary sampling units") such as schools, clinics or neighborhoods.

The current version of PROC LCA can now accomodate clusters and weights using the pseudo-maximum-likelihood approach (Skinner, 1989; Vermunt & Magidson, 2005b, pp. 98-100). Under this approach, sampling weights are first standardized to have an average value of 1 over all of the individuals being analyzed; they are then used as if they were frequency weights in calculating the estimates. Clustering is ignored for estimation purposes, but is taken into account in calculating standard errors by using a "robust" or "sandwich" style covariance estimate.

***Note:*** PROC LCA assumes that all of the data are from the same stratum in the sampling sense.

***Note:*** Even if the GROUPS statement is used, the weights are standardized to average to 1 across the whole analyzed dataset, not within each group separately. Users who wish to take a different approach may standardize weights as they wish prior to conducting the latent class analysis, and then use the ORIG_WEIGHTS option (see p. 13) to specify that original weights be used.

***Note:*** Latent GOLD (Vermunt & Magidson, 2005a, 2005b) uses the pseudolikelihood approach by default to handle sampling weights and clustering. The pseudolikelihood approach is also one of the two approaches available in MPlus for complex survey data (see Asparouhov, 2005; Muthen & Muthen, 2010, p. 233).

***Note:*** When weights or clusters are present, inference is done using the "pseudo" or "weighted" log-likelihood function, because the true likelihood taking sampling into account may be difficult to find. Therefore, in PROC LCA the $G^2$, AIC, BIC, CAIC, ABIC, and entropy statistics are also based on the log pseudolikelihood. However, the classic literature on these criteria generally assumes that they are based on a true log-likelihood from a model with equally weighted independent observations. This may mean that they are more difficult to interpret or must be interpreted with more caution because their statistical properties are largely unknown (Vermunt & Magidson 2007). However, they may still be useful as heuristics (Wedel, ter Hofstede, & Steenkamp, 1998).

***Note:*** When weights or clusters are present, the log-likelihood test for the significance of a covariate is corrected for the effects of the weights and clusters as recommended by Satorra and Bentler (1988) and Asparouhov and Muthén (2005).

# 5 PROC LCA Syntax

The following statements are available in PROC LCA. Only the bold lines are required; the others are optional.

**PROC LCA** < *options* >;
**NCLASS** *value*;
**ITEMS** *variables*;
**CATEGORIES** *values*;
ID *variables*;
GROUPS *variables*;
GROUPNAMES *labels*;
MEASUREMENT *keyword*;
COVARIATES *variables*;
REFERENCE *value*;
BINARY *value*;
CORES *value*;
BETA PRIOR= *value*;
GAMMA PRIOR= *value*;
RHO PRIOR= *value*;
FREQ *variable*;
WEIGHT *variable*;
CLUSTERS *variable*;
ESTIMATION *estimation-method*;
SEED *value*;
SEED_DRAWS *value*;
NSTARTS *value*;
MAXITER *value*;
CRITERION *value*;
**RUN**;

## 5.1 Invoking the LCA Procedure

To begin the LCA procedure, use the following line of SAS code:

**PROC LCA DATA =** *SAS-data-set* < *options* >;

The data file can contain more variables than will be used in the analysis. It must contain at least 2 categorical variables to be used as indicators for the latent class model. The data file can be organized using one record per individual or aggregated with one record per response pattern.

If data are aggregated, the file must contain a frequency count variable. The first 12 characters of variable names will be displayed in the output.

There are several options that may be specified in the PROC LCA statement. The options mainly concern input and output.

## 5.2 Options for Input

**START =** *SAS-data-set*

The START option allows the user to specify a SAS data file containing starting values for the parameters. This data file must contain starting values for the $\gamma$ and $\rho$ parameters (starting values for $\beta$ parameters are optional). If starting values for the $\rho$ parameters are of main interest, then the user can simply provide "flat" starting values (1/NCLASS) for the $\gamma$ parameters and starting values of 0 for all $\beta$s (these are the defaults). The structure of this file must be identical to that of a file created with the OUTPARAM option (see page 15), except that rows for the $\beta$ parameters are optional. Appendix 4 provides an example in which starting values are provided in a SAS data set.

***Note*** : If the START option is not invoked, the SEED statement (see page 21) must be included. If the START option is invoked, the SEED statement and the NSTARTS statement may not be included. SEED and START may not be specified together.

***User Tip*** : When using the START option to specify starting values, a SAS data file containing no rows for the $\beta$ parameters can be used for models with no covariates, as well as models with any number of covariates.

**RESTRICT =** *SAS-data-set*

The RESTRICT option allows the user to specify a SAS data file containing parameter restrictions. Parameter restrictions for the $\rho$ parameters can be useful to help achieve model identification or to test specific hypotheses about the measurement of the latent class variable. Parameter restrictions for the $\gamma$ parameters can be used to test hypotheses about the prevalence of latent classes, or to fix the probability of membership in a latent class to zero for a particular group. The SAS data file containing parameter restrictions must have a structure identical to that of a file created with the OUTPARAM option (see page 15), except that rows for the $\beta$ parameters are optional. The file must specify a restriction option, indicated by an integer of value 0 or higher, corresponding to each parameter. Appendix 4 provides an example in which restrictions are provided in a SAS data set.

The following restrictions for $\rho$ and $\gamma$ parameters are possible.

- *A parameter may be fixed to a specific value.* A value of 0 in the parameter restriction file indicates that the parameter is to be fixed. A parameter that is fixed is not estimated but remains at the starting value provided. If the user wishes to fix parameter estimates to a specific value, then the START option must be used in conjunction with the RESTRICT option.
- *A parameter may be freely estimated with no restrictions.* A value of 1 in the parameter

restriction file indicates that the parameter is to be freely estimated (this is also the default when the RESTRICT option is not used).

- *A parameter may form part of an equivalence set.* Integers of value 2 or greater specify an equivalence set; estimates for all parameters with the same value are constrained equal to one another and only one parameter is estimated for each set. ***Note:*** This must be restricted separately for $\rho$ and $\gamma$ parameters as needed.

Restrictions may not be placed on $\beta$ parameters. If the SAS data file contains rows for the $\beta$ parameters, all restriction values for these parameters should be 1, indicating free estimation.

***Note*** : The RESRICT data file should be in order first by parameter type, then by group, then by response category, and last by variable, as in the example in Appendix 4. Optionally, the file could instead be in order by group, then by parameter type, then by response category, and last by variable. If they are given in any other arbitrary order then the restrictions may be interpreted incorrectly by the PROC.

***Note*** : If an equivalence set is imposed in the $\gamma$ parameters, then covariates may not be used to predict class membership.

***Note*** : There are a few kinds of restrictions which still allow estimates to be computed but for which standard errors are unavailable. These are: (1) One or more $\gamma$s are preset to constants. (2) Some, but not all, $\gamma$s are put in equivalence sets. (3) A $\rho$ in a polychotomous item (>2 categories) is constrained but another $\rho$ in the same item is free.

***Note*** : If the RESTRICT statement is used then the CLUSTERS statement may not be used.

***User Tip*** : For convenience, the MEASUREMENT statement (described on page 18) can be used to restrict $\rho$ parameters to be invariant across groups without using the RESTRICT option. If both the RESTRICT option and the MEASUREMENT statement are used, restrictions corresponding to $\rho$ parameters for Group 1 that are provided in the SAS data file are applied to all subsequent groups. Additional information on the use of parameter restrictions can be found in separate documentation on the Web (WinLTA General Users' Guide, available at *methodology.psu.edu*).

**ORIG_WEIGHTS**

This option is only relevant if sampling weights are being used. If this option is not provided, the weights are standardized to average to 1 over the subjects included in the analysis. Thus, they are assumed to express the relative importance of each subject, but don't change the overall sample size. The ORIG_WEIGHTS option can be used if it is desired that the weights be implemented "as-is." Users may wish to invoke this option if, for example, they wish to standardize weights to average to 1 within each GROUP separately.

## 5.3 Options for Output

**NOBETATEST**

This suppresses tests of significance for covariates. This option has meaning only when the COVARIATES statement is used. If significance tests for covariates are not of interest, then invoking this option is recommended as it speeds up model estimation.

**VERBOSE_OUTPUT**

Including this option produces output that includes the following: the number of $\gamma$ and $\rho$ parameters estimated in the specified model, restrictions for all parameters, starting values for all parameters, and the iteration history (this shows the MAD and log likelihood at each iteration). If NSTARTS (see page 22) is used, seed and log likelihood for each start will be displayed. Invoking this option can result in lengthy output.

**OUTEST =** *SAS-data-set*

This option produces an output SAS data file with the specified name containing final parameter estimates, as well as the log likelihood value, degrees of freedom, and fit indices. (The degrees of freedom may not be shown if the value is extremely high—for example, over a million—as may happen with a very complicated model.) The format of this file is one record. Each parameter estimate has a unique variable name, which can be found in the data file. The other fields in the OUTEST file are defined below.

| **Variable Name** | **Description** |
|---|---|
| Log likelihood: LOG_LIKELIHOOD | The log likelihood of the fitted model |
| Degrees of freedom: DEGREES_OF_FREEDOM | The degrees of freedom of the fitted model. In models with no covariates, this is the number of cells in the contingency table, minus the number of parameters that are freely estimated, minus 1. |
| $G^2$: G_SQUARED | The deviance statistic |
| AIC: AIC | The Akaike Information Criterion (Akaike, 1973; see Lin & Dayton, 1997) |
| BIC: BIC | The Schwarz Bayesian Information Criterion (Schwarz, 1978; see Lin & Dayton, 1997) |
| CAIC: CAIC | The "consistent AIC" (Bozdogan, 1987; see Lin & Dayton, 1997) |
| Adjusted BIC: ABIC | The adjusted BIC using Rissanen's sample size adjustment (see Sclove, 1987) |
| Raw entropy: ENTROPYRAW | The mathematical entropy of the class partitioning, equal to $S = -\sum_{i=1}^{n}\sum_{k=1}^{L} p_{ik} \log p_i k$ where $L$ is number of classes. |
| Entropy statistic: ENTROPY | The scaled relative entropy $1 - S / (n \log L)$ (Ramaswamy et al., 1993). This is the entropy statistic reported in some packages such as Mplus (Muthén, 2004). |
| Multivariate design effect: | (Only when WEIGHTs and/or CLUSTERs are used.) |

| | |
|---|---|
| DESIGNEFFECT | The multivariate design effect based on the matrices in the sandwich covariance estimate, if available (Skinner, Holt, & Smith, 1989; Vermunt & Magidson, 2005b, p. 100) |

Except for LOG_LIKELIHOOD, these fit statistics are not provided when covariates are included in the model.

The contents of this file also are displayed in the SAS output (with less precision); this option is typically needed only when parameter estimates are to be used in subsequent analyses or when high precision is needed.

***Note***: When WEIGHTs or CLUSTERs are used, the fit statistics are based on a pseudo-likelihood rather than a true likelihood, which may complicate their interpretation (see Vermunt & Magidson, 2007; Wedel et al., 2008).

**OUTPARAM =** *SAS-data-set*

This option produces an output SAS data file with the specified name containing final parameter estimates. Contents of this file also appear in the file created by the OUTEST option, although the structure of the file is different and estimates are less precise. In this file, parameter estimates are presented in a user-friendly format. Estimates can be identified by the first four columns: Parameter Type (PARAM), Group Number (GROUP), Variable Name (VARIABLE), and Response Category (RESPCAT). Values for PARAM are $\gamma$, $\beta$, or $\rho$. The number of lines in each parameter set depends on the number of groups, covariates, indicators, and the number of response categories for each indicator. In each record, the final parameter estimates for each latent class are presented for that particular combination of Parameter Type, Group Number, Variable Name, and Response Category. The user may wish to generate this data file when greater precision of parameter estimates is needed than what is presented in the SAS output.

***User Tip:*** If starting values and/or parameter restrictions are provided by the user, they must be presented in a SAS data file with the exact structure of this OUTPARAM data set (with one exception: SAS data files specifying starting values or parameter restrictions need not include rows for $\beta$ parameters). The user may wish to fit a preliminary latent class model including the OUTPARAM option, rename this SAS data file, modify the new file by replacing the preliminary parameter estimates with either starting values or parameter restrictions, and rerun the latent class model including the START and/or RESTRICT option (see above for details on the START and RESTRICT options). This practice will ensure that the structure of the starting values or restrictions SAS data file is correct.

**OUTPOST =** *SAS-data-set*

This option produces an output SAS data file with the specified name that contains the posterior probabilities of latent class membership. The format of this file is the same as that of the original SAS data file (one record per individual, or aggregated if the FREQ option is used). The posterior probabilities appear in newly created variables named POSTLC1, POSTLC2, etc. In addition, the value in the newly created variable named BEST indicates the latent class for which

each individual has the highest posterior probability of membership. This is often refered to as the maximum-probability assignment rule (see Bray, Lanza, & Tan, 2012 for recommendations on how to use posterior probability-based assignments). When data are not aggregated, the OUTPOST data file contains the following variables: items indicating the latent class variable (listed in the ITEMS statement), the grouping variable, the covariates, the posterior probabilities, the latent class assignment based on the maximum-probability assignment rule, and any variables specified in the ID statement. When data are aggregated, the OUTPOST data file contains the following variables: items indicating the latent class variable (listed in the ITEMS statement), the grouping variable, the covariates, the count variable (listed in the FREQ statement), the posterior probabilities, and the latent class assignment based on the maximum-probability assignment rule.

***Changes since version 1.3.0:*** The "Best" variable (described above) has been added to the file. Also, if used in conjuction with the SEED_DRAWS statement, the OUTPOST file will also include 20 columns of random draws from the multinomial distribution defined by each individual's posterior probabilities (Bandeen-Roche et al., 1997; Wang, Brown, & Bandeen-Roche, 2005). This is often refered to as assignment based on multiple pseudo-class draws (see Bray et al., 2012 for recommendations on how to use posterior probability-based assignments). The 20 pseudo-class draw assignments are indicated in newly created variables named DRAW_1 through DRAW_20.

***Change since version 1.2.3:*** OUTPOST will no longer run unless an ID variable is supplied. This change is intended to prevent accidentally associating posterior probabilities with the wrong cases.

***User Tip:*** To quickly add a sequential row ID variable to a SAS dataset, one can use the following code: DATA [*datasetname*]; SET [*datasetname*]; DO ID = 1, _N_; END; OUTPUT; RUN;

**OUTSTDERR**

This option produces an output SAS data file containing standard errors for estimated parameters.

**OUTSEEDS**

This option should be used only when NSTARTS is present. It produces an output SAS data file containing random seeds, log likelihood values and $\gamma$ parameters for each start. This can be used to assess model identification.

**NOPRINT**

This option suppresses the printing of output. This may be useful when the parameter estimates are being saved to a file (see OUTEST and OUTPARAM options) and no additional output is needed.

**OUTCOVB**

This option produces an output data file containing the estimated covariance matric of the beta parameter estimates. It is for use with the LCA Distal Outcomes Macro. Its use is described in the *LCA Distal Outcomes Macro Users' Guide Version 3.0* or higher.

## 5.4 Required Statements for PROC LCA

The following statements are required in order to specify the model to be fit by PROC LCA:

**NCLASS** *value*;

This statement specifies the number of latent classes in the model to be estimated. Valid values are integers greater than or equal to 1.

**ITEMS** *variables*;

This statement lists the categorical variables to be used as indicators of the latent classes. Two or more variables must be specified, and the number of arguments in the CATEGORIES statement must equal the number of variables listed in the ITEMS statement. Each indicator must be coded with sequential integer values from 1 to *R*, where *R* is the number of response categories for that particular item. Missing values are permitted and should be coded as SAS system missing values (.).

**CATEGORIES** *values*;

This statement lists the number of response categories in each of the variables listed in the ITEMS statement. Integer values must be listed in the same order as the variables listed in the ITEMS statement. Values must be between 2 and 99.

## 5.5 Optional Statements for PROC LCA

**ID** *variables*;

The ID statement is used to specify one or more variables in the analysis data set that are to be appended to the OUTPOST SAS data file. It is used only in conjunction with the OUTPOST option of the PROC LCA statement, which allows the user to save posterior probabilities in a SAS data file (see section 5.3 Options for Output, OUTPOST option). When data are not aggregated, the OUTPOST data file contains the following variables: items indicating the latent class variable (listed in the ITEMS statement), the grouping variable, the covariates, the posterior probabilities, the latent class assignment based on the maximum-probability assignment rule, and any variables listed in the ID statement. Typically, when data are not aggregated, a case identifier exists in the analysis data set. By listing the case identifier in the ID statement, this identifier is carried through to the OUTPOST data file, allowing the user to merge the SAS data file containing posterior probabilities to other data files. See Appendix 5 for an example using the ID statement. Note that more than one variable can be specified in the ID statement; all variables listed here are included in the OUTPOST data file.

**GROUPS** *variable*;

Multiple-groups latent class analysis can be conducted using PROC LCA. The grouping variable is specified in the GROUPS statement. Only one grouping variable may be specified, although the user can cross several categorical variables to create a single grouping variable. The grouping variable must be coded with sequential integer values from 1 to the number of groups.

When the GROUPS statement is used, the user may wish to label the groups using the GROUPNAMES statement. Cases with missing data for the grouping variable will be deleted automatically. The number of cases used in the analysis will be noted in the output file and the number of cases read in and the number of deleted cases will be noted in the log file.

***User Tip:*** If the GROUPS and CLUSTERS statements are both used, then the MEASUREMENT statement must also be used to specify measurement invariance. In other words, if the data arises from a cluster sampling scheme, then PROC LCA requires the assumption of measurement invariance across groups.

**GROUPNAMES** *labels*;

This statement allows the user to specify labels (up to 12 characters for each) for the different levels of the grouping variable specified in the GROUPS statement. The number of labels listed in the GROUPNAMES statement must be equal to the number of groups, and the order of the labels must correspond to the order of the integers denoting the groups. This statement should only be used in conjunction with the GROUPS statement.

**MEASUREMENT** *keyword*;

When a grouping variable is provided in the GROUPS statement, the user can use the MEASUREMENT statement to impose measurement invariance across all groups, without having to use the RESTRICT option. The keyword GROUPS restricts estimation so that all $\rho$ parameters (class-specific item response probabilities) are equal across groups. Appendix 5 shows an example of the MEASUREMENT statement.

**COVARIATES** *variables*;

One or more covariates can be incorporated in the latent class model by specifying the variable names in the COVARIATES statement. The $\gamma$ parameters (probabilities of latent class membership) will depend on the values or levels of the covariates. (The $\rho$ parameters [item-response probabilities] will not depend on the values or levels of the covariates.) It is strongly recommended that the user first run the model without covariates to determine the latent structure (to select the number of latent classes), explore issues such as measurement invariance, and assess model fit. Note that covariates are treated as numeric (continuous variables and dummy-coded [i.e., dichotomous] variables are recommended). Cases with missing data for a covariate will be deleted automatically. The number of cases used in the analysis will be noted in the output file; the number of cases read in and the number of deleted cases will be noted in the log file. See Appendix 5 for an example using covariates.

Note that when the COVARIATES statement is specified, it is not possible to specify equivalence sets in the $\gamma$ parameters. However, individual $\gamma$ parameters may be fixed to their corresponding starting values.

**REFERENCE** *value*;

Use only in conjunction with the COVARIATES statement. The REFERENCE statement specifies the number of the latent class (an integer) to serve as the reference group for logistic regression. The minimum value is 1 and the maximum value is the number of classes specified in

the NCLASS statement. The default value is 1.

***Note:*** When random starting values are used, the order of the latent classes is random. When using the SEED statement, the user may wish to estimate a model, examine the output to choose the reference class, then specify that reference class in the syntax and rerun the model using the same SEED value. Alternatively, the user may wish to use the START option in the PROC LCA statement so that the expected ordering of the latent classes can be known.

**BINARY** *value;*

Use only in conjunction with the COVARIATES statement, in place of the REFERENCE statement. By default, PROC LCA uses baseline-category multinomial logistic regression to predict latent class membership. However, a binary logistic model may be specified using the BINARY statement.

Use this statement to specify the number of the latent class to serve as the comparison group for binary logistic regression. The remaining latent statuses will be combined to form the reference group for binary logistic regression. The minimum value is 1 and the maximum value is the number of classes specified in the NCLASS statement.

**CORES** *value*;

Specifies that the computational work should be divided among *value* different processors (cores), for multicore computers. The default value of 1 is assumed if this statement is not specified. Other common values are 2 or 4 depending on your computer.

**"STABILIZE"**

This statement is no longer used. To stabilize the $\beta$ parameters in a model with covariates and sparse data, use syntax BETA PRIOR = *value;*. **Setting value equal to the number of classes as specified by NCLASS (e.g., BETA PRIOR=4 in a four-class LCA model) will give the same prior strength as the STABILIZE command did in version 1.1.5.**

**BETA PRIOR** = *value*;

This statement is used only in conjunction with the COVARIATES statement, to invoke a stabilizing prior distribution on the $\beta$ parameters. It creates a data-derived prior which is used in the estimation of each logistic model specified by the user. The value provided, which must be a positive real number, controls the strength of the prior (and how strongly we want it to influence the $\beta$ estimates). A strength of 0 would mean no prior (ordinary maximum-likelihood estimation). A strength of 1 is recommended if a prior is desired.

***User Tip:*** If estimation of a logit model fails, the problem is likely due to sparseness in the data. The recommended course of action is to invoke the BETA PRIOR statement, which will solve most sparseness-related estimation problems. In extreme cases, this approach may not suffice and additional measures must be taken. One option is to reduce the number of parameters in the logit model by switching from a baseline-category multinomial logit model to a binary logit model. Also, be sure to check that no class membership probability is estimated at zero for one of the groups. (These should be examined in the model with no covariates, fit only to individuals who provided data on the covariate(s).) Any class membership probabilities that are estimated at

a value very close to 0 can be fixed to 0 using the RESTRICT option. This eliminates the empty class from the logistic regression for that group.

***Note:*** This approach is a practical solution for stabilizing the estimation of logit parameters when one or more of the $\beta$ estimates diverge to infinity due to insufficient information for estimation. Sparseness is more likely to cause estimation problems when the sample size is small, one or more groups is small, one of the latent classes has a very small class membership probability, or when membership in one of the classes is essentially zero for some level of a covariate. This last condition can be difficult to identify, as true class membership is unknown. For more information about the prior used here, see Clogg, Rubin, Schenker, Schultz, and Weidman (1991).

**GAMMA PRIOR** = *value;*

This statement invokes a data-derived prior on the $\gamma$ estimates. The value provided with GAMMA PRIOR, which must be a positive real number, controls the strength of the prior (and how strongly we want it to influence the $\gamma$ estimates). We recommend a strength of 1 as standard. The $\gamma$-stabilizing prior strength in PROC LCA is similar to the "Bayes constant" for "latent variables" in the latent class clustering functionality in LatentGOLD (Vermunt & Magidson, 2005a, 2005b). It essentially adds a small number of pseudo-cases to each class, in order to improve estimation overall by biasing $\gamma$ estimates away from zero. The specified strength is the total number of pseudo-cases (if there is no grouping variable) or the total number per group (if there is). The GAMMA PRIOR statement can be used when there are no covariates; if your model has covariates use BETA PRIOR instead.

**RHO PRIOR** = *value*;

This statement invokes a data-derived prior on the $\rho$ estimates. The $\rho$-stabilizing prior strength in PROC LCA is somewhat similar to the "Bayes constant" for "categorical variables" in the latent class clustering functionality in LatentGOLD (Vermunt & Magidson, 2005a, 2005b). The value must be a positive real number. We recommend a strength of 1 as standard, although other values can also be used. This prior acts somewhat like adding a small number of pseudo-cases to each response category for each class, in order to improve estimation overall by biasing it away from solutions in which some $\rho$s are zero or one. This is important if standard errors are desired because standard errors cannot be calculated for models with estimates on the boundary of the parameter space (parameters estimated at zero or one without a prior). The strength given is the total number of pseudo-cases (if there is no grouping variable) or the total number of pseudo-cases per group (if there is).

**FREQ** *variable;*

PROC LCA can analyze data with one record per case or data that are aggregated into response patterns (with a count variable). The FREQ statement must be used if data are aggregated. The variable containing the count variable is specified here. If data are not in such aggregated form, this statement should not be used. Frequency weights will usually be integers (whole numbers) but are not required to be. They are required to be greater than zero. Appendix 3 shows an example of PROC LCA being used with aggregated data.

For aggregated data with 4 items and a **Count** variable, the data could be coded like this. The **freq** option would be used to indicate that the data are aggregated.

| Item 1 | Item 2 | Item 3 | Item 4 | Count |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 44 |
| 1 | 1 | 1 | 2 | 162 |
| 1 | 1 | 2 | 2 | 73 |

**WEIGHT** *variable;*

This statement indicates that inverse-probability sampling weights should be used to adjust the data. By default, sampling weights are standardized before using them, so that they average to 1 over the subjects used in the analysis. (Cases that are deleted because they are missing covariates, are missing the grouping variable, are missing weights or frequencies, or are missing all of the indicators, are not included in this averaging.) Users who do not want weights to be standardized, or wish to do this manually, can use the ORIG_WEIGHTS option.

**CLUSTERS** *variable;*

This statement tells PROC LCA that the subjects are not independent random draws, but are nested within clusters (primary sampling units) such as schools or classrooms. The schools or classrooms, rather than the individuals, are then assumed to be independent of each other. *variable* gives the name of a SAS variable which must consist of positive integers (whole numbers), used as identification numbers for the cluster. For example, everyone having 2 in their cluster ID variable is assumed to be nested inside cluster 2.

**ESTIMATION** *estimation-method;*

This statement is used to specify the estimation method to be employed. Currently, the only method available is EM (expectation-maximization), which produces maximum-likelihood estimates for all parameters in the model. Thus this statement is not currently needed, but is optionally allowed in order to protect future version compatibility.

**SEED** *value;*

Random starting values for the $\rho$ parameters can be generated in PROC LCA by specifying a positive integer value in the SEED statement. (Default starting values of 1/NCLASS for $\gamma$ parameters and 0 for $\beta$ parameters are used.) An integer seed to generate the random values allows the user to replicate the analysis at a later date. If the START option is not used then a random number generator seed must be provided in the SEED statement. See Appendix 3 for an example using this statement.

***Note:*** The seed should be an integer (whole number) greater than 1 and less than 2,000,000,000. We use arbitrary six-digit seeds in our examples in the Appendices. The value of the seed has no substantive meaning and is used to generate random values. We ask the user to specify the seed instead of simply using an automatically generated seed as many software packages do, because by saving the seed the user can generate the same sequence of random

values when an analysis is completed. If the SEED statement is not included, the START option in the PROC LCA statement must be included.

***User Tip:*** If 0 or a negative number is provided as the SEED value, and if the NSTARTS statement is also being used, PROC LCA will generate a seed automatically using the computer's current system time. This produces different starting values each time you run the procedure.

**SEED_DRAWS** *value;*

This statement can only be used in conjunction with the OUTPOST option. The user provides an integer that serves as a random seed for generating posterior class membership draws for each individual in the OUTPOST file. If SEED_DRAWS is not used, the random posterior draws are not generated in the OUTPOST file.

***Note:*** The seed should be an integer (whole number) greater than 1 and less than 9,999,999,999.

**NSTARTS** *value*;

This command allows the model to be fit several times, in order to try to find the best estimates and avoid suboptimal local maxima of the likelihood function.

***Change since version 1.3.0:*** NSTARTS can now be used even when covariates are present in the model.

***User Tip:*** It is good practice to check the identification of all models, both those without and with covariates. . NSTARTS can be used to find the optimal seed or a good set of starting values, which can then be used to replicate the model at a later date. To do this, the optimal seed will be noted in the output file or the OUTPARAM option can be specified to create a SAS data file that can then be used to provide starting values with the START option.

***Note:*** If the user specifies an NSTARTS value greater than one, a starting value in SEED is still needed. However, this random seed is not used directly for creating starting values, but instead is used for generating a set of seeds to be used in the repeated estimation.

**MAXITER** *value*;

The MAXITER statement allows the user to specify the maximum number of iterations in the EM estimation procedure. The default value is 5000. If convergence is reached before the value specified in the MAXITER statement, the procedure will terminate normally.

**CRITERION** *value*;

The CRITERION statement allows the user to specify the maximum absolute deviation convergence criterion for the estimation procedure. The default value is 0.000001.

**Table 1: Summary of PROC LCA Syntax**

| Statement | Description & Default (if applicable) |
|---|---|
| **PROC LCA** | Invokes the procedure. |
| **NCLASS** | Specifies number of latent classes. |
| **ITEMS** | Declares variables that indicate latent class variable. |
| **CATEGORIES** | Specifies number of response categories in items. |
| ID | Declares identifier and other variables to retain in posterior probabilities file. |
| GROUPS | Declares categorical grouping variable. |
| GROUPNAMES | Specifies a label for each group. |
| MEASUREMENT | Invokes measurement invariance across groups. |
| COVARIATES | Declares variables to include as covariates. |
| REFERENCE | Specifies latent class to use as reference group in prediction from covariates. Default: 1 |
| BINARY | Specifies latent class to use as comparison group in prediction from covariates, and specifies that binary logistic regression is to be used. |
| CORES | Divides work between multiple cores on a multiprocessor computer. Default: 1. |
| BETA PRIOR= | Invokes a stabilizing prior for the $\beta$ parameters. Prior strength must be specified; as a standard we recommend BETA PRIOR=1. |
| GAMMA PRIOR= | Invokes a stabilizing prior for the GAMMA parameters. Prior strength must be specified; as a standard we recommend GAMMA PRIOR=1. |
| RHO PRIOR= | Invokes a stabilizing prior for the RHO parameters. Prior strength must be specified; as a standard we recommend RHO PRIOR=1. |
| CLUSTERS | Declares a cluster ID (primary sampling unit) and tells PROC LCA that the data are clustered. |
| FREQ | Identifies the frequency count variable, to use when data are aggregated. |
| WEIGHT | Identifies the sampling weight variable, to use with complex survey data. |
| ESTIMATION | Specifies estimation procedure. Default: EM. |
| SEED | Specifies seed for random number generator. * |
| SEED_DRAWS | Specifies seed for generating posterior class membership draws. |
| NSTARTS | Specifies the number of different random starting values to use. |
| MAXITER | Specifies maximum number of iterations. Default: 5000 |
| CRITERION | Specifies convergence criterion for MAD. Default: 0.000001 |

* SEED statement required if START option is not included in the PROC LCA statement or if the NSTARTS statement is used. START option and SEED statement may not be used together.

# 6 PROC LTA Syntax

The following statements are available in PROC LTA. Only the first five lines are required; the others are optional.

**PROC LTA** < *options* >;
**NSTATUS** *value*;
**NTIMES** *value*;
**ITEMS** *time1_variables time2_variables [etc.]*;
**CATEGORIES** *values*;
ID *variables*;
GROUPS *variable*;
GROUPNAMES *labels*;
MEASUREMENT *keywords*;
COVARIATES1 *variables*;
COVARIATES2 *time1-2_variables time2-3_variables [etc.]*;
REFERENCE1 *value*;
REFERENCE2 *value(s)*;
BINARY1 *value*;
BINARY2 *value(s)*;
CORES *value*;
BETA PRIOR = *value*;
FREQ variable;
ESTIMATION estimation-method;
SEED *value*;
MAXITER *value*;
CRITERION *value*;
**RUN;**

Syntax for PROC LTA is a direct extension of that used in PROC LCA. In the descriptions below, the reader is referred to the previous description in "PROC LCA Syntax" for options that are identical in the two procedures.

## 6.1 Invoking the LTA Procedure

To begin the LTA procedure, use the following line of SAS code:

**PROC LTA DATA =** *SAS-data-set* < *options* >;

The following options are available. They are very similar to the analogous options described for

the PROC LCA statement.

**OUTPOST** saves posterior probabilities as a dataset. Unlike the OUTPOST file in PROC LCA, OUTPOST does NOT include 20 columns of random draws from the multinomial distribution defined by each individual's posterior probabilities.
**OUTEST** saves parameter estimates and some fit statistics as a dataset.
**OUTPARAM** saves parameter estimates as a dataset in a different format.
**START** is used to specify starting values.
**RESTRICT** is used to specify restriction sets.
**NOPRINT** requests that no output be shown on the screen.
**VERBOSE_OUTPUT** makes the output more detailed.
**NOBETATEST** requests that the significance test for logistic regression coefficients for covariates not be done.

The following restrictions for $\delta$, $\tau$, and $\rho$ parameters are possible: a parameter may be fixed to its starting value, a parameter may be freely estimated, or a parameter may form part of an equivalence set.

Including the VERBOSE_OUTPUT option in PROC LTA produces output that includes the following: the number of $\delta$, $\tau$ and $\rho$ parameters estimated in the specified model, restrictions for all parameters, starting values for all parameters, and the iteration history (this shows the MAD and log likelihood at each iteration). Invoking this option can result in lengthy output. See Appendix 6 for an example of latent transition analysis with user-provided starting values and parameter restrictions.

## 6.2 Required Statements for PROC LTA

The following statements are required in order to specify the model to be fit by PROC LTA:

**NSTATUS** *value*;

This statement specifies the number of latent statuses at each time that are to be estimated. Valid values are integers greater than or equal to 2. See Appendices 6, 7, 8, and 9 for examples using this statement.

**NTIMES** *value*;

This statement specifies the number of times (occasions of measurement) in the model. Valid values are integers greater than or equal to 2. See Appendices 6, 7, 8, and 9 for examples using this statement.

**ITEMS** *time1_variables time2_variables [etc.]*;

This statement lists the categorical variables to be used as indicators of the latent statuses at each time. The number of times must correspond to the value specified in the NTIMES statement. The number of arguments in the CATEGORIES statement must equal the number of variables listed in the ITEMS statement divided by the number given by NTIMES. For example, if there are 4 arguments in the CATEGORIES statement (corresponding to four items measured

over time) and NTIMES is 3, then there must be $4 \times 3 = 12$ variables listed in the ITEMS statement. The variables listed in the ITEMS statement should be arranged according to time, so that all Time 1 variables are listed first, and then all Time 2 variables, and so on.

As in PROC LCA, each indicator must be coded with sequential integer values from 1 to $R$, where $R$ is the number of response categories for that particular item. Missing values are permitted and should be coded as SAS system missing values (.).

**CATEGORIES** *values*;

This statement lists the number of response categories for each of the Time 1 variables listed in the ITEMS statement. Integer values must be listed in the same order as the Time 1 variables listed in the ITEMS statement. For each item measured over time, the number of categories is assumed to be the same at all times. Values must be between 2 and 99.

## 6.3 Optional Statements for PROC LTA

The following statements are optional:

**ID** *variables*;

See description under PROC LCA Syntax.

**GROUPS** *variable*;

See description under PROC LCA Syntax.

**GROUPNAMES** *labels*;

See description under PROC LCA Syntax.

**MEASUREMENT** *keyword*;

This statement allows the user to impose certain common parameter restrictions without having to use the RESTRICT option. The TIMES keyword restricts the $\rho$ parameters to be equal across all times. When a grouping variable is provided in the GROUPS statement, the groups keyword can be used to constrain the $\rho$ parameters to be equal across groups. These keywords may be used individually or together. Appendix 8 shows an example of the MEASUREMENT statement using both keywords.

**COVARIATES1** *variables*;
**COVARIATES2** $time_{1\rightarrow 2}$*variables* $time_{2\rightarrow 3}$*variables* [etc.];

The COVARIATES1 and COVARIATES2 statements are used to specify one or more covariates to predict latent status membership. It is strongly recommended that before introducing covariates, the user first run the model without covariates to determine the latent structure (to select the number of latent statuses), explore issues such as measurement invariance, and assess model fit.

The COVARIATES1 statement is used to specify one or more covariates for predicting latent status membership at Time 1 (i.e., determining the subject-specific $\delta$ parameters). This is

analogous to the COVARIATES statement in PROC LCA, and may be used with or without the COVARIATES2 statement.

Note that when the COVARIATES1 statement is specified, it is not possible to specify equivalence sets in the $\delta$ parameters. However, individual $\delta$ parameters may be fixed to their corresponding starting values.

The COVARIATES2 statement is used to specify predictors of transition probabilities ($\tau$ parameters). A set of variables must be specified for each transition probability matrix, i.e., Time 1 to Time 2, Time 2 to Time 3, and so on (there are NTIMES - 1 matrices). Each set must include the same number of variables (typically, these variables will be the same across times or time-varying covariates). COVARIATES2 may be used with or without the COVARIATES1 statement.

Note that when the COVARIATES2 statement is specified, it is not possible to specify equivalence sets in the $\tau$ parameters. However, individual $\tau$ parameters may be fixed to their corresponding starting values.

Latent status membership probabilities at Time 1 ($\delta$ parameters) will depend on the values or levels of the covariates given in the COVARIATES1 statement. The probability of transitioning from a latent status at Time $t$ to a latent status at Time $t+1$ ($\tau$ parameters) will depend on the values or levels of the covariates provided in the COVARIATES2 statement for that transition probability matrix. Note that covariates are treated as numeric, so they should be coded as continuous variables or dummy-coded (i.e., dichotomous, coded 0 or 1) variables. Cases with missing data for a covariate will be deleted automatically. The number of cases used in the analysis will be noted in the output file and the number of cases read in and the number of deleted cases will be noted in the log file. See Appendices 8 and 9 for examples using these statements.

**REFERENCE1** *value*;
**REFERENCE2** *value(s)*;

Use REFERENCE1 only in conjunction with the COVARIATES1 statement; use REFERENCE2 only in conjunction with the COVARIATES2 statement.

The REFERENCE1 statement specifies the latent status at Time 1 to serve as the reference group for baseline-category multinomial logistic regression. The minimum value is 1 and the maximum value is the number of statuses specified in the NSTATUS statement. The default value is 1, corresponding to the first latent status. Appendix 8 demonstrates the use of the REFERENCE1 statement.

If COVARIATES2 is specified, there is a logistic regression model corresponding to each row of each transition probability matrix. The REFERENCE2 statement specifies the latent status to be used as a reference group for the baseline-category multinomial logistic regression corresponding to each logistic regression. For example, if there are four latent statuses and three times, then there would be two $\tau$ parameter matrices (Time 1 to Time 2, and Time 2 to Time 3), each with four rows (corresponding to the previous time's latent status memberships). In this example, a total of eight logistic regression models are estimated in the prediction of the transition probabilities. Eight values can be provided in the REFERENCE2 statement to specify the reference group for each of these regression models. The first four values correspond to each row of the Time 1 to Time 2 transition probability matrix, and the last four values correspond to each row of the Time 2 to Time 3 matrix. If in this example we wish to specify that status 4 is to be used as the

reference in all cases, then the REFERENCE2 statement would be:

**REFERENCE2** 4 4 4 4 4 4 4 4;

Equivalently, when all values contain the same number, one may provide a single value as follows:

**REFERENCE2** 4;

The default reference groups for REFERENCE2 are the diagonal elements of the transition probability matrices. For the example above, an equivalent statement to the default would be:

**REFERENCE2** 1 2 3 4 1 2 3 4;

An important additional feature of the REFERENCE2 statement gives the user the option of suppressing the estimation of one or more of the logistic regressions (i.e., skipping the prediction of one or more rows of the transition probability matrices). When a particular logistic regression is not estimable due to sparseness, this feature allows the user to skip the problematic logistic model but estimate coefficients in the remaining logistic regressions. A logistic regression is suppressed by providing a zero (0) for the corresponding value in the REFERENCE2 statement, and more than one zero can be specified. In the example above, suppose the user wishes to apply covariates only to the third row of the transition probability matrix for Time 2 to Time 3, but use the default reference statuses elsewhere. This is specified by:

**REFERENCE2** 1 2 3 4 0 0 3 0;

When the logistic regression corresponding to a row of the $\tau$ matrix is suppressed, the $\tau$ parameters for that row are estimated using the standard EM algorithm.

***Note:*** Only free or fixed $\tau$ parameters (not constrained $\tau$ parameters) may be included in rows that are skipped using the REFERENCE2 statement.

***Note:*** The REFERENCE1 statement may not be used with the BINARY1 statement. The REFERENCE2 statement may not be used with the BINARY2 statement. However, the REFERENCE1 statement may be used with the BINARY2 statement, and the BINARY1 statement may be used with the REFERENCE2 statement.

***User Tip:*** If estimation of a model involving covariates fails, the BETA PRIOR option may suffice to stabilize the logistic regression. However, in cases of extreme sparseness, certain parts of a model may still fail. When estimation of a model with covariates fails, we recommend first trying the BETA PRIOR option, and if estimation still fails, reducing the number of logit models being estimated (by skipping rows using the reference statements) or reducing the number of parameters estimated in one or more logit models (by switching to a binary logit model, described on page 19).

**BINARY1** *value*;
**BINARY2** *value(s)*;

See description under PROC LCA Syntax.

**CORES** *value*;

See description in PROC LCA Syntax.

"**STABILIZE** *value*;"

The STABILIZE statement is no longer used in PROC LTA. Instead, use "BETA PRIOR =

*value;*". **Setting *value* equal to 1 in an LTA model will give the same prior strength as the STABILIZE command did in version 1.1.5.**

**BETA PRIOR** = *value;*

This is similar to the analogous statement in PROC LCA. Stabilizing priors for other model parameters are not yet available.

**FREQ** *variable*;

This statement works the same here as in PROC LCA; see description in PROC LCA syntax.

**ESTIMATION** *estimation_method*;

This statement works the same here as in PROC LCA; see description in PROC LCA syntax.

**SEED** *value*;

This statement works the same here as in PROC LCA; see description in PROC LCA syntax. Note that either the SEED statement or the START option in the PROC LTA statement must be included.

**MAXITER** *value;*

See description in PROC LCA Syntax.

**CRITERION** *value*;

See description in PROC LCA Syntax.

**Table 2: Summary of PROC LTA Syntax**

| Statement | Description & Default (if applicable) |
|---|---|
| **PROC LTA** | Invokes the procedure. |
| **NSTATUS** | Specifies number of latent statuses. |
| **NTIMES** | Specifies the number of times. |
| **ITEMS** | Declares variables that indicate dynamic latent status variable. |
| **CATEGORIES** | Specifies number of response categories in items for all times. |
| ID | Declares identifier and other variables to retain in posterior probabilities file. |
| GROUPS | Declares categorical grouping variable. |
| GROUPNAMES | Specifies a label for each group. |
| MEASUREMENT | Invokes measurement invariance across groups and/or times. |
| COVARIATES1 | Declares variables to include as covariates for time 1. |
| COVARIATES2 | Declares variables to include as covariates for transitions. |
| REFERENCE1 | Specifies latent status to use as reference group in prediction from COVARIATES1. Default: 1 |
| REFERENCE2 | Specifies latent statuses to use as reference group in prediction from COVARIATES2. Default: diagonal elements of transition matrix |
| BINARY1 | Specifies latent status to use as comparison group in prediction from COVARIATES1, and that binary logistic regression is to be used. |
| BINARY2 | Specifies latent statuses to use as comparison group in prediction from COVARIATES2, and that binary logistic regression is to be used. |
| CORES | Divides work between multiple cores on a multiprocessor computer. Default: 1 |
| BETA PRIOR= | Invokes a stabilizing prior for the logistic regressions relating covariates to class memberships. Prior strength must be specified; as a standard we recommend BETA PRIOR=1. |
| FREQ | Identifies the frequency count variable, to use when data are aggregated. |
| ESTIMATION | Specifies estimation procedure. Default: EM. |
| SEED | Specifies seed for random number generator. * |
| MAXITER | Specifies maximum number of iterations. Default: 5000 |
| CRITERION | Specifies convergence criterion for maximum absolute deviation. Default: 0.000001 |

* SEED statement required if the START option is not included in the PROC LTA statement.

# 7 Appendices: Examples of Use

## 7.1 Appendix 1: Tutorial Example of Using PROC LCA

This example is based on simulated data which emulate the results of the adolescent delinquent behaviors example shown in Table 1.3 on page 12 of Collins and Lanza (2010). Suppose that roughly 2000 adolescents are asked whether or not they have engaged in six categories of delinquent behavior. The data is shown below. Gender is coded 1=male, 2=female. The behaviors are coded as 1=yes, 2=no. The Count column indicates how many gave each pattern of responses.

```
DATA Delinquents;
 INPUT Gender     Lie    Rowdy    Vandal    Shoplift    Steal    Fight    Count;
 Datalines;
        1          1       1         1          1          1        1        40
        1          1       1         1          1          1        2        29
        1          1       1         1          1          2        1         7
        1          1       1         1          1          2        2        14
        1          1       1         1          2          1        1         9
        1          1       1         1          2          1        2         4
        1          1       1         1          2          2        1        20
        1          1       1         1          2          2        2        25
        1          1       1         2          1          1        1        19
        1          1       1         2          1          1        2        49
        1          1       1         2          1          2        1         8
        1          1       1         2          1          2        2        11
        1          1       1         2          2          1        1         2
        1          1       1         2          2          1        2         8
        1          1       1         2          2          2        1        38
        1          1       1         2          2          2        2        88
        1          1       2         1          1          1        2        16
        1          1       2         1          1          2        1         1
        1          1       2         1          1          2        2         3
        1          1       2         1          2          1        2         1
        1          1       2         1          2          2        1         1
        1          1       2         1          2          2        2         7
        1          1       2         2          1          1        1         7
        1          1       2         2          1          1        2        37
        1          1       2         2          1          2        1         2
        1          1       2         2          1          2        2        13
        1          1       2         2          2          1        1         1
        1          1       2         2          2          1        2         1
        1          1       2         2          2          2        1        12
        1          1       2         2          2          2        2       108
        1          2       1         1          1          1        2         9
        1          2       1         1          1          2        1         1
        1          2       1         1          1          2        2         2
        1          2       1         1          2          1        1         1
        1          2       1         1          2          1        2         1
        1          2       1         1          2          2        1         6
```

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 2 | 2 | 2 | 8 |
| 1 | 2 | 1 | 2 | 1 | 1 | 1 | 4 |
| 1 | 2 | 1 | 2 | 1 | 1 | 2 | 12 |
| 1 | 2 | 1 | 2 | 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 2 | 1 | 2 | 2 | 4 |
| 1 | 2 | 1 | 2 | 2 | 1 | 1 | 2 |
| 1 | 2 | 1 | 2 | 2 | 1 | 2 | 1 |
| 1 | 2 | 1 | 2 | 2 | 2 | 1 | 11 |
| 1 | 2 | 1 | 2 | 2 | 2 | 2 | 55 |
| 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| 1 | 2 | 2 | 1 | 1 | 1 | 2 | 3 |
| 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 |
| 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 |
| 1 | 2 | 2 | 1 | 2 | 2 | 2 | 4 |
| 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 |
| 1 | 2 | 2 | 2 | 1 | 1 | 2 | 4 |
| 1 | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| 1 | 2 | 2 | 2 | 1 | 2 | 2 | 9 |
| 1 | 2 | 2 | 2 | 2 | 1 | 2 | 1 |
| 1 | 2 | 2 | 2 | 2 | 2 | 1 | 12 |
| 1 | 2 | 2 | 2 | 2 | 2 | 2 | 246 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 7 |
| 2 | 1 | 1 | 1 | 1 | 1 | 2 | 24 |
| 2 | 1 | 1 | 1 | 1 | 2 | 1 | 4 |
| 2 | 1 | 1 | 1 | 1 | 2 | 2 | 7 |
| 2 | 1 | 1 | 1 | 2 | 1 | 1 | 4 |
| 2 | 1 | 1 | 1 | 2 | 1 | 2 | 4 |
| 2 | 1 | 1 | 1 | 2 | 2 | 1 | 11 |
| 2 | 1 | 1 | 1 | 2 | 2 | 2 | 32 |
| 2 | 1 | 1 | 2 | 1 | 1 | 1 | 12 |
| 2 | 1 | 1 | 2 | 1 | 1 | 2 | 37 |
| 2 | 1 | 1 | 2 | 1 | 2 | 1 | 4 |
| 2 | 1 | 1 | 2 | 1 | 2 | 2 | 11 |
| 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 |
| 2 | 1 | 1 | 2 | 2 | 1 | 2 | 6 |
| 2 | 1 | 1 | 2 | 2 | 2 | 1 | 46 |
| 2 | 1 | 1 | 2 | 2 | 2 | 2 | 147 |
| 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 2 | 1 | 1 | 1 | 2 | 6 |
| 2 | 1 | 2 | 1 | 1 | 2 | 2 | 3 |
| 2 | 1 | 2 | 1 | 2 | 2 | 1 | 1 |
| 2 | 1 | 2 | 1 | 2 | 2 | 2 | 10 |
| 2 | 1 | 2 | 2 | 1 | 1 | 1 | 3 |
| 2 | 1 | 2 | 2 | 1 | 1 | 2 | 23 |
| 2 | 1 | 2 | 2 | 1 | 2 | 1 | 1 |
| 2 | 1 | 2 | 2 | 1 | 2 | 2 | 12 |
| 2 | 1 | 2 | 2 | 2 | 1 | 1 | 1 |
| 2 | 1 | 2 | 2 | 2 | 1 | 2 | 1 |
| 2 | 1 | 2 | 2 | 2 | 2 | 1 | 15 |
| 2 | 1 | 2 | 2 | 2 | 2 | 2 | 149 |
| 2 | 2 | 1 | 1 | 1 | 1 | 1 | 3 |
| 2 | 2 | 1 | 1 | 1 | 1 | 2 | 4 |
| 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 |
| 2 | 2 | 1 | 1 | 2 | 2 | 1 | 4 |
| 2 | 2 | 1 | 1 | 2 | 2 | 2 | 6 |
| 2 | 2 | 1 | 2 | 1 | 1 | 1 | 2 |

```
2          2      1        2          1          1          2          9
2          2      1        2          1          2          2          6
2          2      1        2          2          1          1          1
2          2      1        2          2          1          2          2
2          2      1        2          2          2          1         12
2          2      1        2          2          2          2         89
2          2      2        1          1          1          2          2
2          2      2        1          1          2          2          1
2          2      2        1          2          2          1          1
2          2      2        1          2          2          2          4
2          2      2        2          1          1          1          1
2          2      2        2          1          1          2          3
2          2      2        2          1          2          1          1
2          2      2        2          1          2          2         11
2          2      2        2          2          1          2          1
2          2      2        2          2          2          1         13
2          2      2        2          2          2          2        246

;
RUN;
```

The following code tells PROC LCA to fit a four-class model to these data, using the best of 20 random starting values generated from an initial seed of 1000 (an arbitrarily chosen number). Gender is included as a grouping variable with the assumption of measurement invariance.

```
PROC   LCA   DATA=Delinquents   OUTEST=est1   OUTPARAM=par1   OUTSTDERR=sd1
          OUTSEEDS=seeds1;
     TITLE 'Analysis of simulated delinquency data';
     NCLASS 4;
     ITEMS Lie Rowdy Vandal Shoplift Steal Fight;
     FREQ Count;
     CATEGORIES 2 2 2 2 2 2;
     GROUPS Gender;
     MEASUREMENT Group;
     SEED 1000;
     NSTARTS 20;
RUN;
```

The output suggests a well-identified solution (100% of the seeds give the same answer).

```
Seed selected for best fitted model: 292806558
Percentage of seeds associated with best fitted model: 100.00
```

However, standard errors cannot be provided because two of the rho estimates are on the boundary of the parameter space (one of them is 0 and the other is 1).

```
(Standard errors could not be computed; please see the log file for details.)
```

We can fix this problem by adding a weak but adequate (prior strength 1, which is recommended as standard) stabilizing prior on the $\rho$s.

```
 PROC    LCA    DATA=Delinquents    OUTEST=est1    OUTPARAM=par1    OUTSTDERR=sd1
            OUTSEEDS=seeds1;
      TITLE 'Analysis of simulated delinquency data';
      NCLASS 4;
      ITEMS Lie Rowdy Vandal Shoplift Steal Fight;
      FREQ Count;
      CATEGORIES 2 2 2 2 2 2;
      GROUP Gender;
      MEASUREMENT Group;
      RHO PRIOR=1;
      SEED 1000;
      NSTARTS 20;
RUN;
```

Using the stabilizing prior, the $\rho$s become much less likely to be on the boundary. In fact, in this case we can now get standard errors for all parameters.

```
Log-likelihood:    -5739.14
G-squared:           78.63
AIC:                138.63
BIC:                306.53
CAIC:               336.53
Adjusted BIC:        211.22
Entropy:              0.74
Degrees of freedom:      97


                            Parameter Estimates
Class membership probabilities: Gamma estimates (standard errors)
 Class:             1             2             3             4
  Group 1           0.4345        0.2346        0.2184        0.1125
                    (0.0266)      (0.0265)      (0.0258)      (0.0251)
  Group 2           0.4840        0.3294        0.1542        0.0325
                    (0.0350)      (0.0349)      (0.0190)      (0.0153)

Item response probabilities: Rho estimates (standard errors)
(All groups)
Response category 1:
 Class:                 1             2             3             4
   Lie:                 0.3135        0.7883        0.8156        0.9143
                        (0.0230)      (0.0325)      (0.0241)      (0.0450)
   Rowdy:               0.1581        0.8338        0.6181        0.9964
                        (0.0277)      (0.0393)      (0.0413)      (0.0101)
   Vandal:              0.0120        0.2251        0.3008        0.7440
                        (0.0070)      (0.0262)      (0.0363)      (0.0857)
   Shoplift:            0.0383        0.0096        0.9758        0.7522
                        (0.0092)      (0.0242)      (0.0278)      (0.0657)
   Steal:               0.0021        0.0383        0.7955        0.7294
                        (0.0025)      (0.0180)      (0.0401)      (0.0660)
   Fight:               0.0433        0.2938        0.1585        0.6778
                        (0.0103)      (0.0280)      (0.0322)      (0.1030)
```

Response category 2:

| Class: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Lie: | 0.6865 | 0.2117 | 0.1844 | 0.0857 |
| | (0.0230) | (0.0325) | (0.0241) | (0.0450) |
| Rowdy: | 0.8419 | 0.1662 | 0.3819 | 0.0036 |
| | (0.0277) | (0.0393) | (0.0413) | (0.0101) |
| Vandal: | 0.9880 | 0.7749 | 0.6992 | 0.2560 |
| | (0.0070) | (0.0262) | (0.0363) | (0.0857) |
| Shoplift: | 0.9617 | 0.9904 | 0.0242 | 0.2478 |
| | (0.0092) | (0.0242) | (0.0278) | (0.0657) |
| Steal: | 0.9979 | 0.9617 | 0.2045 | 0.2706 |
| | (0.0025) | (0.0180) | (0.0401) | (0.0660) |
| Fight: | 0.9567 | 0.7062 | 0.8415 | 0.3222 |
| | (0.0103) | (0.0280) | (0.0322) | (0.1030) |

## 7.2 Appendix 2: Complex Sample LCA

In this hypothetical example, 200 adolescents in a city are asked whether in the past year they have used alcohol, tobacco, marijuana, cocaine, or illegal drugs other than cocaine. The adolescents are taken via a cluster sample within 40 local schools. The data are shown below . The numbers represent cluster ID, subject ID, gender (0=male,1=female), sampling weight, and the five substances (1=yes, 2=no).

```
DATA APP2;
INPUT SchoolID SubjectID Gender SamplingWeight Alcohol Tobacco Marijuana Cocaine OtherHard @@;
DATALINES;
1 1 1 18.3 1 2 2 2 2    11 51 0 16.8 1 1 1 2 2    22 101 0 15.6 1 1 1 2 2    30 151 0 16.8 2 2 2 2 2
1 2 1 17.5 2 2 2 2 2    11 52 0 19.2 1 1 1 1 2    22 102 0 15.3 2 1 2 2 2    30 152 0 15.7 2 2 2 2 2
1 3 1 12.4 2 2 2 2 2    11 53 1 17.3 2 2 2 2 2    22 103 1 17.4 1 1 2 2 2    30 153 0 11.7 1 2 2 2 2
2 4 1 10.6 2 2 2 2 2    12 54 1 18.7 2 2 2 2 2    22 104 1 12.9 2 1 1 2 2    30 154 0 14.6 1 2 2 2 2
2 5 1 13.1 2 2 2 2 2    12 55 1 14.3 1 2 2 2 2    22 105 0 11.3 1 2 2 2 2    30 155 0 11.3 1 2 2 1 2
2 6 1 10.4 2 1 2 2 2    12 56 0 12.1 2 2 2 2 2    23 106 0 17.5 1 2 2 2 2    31 156 1 18.7 1 1 2 2 2
2 7 0 13.0 1 1 1 2 1    12 57 1 14.7 1 1 2 2 2    23 107 1 12.5 1 2 2 2 2    31 157 0 11.4 1 2 2 2 2
2 8 1 10.6 2 1 2 2 2    12 58 1 19.6 2 2 2 2 2    23 108 1 16.7 2 2 2 2 2    31 158 1 15.2 2 2 2 2 2
2 9 0 13.3 1 1 2 2 1    13 59 1 15.0 2 1 1 2 2    23 109 1 13.9 2 1 2 2 2    31 159 1 16.1 2 2 2 2 2
2 10 0 18.5 1 2 2 2 2   13 60 0 17.7 1 1 1 2 1    23 110 0 15.6 1 2 2 2 2    32 160 0 16.4 1 1 2 2 2
3 11 0 11.3 2 2 2 2 2   13 61 1 19.8 2 2 2 2 2    23 111 0 11.6 1 2 2 2 2    33 161 1 10.9 2 1 2 2 2
3 12 0 18.2 2 2 2 2 2   13 62 1 17.1 1 2 1 2 2    24 112 0 19.0 1 1 2 1 2    33 162 1 11.0 2 2 2 2 2
3 13 1 13.2 1 1 2 1 2   13 63 0 19.3 1 1 1 2 2    24 113 1 10.5 1 2 2 2 2    33 163 1 19.7 2 2 2 2 2
3 14 0 10.5 2 1 2 1 2   14 64 0 14.0 2 2 2 2 2    24 114 1 16.9 1 2 2 2 2    33 164 0 14.6 2 2 2 2 2
3 15 0 11.1 1 1 2 2 2   14 65 0 15.5 2 2 2 2 2    24 115 0 11.7 1 2 2 2 2    34 165 0 16.6 2 2 2 2 2
3 16 1 17.1 2 2 2 2 2   14 66 1 13.3 1 2 2 2 2    25 116 1 17.6 2 2 2 2 2    34 166 1 13.3 1 2 1 1 1
4 17 0 15.5 1 2 2 2 2   15 67 1 16.2 2 2 2 2 2    25 117 1 19.8 2 2 2 2 2    34 167 0 18.7 1 1 2 2 2
4 18 0 19.3 2 1 1 2 2   15 68 0 19.6 2 1 2 2 2    25 118 1 13.9 1 2 2 2 2    34 168 0 11.9 1 2 2 2 2
5 19 0 18.5 1 2 2 2 2   15 69 1 16.2 1 2 2 2 2    25 119 1 12.6 1 1 2 2 2    34 169 1 18.9 2 2 2 2 2
6 20 0 15.6 2 2 2 2 2   15 70 0 13.6 2 2 2 2 2    25 120 0 10.5 2 2 2 2 2    34 170 1 15.9 2 2 2 2 2
6 21 1 12.8 1 1 2 2 2   15 71 1 10.8 2 2 2 2 2    25 121 0 19.8 1 1 2 2 2    34 171 1 15.2 2 2 2 2 2
6 22 1 16.4 2 2 2 2 2   16 72 0 11.4 1 2 2 1 2    26 122 0 17.8 2 2 2 2 2    35 172 0 10.6 2 2 2 2 2
6 23 1 19.7 2 2 2 2 2   16 73 1 12.8 1 2 2 2 2    26 123 1 10.4 1 2 2 2 2    35 173 0 16.1 2 2 2 2 2
6 24 1 14.0 2 2 2 2 2   16 74 1 11.7 2 2 1 2 2    26 124 1 16.0 2 2 2 2 2    36 174 1 17.4 2 2 2 2 2
6 25 0 15.0 1 2 2 2 2   16 75 0 17.5 1 2 2 2 2    26 125 1 11.1 1 2 2 2 2    36 175 1 14.3 2 1 2 2 2
7 26 1 11.9 1 1 2 2 2   17 76 0 16.1 2 2 2 2 2    27 126 0 16.0 1 2 2 2 2    36 176 0 18.7 1 1 2 2 2
7 27 0 14.8 1 2 2 2 2   17 77 1 14.0 1 2 2 2 2    27 127 0 15.6 1 2 2 2 2    36 177 1 11.0 2 2 2 2 2
7 28 0 16.7 1 1 1 2 2   17 78 1 17.3 2 2 2 2 2    27 128 0 12.1 2 2 2 2 2    36 178 1 13.6 1 2 2 2 2
7 29 1 19.6 1 2 2 2 2   17 79 1 15.8 2 2 2 2 2    27 129 0 14.2 1 2 2 2 2    37 179 0 14.1 2 1 2 2 2
7 30 0 15.3 2 1 2 2 2   18 80 1 18.0 1 1 2 2 2    27 130 1 11.2 2 2 2 2 2    37 180 1 12.4 2 2 2 2 2
7 31 1 16.3 2 2 2 2 2   18 81 0 12.4 2 2 2 2 2    27 131 1 19.2 1 1 2 2 2    37 181 1 17.0 1 2 2 2 2
7 32 0 11.2 2 2 2 2 2   18 82 1 13.0 1 2 2 2 2    27 132 0 15.0 1 2 2 2 2    37 182 0 17.3 2 2 2 2 2
7 33 1 19.0 2 2 2 2 2   18 83 1 14.6 1 2 2 2 2    27 133 0 16.1 2 2 1 2 2    37 183 1 17.7 2 1 2 2 2
8 34 0 16.1 1 1 2 2 2   18 84 0 11.6 1 2 2 2 2    27 134 0 18.9 1 2 2 2 2    37 184 0 14.0 1 2 2 2 2
8 35 0 18.7 1 1 2 2 1   19 85 0 17.0 2 2 2 2 1    27 135 1 15.9 2 2 2 2 2    37 185 1 13.3 2 1 2 1 2
8 36 0 13.1 1 1 2 2 2   19 86 0 10.6 1 2 2 2 2    27 136 0 12.2 2 2 2 2 2    38 186 1 16.0 2 2 2 2 2
8 37 1 15.4 2 2 2 2 2   19 87 1 13.0 2 2 2 2 2    28 137 0 16.1 1 2 2 1 2    38 187 1 16.1 1 1 2 2 2
9 38 1 12.0 2 2 2 2 2   19 88 1 11.1 2 2 2 2 2    28 138 0 11.2 2 1 1 2 2    38 188 1 18.2 1 2 2 2 2
9 39 0 17.3 2 2 2 2 2   19 89 1 13.9 2 2 2 2 2    28 139 0 13.3 1 1 2 2 1    38 189 1 10.1 1 2 2 2 2
```

```
10 40 1 16.5 2 2 1 2 2   19 90 0 11.8 1 2 2 2 2   28 140 0 16.6 2 1 2 2 2   38 190 0 12.8 1 2 2 2 2
10 41 1 12.1 1 1 2 2 2   20 91 1 12.6 1 2 2 2 2   28 141 1 16.9 1 2 2 2 2   38 191 0 17.8 1 2 2 2 2
10 42 1 16.1 2 2 2 2 2   20 92 0 18.0 2 1 2 2 2   29 142 0 11.4 1 1 2 2 2   39 192 0 14.2 1 2 2 2 2
10 43 1 11.0 1 1 2 2 2   21 93 0 19.2 1 1 2 2 1   29 143 0 19.7 1 2 2 2 1   39 193 1 12.4 2 2 2 2 2
10 44 1 18.6 2 1 2 2 2   21 94 1 17.4 1 2 2 2 2   29 144 0 19.6 2 2 2 2 2   39 194 0 11.2 1 1 1 2 2
10 45 0 13.9 1 2 2 2 2   21 95 0 16.6 1 1 2 2 2   29 145 0 12.0 1 1 2 2 2   39 195 1 13.8 2 2 2 2 2
10 46 1 10.3 2 2 2 2 2   21 96 1 11.8 2 2 1 2 2   29 146 0 13.7 1 2 2 2 2   39 196 0 10.3 1 2 1 2 2
10 47 0 18.2 1 1 2 2 2   21 97 0 11.7 1 2 2 2 2   29 147 0 12.0 1 2 2 2 2   39 197 1 18.7 2 1 2 2 2
10 48 1 16.5 2 1 2 2 2   22 98 1 14.3 2 2 2 2 2   30 148 1 19.9 1 2 2 2 2   40 198 0 10.7 2 1 2 2 2
11 49 1 12.6 1 1 2 1 2   22 99 1 12.1 1 2 1 2 2   30 149 1 14.3 2 2 2 2 2   40 199 0 16.9 1 1 2 2 2
11 50 1 10.1 1 2 2 1 1   22 100 1 19.3 1 1 2 2 2  30 150 0 11.2 1 1 2 2 2   40 200 0 14.8 2 1 2 2 2
;
RUN;
```

First we fit a no-covariates model using the following syntax.

```
PROC LCA DATA=APP2 OUTPARAM=param1 OUTPOST=post1 OUTEST=est1
     OUTSTDERR=stderr1;
NCLASS 3;
ID SubjectID;
ITEMS Alcohol Tobacco Marijuana Cocaine OtherHard;
CATEGORIES 2 2 2 2 2;
WEIGHT SamplingWeight;
CLUSTERS SchoolID;
NSTARTS 50;
RHO PRIOR=1;
SEED 1000;
RUN;
```

We get the following output.

```
Log-likelihood:     -390.96
G-squared:           10.99
AIC:                44.99
BIC:               101.06
CAIC:              118.06
Adjusted BIC:        47.20
Entropy:             0.61
Degrees of freedom:      14
(Based on the pseudo-likelihood incorporating weights.)

                        Parameter Estimates

Class membership probabilities: Gamma estimates (standard errors)
Class:                   1         2         3
                    0.1811    0.2054    0.6136
                   (0.1446)  (0.1304)  (0.0524)

Item response probabilities: Rho estimates (standard errors)
  Response category  1:
Class:                   1         2         3
  Alcohol     :      0.9790    0.4498    0.3937
```

```
                  (0.0076)   (0.3286)   (0.0456)
  Tobacco    :      0.6733     0.8432     0.0440
                  (0.1145)   (0.1085)   (0.0330)
  Marijuana  :      0.2247     0.1784     0.0362
                  (0.1394)   (0.0909)   (0.0182)
  Cocaine    :      0.2014     0.0658     0.0004
                  (0.1324)   (0.0492)   (0.0003)
  OtherHard  :      0.2467     0.0010     0.0117
                  (0.1827)   (0.0004)   (0.0117)

  Response category  2:
Class:                   1          2          3
  Alcohol    :      0.0210     0.5502     0.6063
                  (0.0076)   (0.3286)   (0.0456)
  Tobacco    :      0.3267     0.1568     0.9560
                  (0.1145)   (0.1085)   (0.0330)
  Marijuana  :      0.7753     0.8216     0.9638
                  (0.1394)   (0.0909)   (0.0182)
  Cocaine    :      0.7986     0.9342     0.9996
                  (0.1324)   (0.0492)   (0.0003)
  OtherHard  :      0.7533     0.9990     0.9883
                  (0.1827)   (0.0004)   (0.0117)
```

We can also test the relationship of class membership to a covariate such as gender using the following code.

```
PROC LCA DATA=APP2 OUTPARAM=param1 OUTPOST=post1 OUTEST=est1
     OUTSTDERR=stderr1;
NCLASS 3;
ID SubjectID;
ITEMS Alcohol Tobacco Marijuana Cocaine OtherHard;
CATEGORIES 2 2 2 2 2;
WEIGHT SamplingWeight;
CLUSTERS SchoolID;
COVARIATES Gender;
NSTARTS 50;
BETA PRIOR=1;
RHO PRIOR=1;
SEED 2028342676;
RUN;
```

## 7.3 Appendix 3: Minimal PROC LCA Call for Aggregated Data

In this example, the FREQ command is used in order to analyze aggregated (summarized) data.

HIV diagnosis

- two latent classes
- four dichotomous indicators
- aggregated data
- random starting values

```
/*************************************************************************
* Data from Yang and Becker (Biometrics Vol. 53, No. 3, Sept. 1997)      *
*************************************************************************/
TITLE1 'Example 1: HIV Diagnosis';
*Create SAS data file containing data aggregated by response pattern;
*Data contain four indicators of latent class, plus frequency count;
*The indicators are results of four diagnostic tests;
DATA HIV;
INPUT test1 test2 test3 test4 count;
DATALINES;
1 1 1 1 170
1 1 1 2 15
1 2 1 1 6
2 1 1 1 4
2 1 1 2 17
2 1 2 2 83
2 2 1 1 1
2 2 1 2 4
2 2 2 2 128
;
RUN;
*Specify latent class model with two latent classes of diagnosis;
*Data are aggregated, requiring FREQ option;
PROC LCA DATA=HIV;
      TITLE2 'Two-class model of HIV diagnosis';
      NCLASS 2;
      ITEMS test1 test2 test3 test4;
      CATEGORIES 2 2 2 2;
      FREQ count;
      SEED 518165;
RUN;
```

## 7.4 Appendix 4: LCA With User-Provided Starting Values and Parameter Restrictions

Math skills

- five latent classes
- four binary indicators
- aggregated data
- starting values provided in SAS data file
- parameter restrictions provided in SAS data file
- verbose output requested

```
/*************************************************************
*(Example 1 from WinLTA examples -- for documentation, see *
*http://methodology.psu.edu/downloads/winlta)*
*************************************************************/
TITLE1 'Example 2: Math Skills';
*Create SAS data file containing data aggregated by response pattern;
*Data contain four indicators of latent class, plus frequency count;
DATA math;
INPUT addition subtract multipli division count;
DATALINES;
1 1 1 1 408
1 1 1 2 2
1 1 2 1 12
1 1 2 2 1
1 2 1 1 59
1 2 2 1 19
2 1 1 1 417
2 1 1 2 7
2 1 2 1 43
2 1 2 2 3
2 2 1 1 248
2 2 1 2 10
2 2 2 1 204
2 2 2 2 67
;
RUN;
*Create SAS data file containing starting values;
DATA math_start;
INPUT param $ group variable $ respcat estlc1 estlc2 estlc3 estlc4
estlc5;
DATALINES;
 GAMMA  1   .         .   0.2  0.2  0.2  0.2  0.2
 BETA   1   .         .   0.0  0.0  0.0  0.0  0.0
 RHO    1   ADDITION  1   0.8  0.2  0.2  0.2  0.2
 RHO    1   SUBTRACT  1   0.8  0.8  0.2  0.2  0.2
 RHO    1   MULTIPLI  1   0.8  0.8  0.8  0.2  0.2
 RHO    1   DIVISION  1   0.8  0.8  0.8  0.8  0.2
```

```
 RHO    1   ADDITION  2   0.2  0.8  0.8  0.8  0.8
 RHO    1   SUBTRACT  2   0.2  0.2  0.8  0.8  0.8
 RHO    1   MULTIPLI  2   0.2  0.2  0.2  0.8  0.8
 RHO    1   DIVISION  2   0.2  0.2  0.2  0.2  0.8
;
RUN;
*Create SAS data file containing parameter restrictions;
DATA math_restr;
INPUT param $ group variable $ respcat estlc1 estlc2 estlc3 estlc4
estlc5;
DATALINES;
 GAMMA  1   .         .  1   1   1   1   1
 BETA   1   .         .  1   1   1   1   1
 RHO    1   ADDITION  1  2   3   3   3   3
 RHO    1   SUBTRACT  1  2   2   3   3   3
 RHO    1   MULTIPLI  1  2   2   2   3   3
 RHO    1   DIVISION  1  2   2   2   2   3
 RHO    1   ADDITION  2  4   5   5   5   5
 RHO    1   SUBTRACT  2  4   4   5   5   5
 RHO    1   MULTIPLI  2  4   4   4   5   5
 RHO    1   DIVISION  2  4   4   4   4   5
;
RUN;
*Specify latent class model with five categories of math skills;
*Data are aggregated, requiring FREQ option;
PROC LCA DATA=math OUTEST=math_out START=math_start
RESTRICT=math_restr VERBOSE_OUTPUT;
      TITLE2 'Five-class model (Restricted, starting values provided)';
      NCLASS 5;
      ITEMS addition subtract multipli division;
      CATEGORIES 2 2 2 2;
      FREQ count;
      ESTIMATION EM;
      MAXITER 5000;
      CRITERION 0.000001;
RUN;
```

## 7.5 Appendix 5: LCA with Individual-Level Data, Grouping Variable and Covariate

Abortion attitudes

- three latent classes
- six binary indicators
- individual-level data
- grouping variable: sex
- measurement invariance across groups imposed
- one covariate: age
- latent class 1 specified as reference group
- random starting values
- posterior probabilities saved to SAS data file
- parameter estimates saved to SAS data file (1 record)
- parameter estimates saved to SAS data file (familiar format)

```
PROC LCA DATA=ABORTION OUTPOST=ABOR_PP OUTEST=ABOR_EST
         OUTPARAM=ABOR_PARAM;
     TITLE1 'Three-class model with two groups and a covariate (age)';
     TITLE2 'Measurement invariance across groups';
     ID caseid;
     NCLASS 3;
     ITEMS x1 x2 x3 x4 x5 x6;
     CATEGORIES 2 2 2 2 2 2;
     GROUPS sex;
     GROUPNAMES male female;
     MEASUREMENT GROUPS;
     COVARIATES age;
     REFERENCE 1;
     NSTARTS 20;
RUN;
```

## 7.6 Appendix 6: LTA With User-Provided Starting Values and Parameter Restrictions

Math ability over time

- five latent statuses
- four binary indicators
- two times
- aggregated data
- starting values and parameter restrictions provided in SAS data file

```
/* Data from WinLTA example 2 */
/* Aggregated by response pattern */
DATA MATH_CHG_AGG;
INPUT ADDI_T1 SUBT_T1 MULT_T1 DIVI_T1 ADDI_T2 SUBT_T2 MULT_T2 DIVI_T2
      COUNTS;
      DATALINES;
       1  1  1  1  1  1  1  1  217
       1  1  1  1  1  1  1  2  1
       1  1  1  1  1  1  2  1  2
       1  1  1  1  1  2  1  1  14
       1  1  1  1  1  2  1  2  1
       1  1  1  1  1  2  2  1  1
       1  1  1  1  2  1  1  1  133
       1  1  1  1  2  1  2  1  8
       1  1  1  1  2  2  1  1  24
       1  1  1  1  2  2  1  2  1
       1  1  1  1  2  2  2  1  6
       1  1  1  2  1  2  1  1  1
       1  1  1  2  2  1  1  1  1
       1  1  2  1  1  1  1  1  1
       1  1  2  1  1  1  2  1  1
       1  1  2  1  1  2  1  1  1
       1  1  2  1  2  1  1  1  1
       1  1  2  1  2  1  2  1  1
       1  1  2  1  2  2  1  1  1
       1  1  2  1  2  2  2  1  6
       1  1  2  2  2  2  1  1  1
       1  2  1  1  1  1  1  1  11
       1  2  1  1  1  2  1  1  2
       1  2  1  1  1  2  2  1  4
       1  2  1  1  2  1  1  1  17
       1  2  1  1  2  1  2  1  4
       1  2  1  1  2  2  1  1  12
       1  2  1  1  2  2  2  1  8
       1  2  1  1  2  2  2  2  1
       1  2  2  1  1  1  1  1  1
```

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 |
| 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |
| 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 |
| 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 4 |
| 1 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 11 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 100 |
| 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 4 |
| 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 17 |
| 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 3 |
| 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 |
| 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 164 |
| 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 |
| 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 22 |
| 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 63 |
| 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 33 |
| 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 7 |
| 2 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 2 |
| 2 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 3 |
| 2 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 |
| 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 |
| 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 5 |
| 2 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |
| 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 8 |
| 2 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 6 |
| 2 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 20 |
| 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 3 |
| 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 3 |
| 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 10 |
| 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 9 |
| 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 2 |
| 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 |
| 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 40 |
| 2 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 14 |
| 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 73 |
| 2 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 |
| 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 84 |
| 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 12 |
| 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 2 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 3 |
| 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 1 |
| 2 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 3 |
| 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 |
| 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 5 |
| 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 |
| 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 3 |
| 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |

```
      2  2  2  1  2  1  1  1  7
      2  2  2  1  2  1  2  1  12
      2  2  2  1  2  2  1  1  32
      2  2  2  1  2  2  1  2  1
      2  2  2  1  2  2  2  1  90
      2  2  2  1  2  2  2  2  52
      2  2  2  2  1  2  1  1  1
      2  2  2  2  1  2  2  1  1
      2  2  2  2  2  2  1  1  3
      2  2  2  2  2  2  1  2  1
      2  2  2  2  2  2  2  1  17
      2  2  2  2  2  2  2  2  44
;
RUN;
/* Starting values */
DATA MATH_CHG_START;
INPUT PARAM $ GROUP VARIABLE $ TIME STATUS RESPCAT ESTLS1 ESTLS2 ESTLS3
     ESTLS4 ESTLS5;
DATALINES;
 delta      1  .                1  .  .   0.4000  0.3000  0.1000  0.1000  0.1000
 tau        1  .                1  1  .   0.6000  0.1000  0.1000  0.1000  0.1000
 tau        1  .                1  2  .   0.0000  0.6000  0.2000  0.1000  0.1000
 tau        1  .                1  3  .   0.0000  0.0000  0.6000  0.2000  0.2000
 tau        1  .                1  4  .   0.0000  0.0000  0.0000  0.6000  0.4000
 tau        1  .                1  5  .   0.0000  0.0000  0.0000  0.0000  1.0000
 rho        1  ADDITION         1  .  1   0.8000  0.2000  0.2000  0.2000  0.2000
 rho        1  SUBTRACTION      1  .  1   0.8000  0.8000  0.2000  0.2000  0.2000
 rho        1  MULTIPLICATION   1  .  1   0.8000  0.8000  0.8000  0.2000  0.2000
 rho        1  DIVISION         1  .  1   0.8000  0.8000  0.8000  0.8000  0.2000
 rho        1  ADDITION         1  .  2   0.2000  0.8000  0.8000  0.8000  0.8000
 rho        1  SUBTRACTION      1  .  2   0.2000  0.2000  0.8000  0.8000  0.8000
 rho        1  MULTIPLICATION   1  .  2   0.2000  0.2000  0.2000  0.8000  0.8000
 rho        1  DIVISION         1  .  2   0.2000  0.2000  0.2000  0.2000  0.8000
 rho        1  ADDITION         2  .  1   0.8000  0.2000  0.2000  0.2000  0.2000
 rho        1  SU               2  .  1   0.8000  0.8000  0.2000  0.2000  0.2000
 rho        1  MULTIPLICATION   2  .  1   0.8000  0.8000  0.8000  0.2000  0.2000
```

```
rho      1  DIVISION          2  .  1   0.8000  0.8000  0.8000  0.8000  0.2000
rho      1  ADDITION          2  .  2   0.2000  0.8000  0.8000  0.8000  0.8000
rho      1  SUBTRACTION       2  .  2   0.2000  0.2000  0.8000  0.8000  0.8000
rho      1  MULTIPLICATION    2  .  2   0.2000  0.2000  0.2000  0.8000  0.8000
rho      1  DIVISION          2  .  2   0.2000  0.2000  0.2000  0.2000  0.8000
```

**Delta** row represents the latent status membership probabilities at Time 1.
**Tau** rows represent the probability of going into each status if you are in the status listed in the 5th column.
**Rho** rows represent the probability of the response listed in the 6th column for each item for each status and the time listed in the 4th column.

```
/* Parameter restrictions */
DATA MATH_CHG_RESTR;
INPUT PARAM $ GROUP VARIABLE $ TIME STATUS RESPCAT ESTLS1 ESTLS2 ESTLS3
ESTLS4 ESTLS5;
DATALINES;
      delta  1    .         1    .    .    1    1    1    1    1
      tau    1    .         1    1    .    1    1    1    1    1
      tau    1    .         1    2    .    0    1    1    1    1
      tau    1    .         1    3    .    0    0    1    1    1
      tau    1    .         1    4    .    0    0    0    1    1
      tau    1    .         1    5    .    0    0    0    0    1
      rho    1    ADDITION  1    .    1    2    3    3    3    3
      rho    1    SUBTRACT  1    .    1    12   12   13   13   13
      rho    1    MULTIPLI  1    .    1    22   22   22   23   23
      rho    1    DIVISION  1    .    1    32   32   32   32   33
      rho    1    ADDITION  1    .    2    4    5    5    5    5
      rho    1    SUBTRACT  1    .    2    14   14   15   15   15
      rho    1    MULTIPLI  1    .    2    24   24   24   25   25
      rho    1    DIVISION  1    .    2    34   34   34   34   35
      rho    1    ADDITION  2    .    1    2    3    3    3    3
      rho    1    SUBTRACT  2    .    1    12   12   13   13   13
      rho    1    MULTIPLI  2    .    1    22   22   22   23   23
      rho    1    DIVISION  2    .    1    32   32   32   32   33
      rho    1    ADDITION  2    .    2    4    5    5    5    5
      rho    1    SUBTRACT  2    .    2    14   14   15   15   15
      rho    1    MULTIPLI  2    .    2    24   24   24   25   25
      rho    1    DIVISION  2    .    2    34   34   34   34   35
; RUN;
```

```
PROC LTA DATA=MATH_CHG_AGG START=MATH_CHG_START
          RESTRICT=MATH_CHG_RESTR VERBOSE_OUTPUT;
     TITLE1 '2 times, 5 statuses, aggregated, no missing data, starts,
     restrict';
     NSTATUS 5;
     NTIMES 2;
     ITEMS  ADDI_T1  SUBT_T1  MULT_T1  DIVI_T1  ADDI_T2  SUBT_T2  MULT_T2
     DIVI_T2;
     FREQ COUNTS;
     CATEGORIES 2 2 2 2;
RUN;
```

## 7.7 Appendix 7: LTA With Measurement Invariance Across Times

Sexual risk behavior over time

- five latent statuses
- three times
- four indicators
- individual-level data
- measurement invariance across times imposed
- no covariates
- random starting values
- posterior probabilities saved to SAS data file
- parameter estimates saved to SAS data file (familiar format)
- verbose output requested

```
PROC LTA DATA=NLSY_RECODED OUTPOST=NLSY_POST OUTPARAM=NLSY_PARAM
          VERBOSE_OUTPUT;
     TITLE1 'NLSY dating and sexual risk behavior';
     TITLE2 '3 times, no covariates';
     NSTATUS 5;
     NTIMES 3;
     ITEMS datepar_alt_98 sex_yr_98 part_98 expos_98
           datepar_alt_99 sex_yr_99 part_99 expos_99
           datepar_alt_00 sex_yr_00 part_00 expos_00;
     CATEGORIES 3 2 3 2;
     ID CASEID;
     MEASUREMENT TIMES;
     SEED 592667;
RUN;
```

## 7.8 Appendix 8: LTA With Time 1 Covariates

Sexual risk behavior over time

- five latent statuses
- three times
- four indicators
- individual-level data
- grouped by gender
- measurement invariance across times and groups imposed
- three covariates for Time 1
- reference category is 4
- user-provided starting values
- posterior probabilities saved to SAS data file
- parameter estimates saved to SAS data file (familiar format)
- verbose output requested

```
 PROC LTA DATA=NLSY_RECODED START=NLSY_START OUTPOST=NLSY_POST
          OUTPARAM=NLSY_PARAM VERBOSE_OUTPUT;
     TITLE1 'NLSY dating and sexual risk behavior';
     TITLE2 '3 times, covariates1 (cig,drunk,mar), groups (male,
female)';
     NSTATUS 5;
     NTIMES 3;
     ITEMS datepar_alt_98 sex_yr_98 part_98 expos_98
           datepar_alt_99 sex_yr_99 part_99 expos_99
           datepar_alt_00 sex_yr_00 part_00 expos_00;
     CATEGORIES 3 2 3 2;
     COVARIATES1 cig_yr_98 drunk_98 mar_yr_98;
     REFERENCE1 4;
     ID CASEID;
     GROUPS gender;
     GROUPNAMES male female;
     MEASUREMENT TIMES GROUPS;
RUN;
```

## 7.9 Appendix 9: LTA With Time 1 and Transition Covariates

Sexual risk behavior over time

- five latent statuses
- three times
- four indicators
- individual-level data
- measurement invariance across times imposed
- one covariate for Time 1
- reference category for Time 1 is 4
- one covariate for each transition
- binary logistic regression for transitions
- user-provided starting values
- posterior probabilities saved to SAS data file
- parameter estimates saved to SAS data file (familiar format)
- verbose output requested

```
 PROC LTA DATA=NLSY_RECODED START=NLSY_START OUTPOST=NLSY_POST
          OUTPARAM=NLSY_PARAM VERBOSE_OUTPUT;
      TITLE1 'NLSY dating and sexual risk behavior';
      TITLE2 '3 times, covariates1 and covariates2 (drunk)';
      NSTATUS 5;
      NTIMES 3;
      ITEMS datepar_alt_98 sex_yr_98 part_98 expos_98
            datepar_alt_99 sex_yr_99 part_99 expos_99
            datepar_alt_00 sex_yr_00 part_00 expos_00;
      CATEGORIES 3 2 3 2;
      COVARIATES1 drunk_98;
      COVARIATES2 drunk_98 drunk_98;
      REFERENCE1 4;
      ID CASEID;
      BINARY2 0 5 5 0 5 0 5 5 0 5;
      MEASUREMENT TIMES;
RUN;
```

# References


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest, Hungary: Akademiai Kiado.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411-434.

Asparouhov, T. and Muthén, B. (2005) *Multivariate statistical modeling with survey data.* Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference. Washington, DC: Office of Management and Budget.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association, 92* (440), 1375-1386.

Bozdogan, H., (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

Bray, B. C., Lanza, S. T., & Tan, X. (2012) *An introduction to eliminating bias in classify-analyze approaches for latent class analysis* (Technical Report No. 12-118). University Park: The Methodology Center, Penn State. methodology.psu.edu/media/techreports/12-118.pdf

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association, 86,* 68-78.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York, NY: Wiley.

Lanza, S. T., & Collins, L. M. (2008). A new SAS procedure for latent transition analysis: Transitions in dating and sexual risk behavior. *Developmental Psychology, 44*(2), 446-456.

Lanza, S. T., Collins, L. M., Lemmon, D., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling, 14*(4), 671-694.

Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). Latent class and latent transition analysis. In J. A. Schinka, W. F. Velicer & I. B. Weiner (Eds.), *Handbook of psychology* (2nd ed., Vol. 2, pp. 691-716). Hoboken, NJ: Wiley.

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

Muthén, B. O. (2004). *Mplus technical appendices.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. and Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Ramaswamy, V., Desarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12,* 103-124.

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 308-313.

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

Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate

analysis. *Psychometrika, 52,* 333-343.
Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.) (1989). *Analysis of complex surveys.* New York, NY: Wiley.
Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, and T. M. F. Smith, (Eds.), *Analysis of complex surveys.* New York, NY: Wiley.
Vermunt, J. K., and Magidson, J. (2005a). *Latent GOLD 4.0 user's guide.* Belmont, Massachusetts: Statistical Innovations, Inc.
Vermunt, J. K., and Magidson, J. (2005b). *Technical guide for Latent GOLD 4.0: Basic and advanced.* Belmont, MA: Statistical Innovations Inc.
Vermunt, J.K., & Magidson, J. (2007). Latent class analysis with sampling weights: A maximum likelihood approach. *Sociological Methods and Research, 36*(1), 87-111.
Wang, C., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association, 100*(471), 1054-1076.
Wedel, Michel, Frenkel Ter Hofstede and Jan-Benedict E.M. Steenkamp (1998), Mixture Model Analysis of Complex Samples, *Journal of Classification 15* (2), 225-244.

INDEX