Published in final edited form as:
Struct Equ Modeling. 2007 ; 14(4): 671–694.

PROC LCA: A SAS Procedure for Latent Class Analysis

Stephanie T. Lanza, **Linda M. Collins**, **David R. Lemmon**, and **Joseph L. Schafer**
The Methodology Center, The Pennsylvania State University

Abstract

Latent class analysis (LCA) is a statistical method used to identify a set of discrete, mutually exclusive latent classes of individuals based on their responses to a set of observed categorical variables. In multiple-group LCA, both the measurement part and structural part of the model can vary across groups, and measurement invariance across groups can be empirically tested. LCA with covariates extends the model to include predictors of class membership. In this article, we introduce PROC LCA, a new SAS procedure for conducting LCA, multiple-group LCA, and LCA with covariates. The procedure is demonstrated using data on alcohol use behavior in a national sample of high school seniors.

In the social and behavioral sciences, it is useful to regard many constructs as latent variables. These variables cannot be observed directly and instead must be inferred from multiple observed items. Covariance structure analysis (e.g., the factor model) provides a popular framework for mapping items onto continuous latent variables. Latent class analysis (LCA) provides an analogous framework for measuring categorical latent variables. Whereas the factor model characterizes the latent variable with a continuous (e.g., normal) distribution, the latent class model divides a population into mutually exclusive and exhaustive subgroups (Goodman, 1974; Lazarsfeld & Henry, 1968).

LATENT CLASS ANALYSIS

In traditional LCA, two sets of parameters are estimated: class membership probabilities and item-response probabilities conditional on class membership. Because it is a measurement model, the latent class model estimates and removes measurement error from the vector of latent-class membership probabilities.

Latent class models usually involve categorical indicators (although a version of LCA involving continuous indicators called latent profile analysis [Gibson, 1959] is being used increasingly). When categorical data are used, the latent class model has the advantage of making no assumptions about the distributions of the indicators other than that of local independence; that is, the assumption that within a latent class the indicators are independent. Local independence is directly analogous to the assumption of uncorrelated uniquenesses often

Correspondence should be addressed to Stephanie T. Lanza, The Methodology Center, The Pennsylvania State University, 204 East Calder Way, Suite 400, State College, PA 16801. SLanza@psu.edu.

made in factor analysis. Dependence between indicators in the overall sample is expected; it is assumed that the latent class variable will account for these interrelations.

The latent class model has been applied in many domains. For example, in psychology, the approach has been used to assess temperament (Stern, Arcus, Kagan, Rubin, & Snidman, 1995) and depression (Lanza, Flaherty, & Collins, 2003). In education, teaching style has been modeled using LCA (e.g., Aitkin, Anderson, & Hinde, 1981). In sociology, this approach has been used to express poverty as a multidimensional construct (Dewilde, 2004). Recently LCA (and latent transition analysis, an extension of LCA to repeated measures) has been used increasingly often as a multivariate approach to behavioral research. For example, multidimensional alcohol use including features such as frequency of use, quantity of use, and the existence of heavy episodic use has been modeled by Jackson, Sher, Gotham, and Wood (2001) and Auerbach and Collins (2006). Other aspects of substance use behavior have been modeled by Velicer, Martin, and Collins (1996), Chung, Park, and Lanza (2005), Lanza and Collins (2002, 2006), and Chung, Flaherty, and Schafer (2006).

Two particularly useful extensions of LCA are multiple-group LCA and LCA with covariates. In multiple-group LCA, both the class membership and item-response probabilities can vary across groups, and measurement invariance across groups can be empirically tested. Clogg and Goodman (1984) were the first to introduce a latent class model in which class membership probabilities and item-response probabilities are conditioned on membership in an observed group. An example of sex differences in adolescent depression latent classes appears in Lanza et al. (2003). LCA with covariates extends the model to include predictors of class membership; latent class membership probabilities are predicted by covariates through a logistic link (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Dayton & Macready, 1988). An example of this model appears in Chung et al. (2006), in which latent classes characterized by patterns of marijuana use and attitudes were regressed on year to show historical changes in the prevalence of classes over time.

Software for LCA includes Latent GOLD (Vermunt & Magidson, 2000), PanMark (van de Pol, Langeheine, & de Jong, 1998), WinLTA (Collins, Lanza, Schafer, & Flaherty, 2002), and M*plus* (Muthén & Muthén, 1998–2006). This article introduces PROC LCA, a new SAS procedure for latent class analysis developed for SAS Version 9.1 for Windows.[1] The software is available for download free of charge at http://methodology.psu.edu. This article reviews features of the software and illustrates its use with a series of empirical analyses.

MODEL SPECIFICATION

The following sets of parameters are estimated in the traditional latent class model: γ (gamma) parameters, which represent latent class membership probabilities, and ρ (rho) parameters, which are item-response probabilities conditional on latent class membership. The ρ parameters express the correspondence between the observed items and the latent classes.

If a grouping variable is included, both sets of parameters (γ, ρ) can be conditioned on group. When one or more covariates are included, an additional set of parameters is estimated: β (beta) parameters are logistic regression coefficients for covariates, predicting class membership. When covariates are included, only ρ and β parameters are actually estimated, although the γ parameters are still of interest. The γ parameters are calculated as functions of β and the covariates, and are provided in PROC LCA output. If a grouping variable is included, all parameters are conditioned on group.

[1] Copyright 2002–2003 SAS Institute, Inc. SAS and all other SAS Institute, Inc. product or service names are registered trademarks or trademarks of SAS Institute, Inc., Cary, NC, USA.

Suppose a latent class model with C classes is to be estimated based on a data set including m categorical items, a covariate x, and a grouping variable g. Let $Y_i = (Y_{i1}, \ldots, Y_{iM})$ represent the vector of individual i 's responses to the M items where $Y_{im} = 1, 2, \ldots, r_m$. Let $c_i = 1, 2, \ldots, C$ be the latent class membership of individual i and let $I(y = k)$ be the indicator function that equals 1 if response y equals k and 0 otherwise. Suppose also that g_i represents the value of individual i's group membership, x_i represents the value of the covariate for individual i, and its value can relate to the probability of membership in each latent class, γ. Then the latent class model can be expressed as:

$$P(Y=y|x_i, g)=\sum_{c=1}^{C}\gamma_{c|g}(x_i)\prod_{m=1}^{M}\prod_{k=1}^{r_m}\rho_{mk|cg}^{I(y_{im}=k)} \tag{1}$$

where $\gamma_{c|g}(x_i) = P(C_i = c|x_i, G_i = g)$ is a standard baseline-category multinomial logistic model (Agresti, 2002). For example, with one covariate x the γ parameters are expressed as:

$$\gamma_{c|g}(x_i)=P(C_i=c|x_i, G_i=g)=\frac{\exp\{\beta_{0c|g}+x_i\beta_{1c|g}\}}{1+\sum_{j=1}^{C-1}\exp\{\beta_{0j|g}+x_i\beta_{1j|g}\}} \tag{2}$$

for $c = 1, \ldots, C-1$ with class C as the reference class in the logistic regression. This enables estimation of the log-odds that an individual falls in latent class c relative to reference class C. For example, if Class 2 is the reference class, the log-odds of membership in Class 1 relative to Class 2 for an individual in Group 1 with value x_i on the covariate is:

$$\log\left(\frac{\gamma_{1|1}(x_i)}{\gamma_{2|1}(x_i)}\right)=\beta_{01|1}+\beta_{11|1}x_i. \tag{3}$$

The exponentiated β parameter corresponding to the covariate is an odds ratio, reflecting the increase in odds of class membership (relative to reference class C) corresponding to a one-unit increase in the covariate. Note that multiple covariates can be included simultaneously.

Because class membership probabilities are modeled as functions of the covariates (see Equation 2), and individuals vary with respect to their covariates, there is a vector of estimated class membership probabilities corresponding to each individual (or group of individuals with the same responses to the covariates). The prevalence of each latent class is calculated as the average across participant-specific class membership probabilities.

In PROC LCA, parameters are estimated by maximum likelihood using an EM (expectation-maximization) type procedure. Missing data on the latent class indicators are handled in this procedure, with data assumed to be missing at random (MAR). With missing items, the model given by Equation 1 is modified so that the product over $m = 1, \ldots, M$ is replaced by a product over the items observed for that individual. A test of the null hypothesis that data are missing completely at random appears in the output.

OVERVIEW OF PROC LCA FOR THE SAS SYSTEM

Data Preparation and Exploratory Data Analysis

Coding items measuring the latent class variable—Response categories of items measuring the latent class variable must be coded with sequential integer values from 1 to R, where *R* is the number of response categories for that particular item. The procedure recognizes SAS system missing values (.) as the code for missing data. After the coding has been done, a helpful preliminary step in any LCA is exploring overall relations among pairs of items by conducting cross-tab analyses. For example, this can help reveal highly related items that may be partially redundant indicators of a latent variable, functioning as parallel items.

Coding the grouping variable—Only one grouping variable can be included, although two or more grouping variables can be crossed to create a single grouping variable (e.g., gender and minority status can be crossed to create a four-level grouping variable: female minorities, male minorities, female not minorities, male not minorities). The grouping variable should be coded with consecutive integers starting with 1 (in the preceding example, the groups would be coded 1, 2, 3, 4).

Coding the covariates—All covariates are treated as numeric in the statistical model. Categorical covariates should be coded as a dummy variable (or a set of $r - 1$ dummy variables if there are $r > 2$ response categories). Continuous covariates can be transformed to *z* scores to aid interpretation by producing standardized logistic regression coefficients.

Basic Model Specification: Required Elements

Table 1 summarizes the statements available in PROC LCA. The SAS data file to be analyzed must be specified using the DATA option. The data file can contain more variables than will be used in the analysis and must contain at least two categorical variables to be used as indicators for the latent class model. There are two ways to organize data for use in PROC LCA. Most users will organize the data file in the most commonly used way, with one record per individual. This data structure is required if covariates are to be included in the model. However, PROC LCA also can handle data that are aggregated into response patterns (i.e., one record for each unique set of responses to all items). Aggregated data should have one record per response pattern with a count variable indicating the number of participants with that particular pattern. An item indicating a grouping variable can be included in the response pattern, but covariates cannot be included. The count variable must be specified in the FREQ statement when data are aggregated into response patterns (this statement should not be used if data are not aggregated). One advantage to aggregating data is that the estimation time can be reduced. An example using the FREQ statement can be found in the *PROC LCA User's Guide* (Lanza, Lemmon, Schafer, & Collins, 2007).

We recommend that all new analyses begin with fitting a baseline model with no grouping variable or covariates. Appendix A shows an example of SAS syntax to call PROC LCA for a baseline latent class model. Note that title statements can be used within the procedure call. The NCLASS statement is used to specify the number of latent classes that are to be estimated. Next, two or more categorical variables to be used as indicators of the latent classes must be listed in the ITEMS statement.

The CATEGORIES statement is used to list the number of response categories in each of the variables listed in the ITEMS statement. The number of arguments here must equal the number of variables listed in the ITEMS statement, and the order of the numbers must correspond to the order of the listed items. The range of valid values is from 2 to 99.

The baseline model specified in Appendix A uses the SEED statement to specify a seed for generating random starting values. Starting values are discussed later, but it is worth noting that estimation requires either a random seed or a SAS data set containing user-specified starting values.

Model Assessment

A good starting point for identifying an optimal baseline model is to fit a sequence of models with two classes, three classes, and so on. A variety of tools can be used together for model selection, including the likelihood-ratio G^2 statistic, Akaike's Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). In addition, model interpretability should be considered. For example, each class should be distinguishable from the others on the basis of the item-response probabilities, no class should be trivial in size (i.e., with a near-zero probability of membership), and it should be possible to assign a meaningful label to each class.

Although models with different numbers of latent classes are technically nested, the distribution of the likelihood ratio statistic comparing two models should not be compared to a chi-square; the difference G^2 statistic can be used only in a rough way to compare model fit. The AIC and BIC are penalized log-likelihood model information criteria that can be used to compare competing models (e.g., models with different numbers of latent classes) fit to the same data. A smaller AIC and BIC for a particular model suggests that the trade-off between fit and parsimony is preferable.

Expanded Model Specification: Multiple Groups LCA and LCA With Covariates

Multiple-group LCA—A grouping variable can be used in LCA in much the same way as in multiple-group structural equation modeling. All parameters can be estimated conditional on group membership, allowing class membership probabilities and item-response probabilities to differ across groups. It is often advisable to establish whether measurement invariance across groups holds before making conclusions about group differences in class membership probabilities. Appendix B shows an example of SAS code for multiple-group LCA.

Multiple-group LCA can be conducted by specifying the grouping variable in the GROUPS statement. When the GROUPS statement is used, the user might wish to provide labels (up to eight characters each) for the groups using the GROUPNAMES statement. The order of the labels must correspond to the order of the integers denoting the groups. Cases with missing data for the grouping variable are automatically deleted. (The number of cases used in the analysis is noted in the output file.)

Measurement invariance across groups—Often when a grouping variable is included it is important to test for measurement invariance across groups. To do this, a model with free estimation of the ρ parameters can be compared to the same model that includes restrictions equating the ρ parameters across groups. (This set of restrictions can be imposed by specifying the keyword *groups* in the MEASUREMENT statement.) Because these two models are nested and the distribution of the likelihood-ratio difference test is asymptotically chi-square, model fit can be compared by examining the difference between the G^2s from each model in the usual way, by comparing the G^2 difference to a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom. A significant p value suggests that the null hypothesis of measurement invariance should be rejected. This implies that the meaning of the latent classes may differ across groups to some extent, and therefore caution should be used in interpreting group differences in the latent class membership probabilities. Examination of the ρ parameters in the unconstrained model can shed light on the nature and extent of group differences in the interpretation of the latent variable. If the group differences in measurement are severe, it may be prudent to conduct the modeling separately for each group, providing group-specific interpretations of the latent classes. Often, however, just one or two specific item-response probabilities may function differently across groups, suggesting possible modifications that can be made to the overall measurement model (e.g., freely estimating one or more ρ parameters in each group). When the null hypothesis is not rejected, measurement invariance can be imposed in the LCA model, implying that there is equivalent meaning of the latent classes across groups.

LCA with covariates—Covariates can be incorporated in the latent class model by specifying the variable names in the COVARIATES statement. The probability of class membership depends on the values or levels of the covariates through multinomial logistic regression, where the dependent variable is latent (latent class membership). When a grouping variable is included in LCA with covariates, the multinomial logistic regression parameters are estimated for each group. The item-response probabilities (ρ parameters) do not depend on the values or levels of the covariates.[2]

The REFERENCE statement specifies the number of the latent class (an integer) to serve as the reference class for the multinomial logistic regression. The default reference class is Class 1.

Cases with missing data for a covariate are automatically deleted. (The number of cases used in the analysis is noted in the output file.) See Appendix C for an example of SAS code for a model involving covariates.

By default, the log-likelihood test for the overall effect of each covariate is reported in the output. For each covariate, a significant result (e.g., $p < .05$) provides evidence that the covariate is a significant predictor of class membership. If significance tests for covariates are not of interest, the NOBETATEST option can be used to suppress them, which saves on the time needed for model estimation.

Estimation Options

PROC LCA employs the EM (expectation-maximization) algorithm to produce maximum likelihood estimates of all model parameters. (The ESTIMATION statement can be used to specify the estimation method to be employed. Currently, the only method available, and the default, is EM.) Based on a data set, a particular model specification, and starting values for the parameters, the algorithm iterates between the Expectation (E) step and the Maximization (M) step until either the convergence criterion is achieved or the maximum number of iterations is reached.

Convergence index and convergence criterion—The convergence index used in PROC LCA is the maximum absolute deviation (MAD) between parameter sets in successive iterations. The EM algorithm stops iterating when MAD falls below a specified value, called the convergence criterion. The default convergence criterion in PROC LCA is MAD = .000001; a different value can be specified using the CRITERION statement. A larger value for the convergence criterion results in convergence in fewer iterations, but noticeable additional improvement in parameter estimates is possible. A smaller value for the convergence criterion requires more iterations to converge, but once convergence is reached, little improvement in estimation is possible.

[2] The fitting paradigm used for this model is based on an assumption of simple random sampling. Although data that do not meet this assumption can, and often are, used for latent class modeling, it is possible for the interpretation of one or more latent classes to change after the addition of covariates. If the ρ parameters change substantially with the addition of a covariate, there is evidence that the data are not representative of the population, and evidence of lack of model fit for the overall population. If this occurs, it might be worthwhile splitting the sample based on levels of the covariate and examining the latent class model separately for each subsample. Fitting latent class models in samples that are not simple random samples is a topic of current research (see, e.g., Asparouhov, 2005).

Maximum number of iterations—The MAXITER statement can be used to specify the maximum number of iterations to be performed by the EM estimation procedure. If convergence is reached prior to the value specified in the MAXITER statement, the procedure terminates normally. Most latent class models converge well before the default value of 5,000 iterations. Very slow convergence can be a sign of underidentification.

Starting values—Random starting values for the ρ parameters can be generated in PROC LCA by specifying any positive integer value in the SEED statement (default starting values of 1/NCLASS for γ parameters and 0 for β parameters are used). Use of an identical seed reproduces an analysis exactly. In most cases, random starting values suffice. However, in some cases the user might wish to provide starting values in a SAS data file by specifying the file-name in the START option. (Note that either the SEED statement or the START option must be included.) The structure of this file must be identical to that of the file created with the OUTPARAM option (see the section on optional output later). This data file must contain starting values for the γ, β, and ρ parameters, although focus should be given to values for the ρ parameters. We recommend using the default starting values for the γ parameters (1/NCLASS) and the β parameters (values of 0), as they will have little effect on estimation. If starting values are provided for ρ parameters using the START option and a SEED is specified, user-provided starting values are ignored. See Appendix A for an example using the SEED statement, and Appendix 2 in the *PROC LCA User's Guide* (Lanza et al., 2007) to see how to provide starting values in a SAS data set.

Parameter restrictions—The RESTRICT option allows the user to specify parameter restrictions for the ρ and γ parameters. A SAS data file containing parameter restrictions is specified here. The structure of this file must be identical to that of the file created with the OUTPARAM option (see the section on optional output later). Restrictions for β parameters are not available; all restriction values for these parameters should be set to the value one, indicating free estimation. Parameter restrictions for the ρ parameters can be useful to help achieve model identification or to test specific hypotheses about the measurement of the latent class variable. Appendix 2 in the *PROC LCA User's Guide* (Lanza et al., 2007) demonstrates user-provided parameter restrictions.

Restrictions provided for each parameter in a SAS data file must be integers of value zero or higher. The following restrictions for ρ and γ parameters are possible:

- Values of zero indicate that a particular parameter is to be fixed to its corresponding starting value. If the user wishes to fix parameter estimates to a specific value, the START option must be used in conjunction with the RESTRICT option.
- Values of one indicate that a particular parameter is to be freely estimated (this is the default for all parameters when the RESTRICT option is not used).
- Integer values of two or greater are used to specify an equivalence set. Estimates for all ρ or γ parameters with the same value are constrained equal to one another; in other words, a single parameter is estimated for each set.

If the MEASUREMENT statement is used in conjunction with the RESTRICT option, user-provided restrictions corresponding to ρ parameters for Group 1 are applied to all subsequent groups. Additional information on the use of parameter restrictions can be found in the *WinLTA User's Guide* (Collins et al., 2002).

Model identification—Model identification is an issue that should be explored in LCA, as in all latent variable models. The optimal solution to a model can be difficult to identify if the amount of information provided is small relative to the number of parameters being estimated. Information refers to several things, including the number of participants and the number of items. In general, complex latent class models (e.g., models with a large number of latent classes, groups, or covariates) require more information than simple ones. The best way to detect identification problems or local optima (i.e., solutions other than the optimal one) is to fit the same model using multiple sets of starting values. This can be done by calling the procedure repeatedly with different seeds specified. (Advanced SAS users easily can set up a macro to call the procedure repeatedly.) Even well-identified models can land on a different solution occasionally; if the solution with the smallest log-likelihood is arrived at using the majority of the seeds, one can have confidence that it is the optimal solution. If an identification problem is observed, providing reasonable parameter restrictions in the ρ parameters often can solve the problem.

Optional Output: Parameter Estimates, Posterior Probabilities, and Verbose Output

Parameter estimates—Parameter estimates are displayed in the PROC LCA output, which is sufficient for many users. In the output, estimates are displayed with four significant digits. Users who wish to see the parameter estimates displayed with greater precision might wish to save the parameter estimates to a SAS file using the OUTEST or OUTPARAM option. These options produce files with identical content but different structures.

The OUTEST option produces a SAS data file with the specified name in which all final parameter estimates are saved in a single record, with a unique variable name for each parameter estimate. The OUTPARAM option produces a SAS data file with the specified name that contains final parameter estimates 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 must be one of the following character strings: 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.

Because the START and RESTRICT options require SAS data files of the same format as the file generated by the OUTPARAM option, an easy way to specify starting values or parameter restrictions is to start by running a preliminary model invoking the OUTPARAM option. Running just one iteration would suffice for this step, which can be accomplished using the MAXITER statement. The SAS data file can then be renamed and modified by replacing the preliminary parameter estimates with either starting values or parameter restrictions. The modified file can be input in a subsequent run using the START or RESTRICT option. This practice ensures that the correct structure for the starting values and restrictions SAS data files is used.

Posterior probabilities—Bayes's theorem can be used to compute each individual's probability of membership in each latent class. The theorem, which is based on individuals' responses to the latent class indicators, values on covariates, and group membership, as well as the estimated model parameters, is:

$$P(C=c|Y=y, x_i, g)=\frac{P(C=c|x_i, g)P(Y=y|C=c, x_i, g)}{P(Y=y|x_i, g)}. \tag{4}$$

These values, referred to as posterior probabilities of latent class membership, can be saved to a SAS data file by specifying a file name in the OUTPOST option. 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). For each latent class, one variable containing the posterior probability of membership in that latent class is appended (the new variables are labeled POSTLC1, POSTLC2, etc.). When data are structured with one record per individual, the ID statement can be used to specify one or more variables in the analysis data set that are to be included in the OUTPOST SAS data file. For example, by listing the case identifier in the ID statement, this identifier is carried through to the OUTPOST data file, allowing the file containing posterior probabilities to be merged with other data files. See Appendix C 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. An application of posterior probabilities is described later.

When data are organized with one record per individual, 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, 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 count variable (listed in the FREQ statement), and the posterior probabilities, which are the same for all individuals with a specific response pattern.

Verbose output—The VERBOSE_OUTPUT option produces more detailed output that includes the following: restrictions used for all parameters, starting values used for all parameters, and the iteration history (this shows the MAD and value of the log-likelihood at each iteration). Listing of the output in SAS can be suppressed using the NOPRINT option. This can be useful when the parameter estimates are being saved to a file (using the OUTEST or OUTPARAM options) but no output is needed.

Postprocessing in SAS

Individuals' posterior probabilities of latent class membership can be useful tools for describing the latent classes and assessing the accuracy with which individuals can be assigned to latent classes. Using the posterior probabilities saved in the optional OUTPOST SAS data file, individuals can be assigned to the class in which they have the highest posterior probability of membership (this is sometimes referred to as the maximum-probability assignment rule). If class membership is treated as known, describing classes can be done in a straightforward manner, for example, by using analysis of variance to test for mean differences across class in some characteristic. However, unless the probability of membership in a particular latent class is one for an individual, there is uncertainty associated with latent class membership. Because this uncertainty generally is not modeled when subsequent analyses are done using latent class assignment as a variable, it is important to interpret the results of such analyses with caution.

Posterior probabilities can be useful indicators of the assignment accuracy of a model. If individuals are assigned using maximum-probability assignment, the average posterior probability of membership can be calculated for membership in each class, given class assignment. An average close to one for the assigned class suggests that one can have high certainty about true class membership for those individuals. The average posterior probabilities can then be used to calculate the odds of correct classification for each class (Nagin, 2005) and the overall entropy (Celeux & Soromenho, 1996). These diagnostic tools are useful for judging the confidence one can have when assigning individuals to classes. When the classes are more distinct, posterior probabilities tend to be closer to one for a single class, and closer to zero for the remaining classes (see Nagin, 2005, for simulation results). Although posterior probabilities close to zero and one are desirable, particularly if class assignment is conducted, they are not necessarily an indicator of model fit.

Note that if there are participants with all missing indicator data, their posterior probabilities will be equal to the overall class membership probabilities. When assigning individuals to classes based on their maximum posterior probability, it might make sense to include only those individuals who have responded to at least one of the indicators.

A STEP-BY-STEP EXAMPLE OF LCA: ALCOHOL BEHAVIOR LATENT CLASSES

The example used here to illustrate PROC LCA explores latent classes of alcohol use behavior among high school seniors in the United States. Data are from the 2004 cohort of the public release of the Monitoring the Future study (Johnston, Bachman, O'Malley, & Schulenberg, 2004). The sample consists of 2,490 high school seniors (48% boys, 52% girls) who answered at least one question on alcohol use. The goals of the study are to explore alcohol use behavior in this population, examine gender differences in the measurement of alcohol use and in alcohol use behavior, and explore whether grades and skipping school are predictive of alcohol behavior class membership.

Seven binary indicators of drinking behavior were used in the latent class model. The indicators measured lifetime alcohol use (more than just a few sips), past-year alcohol use, past-month alcohol use, lifetime drunkenness, past-year drunkenness, past-month drunkenness, and five or more drinks in a row during the last 2 weeks. Each indicator was coded 1 (*behavior not reported*) or 2 (*behavior reported*). Additional variables used in this example were an indicator of sex, a binary variable indicating whether participants have skipped an entire day (or more) of school in the previous month, and a continuous measure of grades. Table 2 shows the distribution of all variables used in the example.

The following three research questions were addressed:

1. Are there underlying types of drinking behavior? In other words, is there a latent class structure that adequately represents the heterogeneity in drinking behavior among high school seniors? If so, what are the types and their corresponding prevalence?
2. Is the measurement of drinking behavior latent classes invariant across sex? In other words, does the same class structure for drinking behavior hold for males and females? This question does not imply that the prevalence of drinking classes would be constant across sex, just the measurement of drinking classes.
3. Are grades or skipping school days predictive of membership in drinking behavior classes?

Research Question 1: Baseline Model Selection

As Table 3 shows, the drop in G^2 relative to the drop in degrees of freedom is substantial with each additional class up to the five-class model; the addition of classes beyond five provides essentially no improvement in fit, so based on the G^2 statistic the five-class model appears best. The AIC and BIC values shown in Table 3 agree with the G^2 statistic, also indicating that the five-class model is the best among these models.

An inspection of the parameter estimates from the five-class model suggests that the classes are distinguishable and nontrivial, and meaningful labels can be assigned to each. Each column of Table 4 shows, for each class, the assigned label and probability of membership, as well as the item-response probabilities for endorsing each item. For example, 17.9% of high school seniors are expected to belong to the nondrinkers class; these individuals are not expected to endorse any of the seven drinking questionnaire items. Similarly, 33.5% are expected to belong to the heavy drinkers class, with very high probabilities of endorsing all seven items. The remaining classes are the experimenters (21.9%), drinkers (9.3%), and occasional bingers (17.4%).

Before selecting a final baseline model, identification should be examined. It helps to look at the log-likelihood value across iterations to see if convergence was achieved smoothly (the iteration history can be printed by specifying the VERBOSE_OUTPUT option). In addition, the estimation should be repeated using different seeds to try different sets of starting values. Models that are identified will have one dominant solution that is arrived at most frequently among various sets of starting values. Solutions should be considered to be identical if the log-likelihood and parameter estimates are replicated, regardless of the somewhat arbitrary order of the latent classes. In this example, the five-class model appeared to be identified, and was selected as the baseline model of alcohol use behavior among high school seniors.

Research Question 2: Multiple-Groups LCA

Once a baseline latent class model is selected, the user might wish to incorporate grouping variables or covariates. Sex was added to the five-class model as a grouping variable. To test whether measurement is invariant across sex, this model was run with all parameters freely estimated and again with item-response probabilities constrained equal across groups (see Appendix B for the model specification). The G^2 statistic was 27.8 (df = 177) for the freely estimated model and 61.4 (df = 212) for the constrained model, resulting in a likelihood-ratio difference test statistic of 33.6 (df = 35). This difference is not statistically significant, providing strong evidence that measurement invariance across sex holds. The item-response probabilities were held equal across sex for all remaining analyses that include sex as a grouping variable.

Because measurement invariance held, sex differences in class membership probabilities (γ parameters) could be interpreted with confidence that the classes have exactly the same meaning for males and females. Males and females were equally likely to belong to the nondrinkers class (18.0% of males, 18.1% of females), experimenters class (22.0% of males, 22.7% of females), and drinkers class (9.5% of males, 9.5% of females), whereas females were more likely to belong to the occasional bingers class (12.9% of males, 21.3% of females) and males were more likely to belong to the heavy drinkers class (37.7% of males, 28.5% of females).

Research Question 3: LCA With Covariates

In the example, grades and an indicator of whether the adolescent had skipped school in the past month were added as covariates. To simplify interpretation, each covariate was added separately in the latent class model, although both covariates (and their interaction, if desired) could be included in a single model. The nondrinkers class (Class 5) was specified as the reference class for the multinomial logistic regression. (Appendix C shows the model specification for both models.)

Both grades (p < .0001) and skipped school (p < .0001) were strong predictors of latent class membership. Table 5 shows the β parameters for the effect of each covariate, as well as odds ratios (exponentiated β parameters). For grades, the inverse of the odds ratio is also shown to facilitate interpretation.

For skipped school (a dummy variable), odds ratios are interpreted as the increase in odds of membership in a particular latent class relative to the reference class given that an adolescent has skipped school in the past month. For example, adolescents who skipped school were 50% more likely to be in the experimenters class than the nondrinkers class, twice as likely to be in the drinkers class than the nondrinkers class, and so on. The most striking finding is that adolescents who skipped school were five times more likely to belong in the heavy drinkers class than the nondrinkers class.

For grades (a standardized variable), odds ratios are interpreted as the increase in odds of membership in a particular latent class relative to the reference class corresponding to a one-unit increase in the covariate. When odds ratios are less than one, the inverse often can be interpreted more easily. For example, for every 1 *SD* lower in grades, adolescents were 20% more likely to be in the experimenters class than the nondrinkers class, 40% more likely to be in the drinkers class than the nondrinkers class, and so on. Again, the strongest effect of grades appears for heavy drinkers, such that for every 1 *SD* lower in grades, adolescents were 60% more likely to belong in this class than the nondrinkers class. The reference class can be respecified to obtain odds ratios for different pairwise comparisons.

Plots of class membership probabilities across levels of the covariates—One helpful tool for interpreting the effect of covariates on the latent class variable is plots showing the prevalence of each latent class across levels of each covariate. To do this, a Microsoft Excel spreadsheet was created with a range of values on the covariates. For skipped school, class membership probabilities are plotted in Figure 1 for those who did not skip school in the previous month, and for those who did skip school. Probabilities of membership in the nondrinkers, experimenters, and drinkers classes were lower for individuals who skipped school, and the probability of membership in the heavy drinkers class was twice as high for those individuals. For grades, class membership probabilities are plotted in Figure 2 for values ranging from −2.0 to 2.0 on the standardized grades variable (increments of 0.1 *SD* were used). Individuals with grades at least 1.5 *SD* above the mean were most likely to belong to the nondrinkers class. The experimenters, drinkers, and bingers classes were not strongly related to grades. Membership in the heavy drinking class was strongly predicted by grades, with nearly half of the lowest graded individuals expected to belong to this class.

CONCLUSIONS

Scientists are increasingly using latent class models to identify underlying subgroups of individuals who share important characteristics or behaviors. PROC LCA provides a simple, convenient approach to estimating these models in the SAS environment. The procedure is available for download free of charge at http://methodology.psu.edu. Multiple-group LCA and LCA with covariates provide two important and useful extensions to the traditional latent class model. PROC LCA provides the basis for future work on additional features and modeling extensions, including a SAS procedure for latent transition analysis, where transitions over time in latent class membership are modeled using longitudinal data.

Acknowledgments

This research was supported by National Institute on Drug Abuse Grants P50 DA 10075 and K05 DA 018206. We would like to thank Kari C. Kugler for helpful comments on a draft of this article and Bobby L. Jones for providing SAS programming expertise.

References

Agresti, A. Categorical data analysis. Vol. 2. New York: Wiley; 2002.

Aitkin M, Anderson D, Hinde J. Statistical modeling of data on teaching styles. Journal of the Royal Statistical Society A 1981;144:419–461.

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19:716–723.

Asparouhov T. Sampling weights in latent variable modeling. Structural Equation Modeling 2005;12:411–434.

Auerbach KJ, Collins LM. A multidimensional developmental model of alcohol use during emerging adulthood. Journal of Studies on Alcohol 2006;67:917–925. [PubMed: 17061010]

Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. Journal of the American Statistical Association 1997;92:1375–1386.

Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification 1996;13:195–212.

Chung H, Flaherty BP, Schafer JL. Latent-class logistic regression: Application to marijuana use and attitudes among high-school seniors. Journal of the Royal Statistical Society: Series A 2006;169:723–743.

Chung H, Park Y, Lanza ST. Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. Statistics in Medicine 2005;24:2895–2910. [PubMed: 16134129]

Clogg CC, Goodman LA. Latent structure analysis of a set of multidimensional contingency tables. Journal of the American Statistical Association 1984;79:762–771.

Collins, LM.; Lanza, ST.; Schafer, JL.; Flaherty, BP. WinLTA user's guide version 3.0. University Park: The Pennsylvania State University, The Methodology Center; 2002.

Dayton CM, Macready GB. Concomitant-variable latent-class models. Journal of the American Statistical Association 1988;83:173–178.

Dewilde C. The multidimensional measurement of poverty in Belgium and Britain: A categorical approach. Social Indicators Research 2004;68:331–369.

Gibson WA. Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. Psychometrika 1959;24:229–252.

Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 1974;61:215–231.

Jackson KM, Sher JJ, Gotham HJ, Wood PK. Transitioning into and out of large-effect drinking in young adulthood. Journal of Abnormal Psychology 2001;100:378–391. [PubMed: 11502081]

Johnston, LD.; Bachman, JG.; O'Malley, PM.; Schulenberg, JE. Monitoring the future: A continuing study of American youth (12th-grade survey). Ann Arbor, MI: Inter-University Consortium for Political and Social Research; 2004. [Computer file]

Lanza ST, Collins LM. Pubertal timing and the stages of substance use in females during early adolescence. Prevention Science 2002;3:69–82. [PubMed: 12002560]

Lanza ST, Collins LM. A mixture model of discontinuous development in heavy drinking from ages 18 to 30: The role of college enrollment. Journal of Studies on Alcohol 2006;67:552–561. [PubMed: 16736075]

Lanza, ST.; Flaherty, BP.; Collins, LM. Latent class and latent transition models. In: Schinka, JA.; Velicer, WF., editors. Handbook of psychology: Vol. 2. Research methods in psychology. Hoboken, NJ: Wiley; 2003. p. 663-685.

Lanza, ST.; Lemmon, D.; Schafer, JL.; Collins, LM. PROC LCA user's guide version 1.1.3 beta. University Park: The Pennsylvania State University, The Methodology Center; 2006.

Lazarsfeld, PF.; Henry, NW. Latent structure analysis. Boston: Houghton-Mifflin; 1968.

Muthén, LK.; Muthén, BO. *M*plus *user's guide*. Vol. 4. Los Angeles: Muthén & Muthén; 1998–2006.

Nagin, DS. Group-based modeling of development. Cambridge, MA: Harvard University Press; 2005.

Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978;6:461–464.

Stern HS, Arcus D, Kagan J, Rubin DB, Snidman N. Using mixture models in temperament research. International Journal of Behavioral Development 1995;18:407–423.

van de Pol, F.; Langeheine, R.; de Jong, W. PANMARK 3 user's manual. Voorburg, The Netherlands: Netherlands Central Bureau of Statistics; 1998.

Velicer WF, Martin RA, Collins LM. Latent transition analysis for longitudinal data. Addiction 1996;91:S197–S209. [PubMed: 8997793]

Vermunt, JK.; Magidson, J. Latent GOLD user's guide. Boston, MA: Statistical Innovations, Inc; 2000.

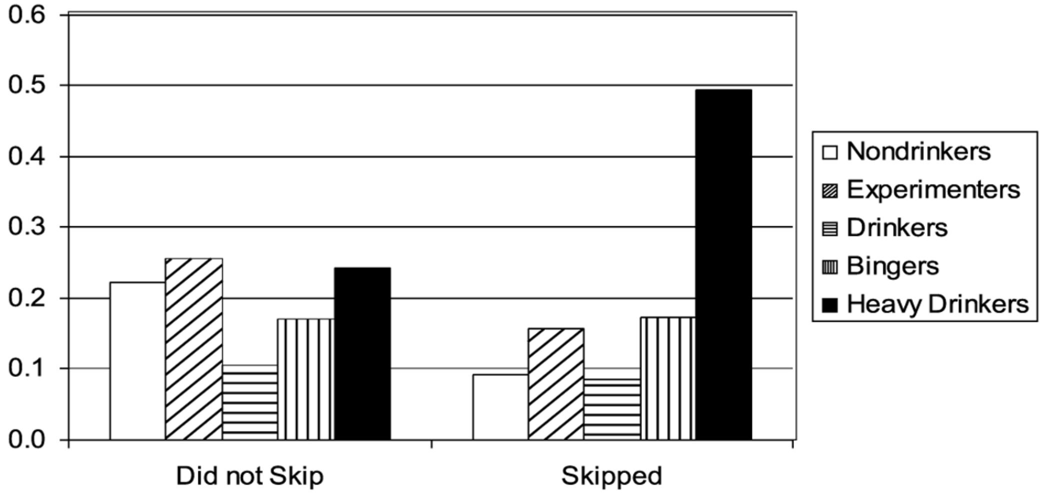

FIGURE 1.
Class membership probabilities as a function of whether adolescents skipped school in the past 30 days.

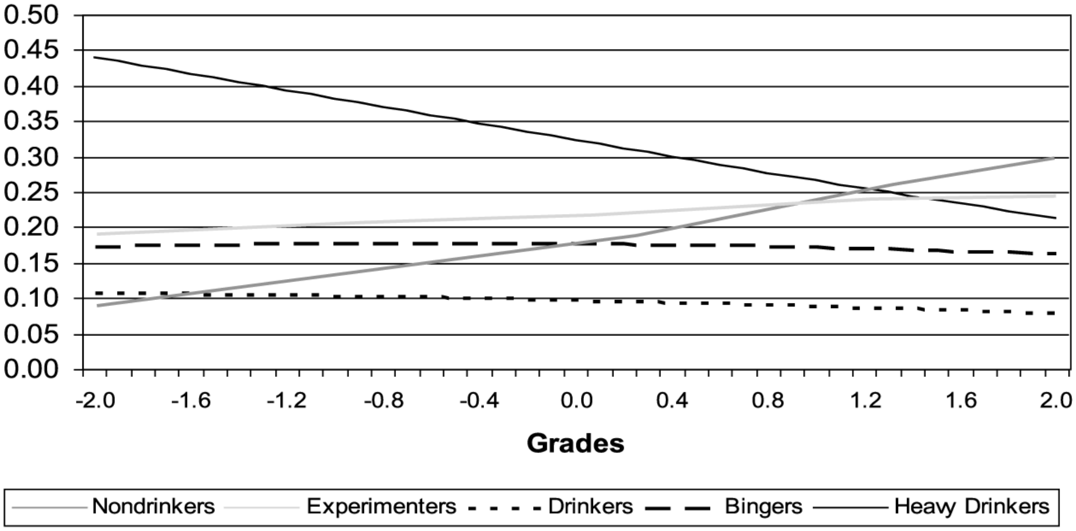

FIGURE 2.
Class membership probabilities as a function of grades in school (standardized).

TABLE 1

Summary of PROC LCA Statements and Options

Syntax	Required	Description
PROC LCA	✓	Invokes the procedure
Options		
DATA	✓	Specifies SAS data file to be analyzed
VERBOSE_OUTPUT		Shows starting values, parameter restrictions, and maximum absolute deviation and log-likelihood at each iteration
OUTEST		Saves parameter estimates to SAS data file in one record
OUTPARAM		Saves parameter estimates to SAS data file
OUTPOST		Saves posterior probabilities to SAS data file
NOPRINT		Suppresses printing of output
START		Allows user to provide starting values
RESTRICT		Allows user to specify parameter restrictions for item-response probabilities
NOBETATEST		Suppresses tests of significance for covariates
Statements		
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 SAS file
GROUPS		Declares categorical grouping variable
GROUPNAMES		Specifies a label for each group
COVARIATES		Declares variables to include as covariates
REFERENCE		Specifies latent class to use as reference group in prediction from covariates
FREQ		Declares frequency count variable, to use when data are aggregated
ESTIMATION		Specifies estimation procedure
SEED	✓[a]	Specifies seed for random number generator
MEASUREMENT		Invokes measurement invariance across groups
MAXITER		Specifies maximum number of iterations
CRITERION		Specifies convergence criterion for maximum absolute deviation

[a] SEED statement is required only if the START option is not included in the PROC LCA statement.

TABLE 2

Descriptive Statistics

Variable in Model	Code	Label	Frequency (Valid %)
Indicators of latent class			
Lifetime alcohol use	1	No	442 (18.1)
	2	Yes	2,001 (81.9)
	.	Missing	47
Past-year alcohol use	1	No	652 (26.8)
	2	Yes	1,784 (73.2)
	.	Missing	54
Past-month alcohol use	1	No	1,235 (50.5)
	2	Yes	1,210 (49.5)
	.	Missing	45
Lifetime drunkenness	1	No	988 (42.7)
	2	Yes	1,325 (57.3)
	.	Missing	177
Past-year drunkenness	1	No	1,182 (51.2)
	2	Yes	1,126 (48.8)
	.	Missing	182
Past-month drunkenness	1	No	1,672 (71.4)
	2	Yes	668 (28.6)
	.	Missing	150
5+ drinks in past 2 weeks	1	No	1,773 (74.2)
	2	Yes	617 (25.8)
	.	Missing	100
Grouping variable			
Sex	1	Male	1,098 (47.7)
	2	Female	1,204 (52.3)
	.	Missing	188
Categorical covariate			
Skipped school	0	No	1,478 (67.1)
	1	Yes	725 (32.9)
	.	Missing	287

	% Missing	*M (SD)*
Continuous covariate		
Grades (standardized)	8.8	0.0 (1.0)

TABLE 3

Comparison of Baseline Models

No. of Classes	Likelihood Ratio G^2	Degrees of Freedom	AIC	BIC
2	2561.5	112	2591.5	2678.8
3	910.7	104	956.7	1090.5
4	209.1	96	271.1	451.5
5	**3.5**	**88**	**81.5**	**308.4**
6	3.5	80	97.5	371.0
7	2.8	72	112.8	432.9

Note. Boldface type indicates the selected model. AIC = Akaike’s Information Criterion; BIC = Bayesian Information Criterion.

TABLE 4

Item-Response Probabilities for Five-Class Model: Probability of Endorsing Item Given Latent Class

	Latent Class				
Item	**Nondrinkers (17.9%)**	**Experimenters (21.9%)**	**Drinkers (9.3%)**	**Occasional Bingers (17.4%)**	**Heavy Drinkers (33.5%)**
Lifetime alcohol use	0.000	1.000	1.000	1.000	1.000
Past year use	0.000	0.607	1.000	1.000	1.000
Past month use	0.000	0.000	1.000	0.385	1.000
Lifetime drunkenness	0.000	0.241	0.293	1.000	1.000
Past year drunkenness	0.000	0.000	0.000	1.000	1.000
Past month drunkenness	0.000	0.000	0.000	0.000	0.924
5+ drinks in past 2 weeks	0.000	0.000	0.161	0.000	0.732

TABLE 5

Parameter Estimates and Odds Ratios for Covariates

	Skipped School		Grades		
Class	**β**	**Odds Ratio**	**β**	**Odds Ratio**	**Inverse Odds Ratio**
Nondrinkers	—	1.0	—	1.0	1.0
Experimenters	0.4	1.5	−0.2	0.8	1.2
Drinkers	0.7	2.0	−0.4	0.7	1.4
Bingers	0.9	2.5	−0.3	0.7	1.4
Heavy drinkers	1.6	5.0	−0.5	0.6	1.6

Note. Dashes indicate the reference class.

APPENDIX A

SAS SYNTAX FOR BASELINE MODEL

Five-Class Model

```
PROC LCA DATA=drugs;
  TITLE2 ‘5-class model, 7 items’;
  NCLASS 5;
  ITEMS alc_life alc_yr alc_mo alc_5up drunk_life
drunk_yr drunk_mo;
  CATEGORIES 2 2 2 2 2 2 2;
  SEED 861551;
RUN;
```

APPENDIX B

SAS SYNTAX FOR MODELS WITH A GROUPING VARIABLE

Five-Class Model With Sex as a Grouping Variable and All Parameters Estimated Freely

```
PROC LCA DATA=drugs;
  TITLE2 ‘5-class model, 7 items, by sex (no
measurement invariance)’;
  ID caseid;
  NCLASS 5;
  ITEMS alc_life alc_yr alc_mo alc_5up drunk_life
drunk_yr drunk_mo;
  CATEGORIES 2 2 2 2 2 2 2;
  GROUPS sex;
  GROUPNAMES male female;
  SEED 861551;
RUN;
```

Five-Class Model With Sex as a Grouping Variable, With Measurement Invariance Imposed Across Groups

```
PROC LCA DATA=drugs;
  TITLE2 ‘5-class model, 7 items, by sex (measurement
invariance)’;
  ID caseid;
  NCLASS 5;
  ITEMS alc_life alc_yr alc_mo alc_5up drunk_life
drunk_yr drunk_mo;
  CATEGORIES 2 2 2 2 2 2 2;
  GROUPS sex;
  GROUPNAMES male female;
  MEASUREMENT groups;
  SEED 861551;
RUN;
```

APPENDIX C

SAS SYNTAX FOR MODELS WITH COVARIATES

Five-Class Model With Grades in School as Covariate

```
PROC LCA data=alcohol;
  TITLE2 ‘5 drinking classes, grades as covariate’;
  ID caseid;
  NCLASS 5;
  ITEMS alc_life alc_yr alc_mo drunk_life drunk_yr
drunk_mo alc_5up;
  CATEGORIES 2 2 2 2 2 2 2;
  COVARIATES grades;
  REFERENCE 5;
  SEED 861551;
RUN;
```

Five-Class Model With Skipped School as Covariate

```
PROC LCA data=alcohol;
  TITLE2 ‘5 drinking classes, skipping school as
covariate’;
  ID caseid;
  NCLASS 5;
  ITEMS alc_life alc_yr alc_mo drunk_life drunk_yr
drunk_mo alc_5up;
  CATEGORIES 2 2 2 2 2 2 2;
  COVARIATES skip;
  REFERENCE 5;
  SEED 861551;
RUN;
```