An Introduction to Elminating Bias in Classify-Analyze Approaches for Latent Class Analysis

Bethany C. Bray[1,2]
Stephanie T. Lanza[2,3]
Xianming Tan[2,4]

[1]Department of Psychology, Virginia Polytechnic Institute and State University
[2]The Methodology Center, The Pennsylvania State University
[3]College of Health and Human Development, The Pennsylvania State University
[4]Biostatistics Core Facility, McGill University Health Centre

Technical Report Series
#12-118

College of Health and Human Development
The Pennsylvania State University

Address correspondence to Bethany C. Bray
The Methodology Center, The Pennsylvania State University
204 E. Calder Way, Suite 400, State College, PA 16801
Phone: (814) 863-9795
bcb178@psu.edu

KEY WORDS: Latent class analysis, Posterior probabilities, Pseudo-class draws, Classify-analyze

The project described was supported by Award Number P50-DA010075 from the National Institute on Drug Abuse. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health**.**

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

An Introduction to Eliminating Bias in Classify-Analyze Approaches for Latent Class Analysis

Bethany C. Bray[1,2], Stephanie T. Lanza[2,3], Xianming Tan[2,4]

[1]Department of Psychology, Virginia Polytechnic Institute and State University

[2]The Methodology Center, The Pennsylvania State University

[3]College of Health and Human Development, The Pennsylvania State University

[4]Biostatistics Core Facility, McGill University Health Centre

The Methodology Center Technical Report No. 12-118

Author Note

Correspondence should be addressed to Bethany C. Bray, The Methodology Center, Penn State, 204 East Calder Way, Suite 400, State College, PA, 16801. E-mail: bcb178@psu.edu; Office: 814-865-1225.

The project described was supported by Award Number P50-DA010075 from the National Institute on Drug Abuse. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (`http://www.cpc.unc.edu/addhealth`). No direct support was received from grant P01-HD31921 for this analysis.

Abstract

Recent advances in latent class analysis (LCA) have resulted in a rapid increase in the application of this method in behavioral research. As scientific questions about the relation between latent class membership and other variables of interest become more complex, however, often they cannot be addressed in the context of the latent class model itself, but must be addressed using a classify-analyze approach. These approaches rely on posterior probabilities to classify individuals prior to the analysis of interest. Typically, the posterior probabilities are generated using a non-inclusive LCA that includes manifest indicators but not other variables of interest that are included in the analysis model; when the analysis model is more general than the classification model, it is expected that the estimated relations between latent class membership and the other variables are attenuated. We propose the use of an inclusive LCA in which all variables included in the analysis model are also included in the classification model. First, a motivating empirical demonstration is presented to illustrate the use of four inclusive and four non-inclusive classify-analyze approaches. Second, a simulation study is presented to assess the performance of these approaches. Performance of inclusive versus non-inclusive approaches is compared, and the impact of different levels of latent class measurement quality, effect size of relation, and sample size are studied. Results show that bias in effect estimation using standard analysis approaches is eliminated using an inclusive classify-analyze approach for LCA with sufficient measurement quality or sample size.

Keywords: latent class analysis, posterior probabilities, pseudo-class draws, classify-analyze

An Introduction to Eliminating Bias in Classify-Analyze Approaches for Latent Class Analysis

As application of latent class analysis (LCA) increases in the behavioral sciences, more complex scientific questions are being posed about the role latent class membership plays in development. For example, scientists are increasingly interested in the effect of latent class membership on later developmental outcomes (e.g., Nylund, Bellmore, Nishina, & Graham, 2007; Petras & Masyn, 2010; Roberts & Ward, 2011; Hardigan & Sangasubana, 2010; Reinke, Herman, Petras, & Ialongo, 2008). In particular, membership in latent classes representing exposure to different combinations of risk factors during adolescence may be related to a negative distal outcome during adulthood like binge drinking (e.g., Lanza & Rhoades, 2011). Further, complex questions arise when theory posits latent class membership acting as a moderator or mediator in a model linking an individual's earlier experiences to later outcomes.

Examining the role played by a latent class variable in a developmental process requires modeling its association with a variety of other variables of interest. Sometimes, the associations between a latent class variable and other variables of interest can be modeled in the context of the latent class model itself. This is desirable because it allows measurement error to be estimated and removed from the estimates of interest. However, there remain many questions that cannot be addressed within the context of the latent class model because the posited relations are more complex than our current understanding of the model can handle. In these cases, a classify-analyze approach (Clogg, 1995) may be required. This involves first classifying individuals into latent classes, and then performing a subsequent analysis using latent class membership as a categorical variable in a larger model of interest. The nature of latent class variables is such than an individual's true class membership cannot be known. Instead, each individual has a probability of membership in each latent class; these probabilities are known as posterior probabilities. All methods for classifying individuals are based on posterior probabilities derived from the

latent class measurement model.

Below we briefly review the latent class model and review current classify-analyze approaches for LCA. We then propose the use of a more inclusive latent class model for deriving the posterior probabilities in order to reduce attenuation of the effects between latent class membership and other variables of interest. Our motivating example involves the identification of risk exposure latent classes on the basis of six characteristics: household poverty, single-parent status, peer cigarette use, peer alcohol use, neighborhood unemployment, and neighborhood poverty (Lanza & Rhoades, 2011). In this example, latent class membership is used to predict later binge drinking. A simulation study is then conducted to evaluate the performance of the more inclusive latent class model.

The Latent Class Model

The latent class model has been described in detail in a variety of resources (e.g., Collins & Lanza, 2010; Lanza, Collins, Lemmon, & Schafer, 2007; Clogg, 1995; Goodman, 1974b) and applied to model a variety of constructs in the psychological and behavioral sciences (e.g., Reboussin, Song, Shrestha, Lohman, & Wolfson, 2006; Biemer & Wiesen, 2002; Loken, 2004). Details of the latent class model relevant to classify-analyze approaches for LCA are briefly discussed below.

The traditional latent class model posits a mutually exclusive and exhaustive underlying set of latent classes (i.e., subgroups) in the population that are inferred from multiple categorical observed variables. Suppose that there are $j = 1, ..., J$ observed variables measuring the latent classes, and that observed variable j has $r_j = 1, ..., R_j$ response categories. Let $\mathbf{y} = (r_1, ..., r_J)$ represent the vector of a particular individual's responses to the J variables. Let C represent the latent variable with $c = 1, ..., K$ latent classes. Finally, $I(y_j = r_j)$ is an indicator function that equals 1 when the response to

variable $j = r_j$, and equals 0 otherwise. Then the latent class model can be expressed as

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{c=1}^{K} \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)}, \tag{1}$$

where γ_c is the probability of membership in latent class c, and $\rho_{j,r_j|c}^{I(y_j=r_j)}$ is the probability of response r_j to item j, conditional on membership in latent class c. The γ parameters represent a vector of latent class membership probabilities that sum to 1. The ρ parameters represent a matrix of item-response probabilities conditional on latent class membership. This traditional model can be extended to include covariates (e.g., Collins & Lanza, 2010; Lanza & Collins, 2008; Chung, Flaherty, & Schafer, 2006). When a covariate X is added to the latent class model to predict latent class membership, the model can be expressed as

$$P(\mathbf{Y} = \mathbf{y}|X = x) = \sum_{c=1}^{K} \gamma_c(x) \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)}, \tag{2}$$

where $\gamma_c(x) = P(C = c|X = x)$ is a standard baseline-category multinomial logistic model (e.g., Agresti, 2002). With a single covariate, $\gamma_c(x)$ can be expressed as

$$\gamma_c(x) = P(C = c|X = x) = \frac{e^{\beta_{0,c}+\beta_{1,c}x}}{1 + \sum_{c'=1}^{K-1} e^{\beta_{0,c'}+\beta_{1,c'}x}} \tag{3}$$

for $c' = 1, ..., K-1$ and reference latent class K. Once the parameters of the latent class model have been estimated, posterior probabilities of membership in each latent class can be obtained for each individual using Bayes' Theorem (e.g., Gelman, Carlin, Stern, & Rubin, 2003; Lanza et al., 2007). Posterior probabilities can be calculated both when covariates are not included in the LCA,

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

and when covariates are included in the LCA,

$$P(C = c|Y = y, x) = \frac{P(C = c|x)P(Y = y|C = c, x)}{P(Y = y|x)}. \tag{5}$$

Cohen's w (Cohen, 1992) can be used as a measure of effect size, indicating the strength of the relation between latent class membership (C) and a categorical distal outcome (Z). The effect size is calculated as follows:

$$w = \sqrt{\sum_{i=1}^{m}\sum_{c=1}^{K}\frac{(\pi_c\pi_{i|c} - \pi_i)^2}{\pi_i}}, \tag{6}$$

where m is the number of categories of the distal outcome, π_c is $P(C = c)$, $\pi_{i|c}$ is $P(Z = i|C = c)$, and π_i is $P(Z = i)$.

Classify-Analyze Approaches for LCA

LCA is a probability-based approach that does not require assignment of individuals to latent classes; this is one of its greatest strengths because it provides a way to account for measurement error in responses to manifest indicators. There are times, however, when scientists must assign individuals to latent classes so that class membership can be used in a subsequent analysis as an exogenous or endogenous variable.

Currently, two popular classify-analyze approaches for LCA are used in practice. Both approaches rely on posterior probabilities to classify individuals. The first approach is a maximum-probability assignment rule (Nagin, 2005, pg. 80), which assigns individuals to the class for which they have the highest posterior probability of membership (Goodman, 2007, 1974a). The subsequent analysis is performed once with latent class membership treated as known. Although this simple method does not take into account uncertainty of class assignment (Clogg, 1995), it minimizes the number of incorrect assignments compared to other approaches (Goodman, 2007).

The second approach is multiple pseudo-class draws (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997). This approach is similar to maximum-probability assignment but accounts for uncertainty in class assignment. Using this method, individuals are classified into latent classes multiple times based on their distributions of posterior probabilities. Often, 20 pseudo-class draws are used; that is, individuals are classified 20 times (Wang, Brown, & Bandeen-Roche, 2005). The subsequent analysis is performed once for each draw (i.e., 20 times) and results are combined across draws using rules derived for multiple imputation for missing data (Rubin, 1987). This technique was originally developed as a diagnostic tool to assess model adequacy (Bandeen-Roche et al., 1997; Wang et al., 2005); that is, the technique was not developed to generate classifications for use in subsequent analysis.

The posterior probabilities upon which both of these approaches depend are calculated from an LCA with a specified number of latent classes. Typically, this LCA only includes manifest indicators and does not include other variables (e.g., moderators, mediators, outcomes) that are included in the subsequent analysis. For example, consider the case of predicting a distal outcome from a latent class variable. A routine approach to this problem is to (1) determine the optimal number of latent classes by fitting and comparing models that only include the manifest indicators of interest, (2) use the parameter estimates from the selected model to calculate posterior probabilities of latent class membership for all individuals (i.e., the classification model), (3) use the posterior probabilities to classify individuals into latent classes using maximum-probability assignment or multiple pseudo-class draws, (4) conduct an analysis to estimate the relation between latent class membership (treated as known) and the distal outcome, for example, by regressing the distal outcome on classification (i.e., the analysis model). With this routine approach, the distal outcome is not included in the classification model (i.e., non-inclusive maximum-probability assignment or non-inclusive multiple pseudo-class draws). From the multiple imputation literature, because the distal outcome was not

included in the classification model but was included in the analysis model, it is expected that the estimated relation between latent class membership and the distal outcome will be attenuated (e.g., Collins, Schafer, & Kam, 2001; Schafer, 1997). In particular, with a non-inclusive approach, we would expect attenuation to increase with the strength of the true relation. This attenuation is starting to be recognized as an important issue (Clark & Muthén, 2009).

We propose an alternative, inclusive approach in which all variables to be included in the analysis model are included as covariates in the classification model. In other words, the LCA used to obtain the posterior probabilities is generalized to include all variables used in the analysis model, ensuring that the imputation (i.e., classification) model is as general as the analysis model. In the current example, the distal outcome would be included as a covariate in the latent class model from which the posterior probabilities are calculated. Because the distal outcome is included in both the classification and analysis models, we expect an inclusive approach (i.e., inclusive maximum-probability assignment or inclusive multiple pseudo-class draws) to produce a more accurate estimate of the relation between latent class membership and the distal outcome. We reiterate, though, that this approach is more general than just the example of using latent class membership to predict a distal outcome.

Purpose of the Current Study

The primary objective of the current study is to compare the performance of the proposed inclusive classify-analyze approach for LCA with the current practice of a non-inclusive approach. First, a motivating empirical demonstration is presented. This demonstration examines the relation between risk exposure latent class membership and later binge drinking. Second, a simulation study based on the demonstration is conducted to examine the inclusive and non-inclusive classify-analyze approaches. This study was designed to mimic the empirical demonstration, and includes five latent classes of risk

exposure and a binary distal outcome. The influence of several factors on the performance of each approach is examined, including quality of the LCA measurement model, strength of the relation between the latent classes and distal outcome (i.e., effect size), and sample size. Both the empirical demonstration and simulation study estimate the effect of latent class membership on the distal outcome using eight different approaches: (1) maximum-probability assignment with a standard (i.e., non-inclusive) LCA; (2) maximum-probability assignment with an inclusive LCA; (3) single pseudo-class draw with a standard LCA; (4) 20 pseudo-class draws with a standard LCA; (5) 40 pseudo-class draws with a standard LCA; (6) single pseudo-class draw with an inclusive LCA; (7) 20 pseudo-class draws with an inclusive LCA; (8) 40 pseudo-class draws with an inclusive LCA.

Empirical Demonstration

The purpose of this empirical demonstration is to illustrate the proposed inclusive classify-analyze approaches for LCA. This demonstration is based on the relatively simple case of predicting a distal outcome from latent class membership. Six manifest variables indicating exposure to various risk factors were used to identify the latent class variable, risk exposure. Risk exposure latent class membership then was used to predict the distal outcome, binge drinking in the past year, using the eight approaches listed above. This demonstration, including the participants and measures, was based on work by Lanza and Rhoades (2011).

Participants

Data were from Wave I and Wave II of the public-use data from the National Longitudinal Study of Adolescent Health (Add Health; Harris, 2009; Harris et al., 2009). The sample consisted of $n = 844$ adolescents who were in 8^{th} grade at Wave I (53% female; *mean* age = 14.5 years, $SD = .86$; 72% White, 20% Black, 3% Asian, 5% Other; 11% Hispanic). Only participants who provided data on exposure to at least one risk factor at

Wave I and provided data on binge drinking at Wave II were included in the sample. This sample was smaller than that used in the original study because the restricted-use data were not included.

Measures

Indicators of Risk Exposure. Measures of the latent class variable, risk exposure, included two indicators of household risk, two indicators of peer risk, and two indicators of neighborhood risk, all assessed at Wave I. For household risk, adolescents were considered to be at risk for *household poverty* if their household income-to-needs ratios were below 1.85; they were considered to be at risk for *single-parent household* if they lived with a parent/caregiver who was widowed, divorced, separated, or never married at the time of assessment. For peer risk, adolescents were considered to be at risk for *peer cigarette use* if one or more of their three best friends smoked at least one cigarette per day; similarly, they were considered to be at risk for *peer alcohol use* if one or more of their three best friends drank alcohol at least once per month. For neighborhood risk, adolescents were considered to be at risk for *neighborhood unemployment* if they lived in a census block where the unemployment rate was greater than 10.9% (Billy, Wenzlow, & Grady, 1998); they were considered to be at risk for *neighborhood poverty* if they lived in a census block where at least 23.9% of the households were living below the poverty level in 1989 (Billy et al., 1998).

Binge Drinking. The distal outcome, binge drinking, was measured using a single indicator, assessed at Wave II. Adolescents were considered to be past-year binge drinkers if they reported drinking five or more drinks in a row on one or more days in the past 12 months; 24.8% of adolescents reported binge drinking.

Analysis

First, LCA was used to confirm that the 5-class model identified by Lanza and Rhoades (2011) was optimal for the public-use sample selected for the current demonstration. Second, posterior probabilities were calculated with the selected model

using both an inclusive LCA (i.e., with binge drinking included as a covariate) and a non-inclusive LCA (i.e., without binge drinking included as a covariate). Third, the eight classify-analyze approaches listed previously were implemented based on posterior probabilities derived from the inclusive and non-inclusive LCAs. These eight approaches were used to calculate the proportion of adolescents reporting past-year binge drinking given latent class membership.

The data analysis was generated using SAS V9 software. Inclusive and non-inclusive LCAs were conducted with PROC LCA (Lanza, Dziak, Huang, Xu, & Collins, 2011); PROC LCA and the corresponding users' guide are available for free download at `methodology.psu.edu/downloads`. Annotated SAS code relating latent class membership to the distal outcome using both inclusive and non-inclusive approaches is available in the Appendix.

Results

First, to confirm that the 5-class model was optimal, LCAs with 1-6 classes were compared based on model fit, parsimony, and stability using the Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwartz, 1989), consistent AIC (CAIC; Bozdogan, 1987), adjusted BIC (a-BIC; Sclove, 1987), bootstrap likelihood ratio test (BLRT; McLachlan & Peel, 2000), entropy R^2 (Celeux & Soromenho, 1996), and G^2 fit statistic. Solution stability was based on the proportion of times the maximum-likelihood solution was selected out of 1000 random sets of starting values (Solution %). As expected, the 5-class model was selected as optimal; a summary of the fit criteria is shown in Table 1.

The overall proportion of adolescents who were exposed to each risk factor and the parameter estimates for the 5-class model are shown in Table 2. The latent classes comprising the 5-class model lent themselves to straightforward interpretations, similar to those by Lanza and Rhoades (2011). The first latent class was labeled Low Risk (with a

prevalence of .41) because members had low probabilities of exposure to all six risk factors. In comparison, the second latent class was labeled Peer Risk (.22) because members had high probabilities of exposure to Peer Cigarette Use and Peer Alcohol Use, but low probabilities of exposure to the other four risk factors. Using a similar approach, the third and fourth latent classes were labeled Economic Risk (.19) and Household & Peer Risk (.13), respectively. The fifth latent class was labeled Multi-Risk (.04) because members had high probabilities of exposure to all six risk factors.

Second, posterior probabilities from the 5-class model were retained as the basis for the non-inclusive approaches. Then, binge drinking was added to the LCA as a covariate in order to generate and retain the posterior probabilities used as the basis for the inclusive approaches. Notably, binge drinking was significantly related to latent class membership ($df = 4$, $p < .0001$). The odds of membership in the Peer Risk ($OR = 5.6$), Economic Risk ($OR = 1.1$), Household & Peer Risk ($OR = 13.2$), and Multi-Risk ($OR = 4.8$) latent classes relative to the Low Risk latent class were higher for binge drinkers compared to those who did not binge drink.

Third, the effect of risk exposure latent class membership on binge drinking was examined using the eight approaches listed previously. Estimates of the proportion of adolescents reporting past-year binge drinking conditional on risk exposure latent class membership are shown in Table 3. Using Cohen's w as a measure of effect size, there was a medium-sized effect of risk exposure on binge drinking ($w = .42$). The proportion of adolescents reporting binge drinking across all latent classes was .25.

As shown in Table 3, adolescents in the Low Risk and Economic Risk latent classes were less likely to report binge drinking compared to adolescents in the Peer Risk, Household & Peer Risk, and Multi-Risk latent classes. The estimates of the proportions, however, differed substantially depending on the approach used. Non-inclusive approaches provided class-specific estimates closer to the marginal proportion of .25 compared to using an inclusive approach. For example, the non-inclusive 20 pseudo-class draws approach

estimated that 39% of Household & Peer Risk adolescents reported binge drinking; the corresponding inclusive approach produced an estimate of 62%. This is consistent with our expectation that proportion estimates are attenuated with a non-inclusive approach.

Additionally, the non-inclusive maximum-probability assignment and multiple pseudo-class draws approaches provided similar estimates, as did the inclusive maximum-probability assignment and multiple pseudo-class draws approaches. For example, using an inclusive approach and a maximum-probability assignment rule, it was estimated that 11% of Low Risk adolescents reported binge drinking, compared to an estimated 11% using the inclusive 20 pseudo-class draws approach. This raises the question of whether multiple pseudo-class draws really do provide more accurate estimates to justify the additional computational burden. To examine the issues of attenuation and performance in detail, a simulation study was conducted using the empirical demonstration as a basis for its design.

Comparing Non-Inclusive and Inclusive Classify-Analyze Approaches: A Simulation Study

The simulation study was designed to assess the performance of the eight classify-analyze approaches for LCA listed previously. The primary objective was to compare the performance of inclusive and non-inclusive approaches. Secondary objectives were to compare maximum-probability assignment to multiple pseudo-class draws, and to compare performance of all approaches across measurement quality conditions, effect sizes, and sample sizes.

Design

Three factors were considered in this study because they were expected to have a large impact on the performance of inclusive and non-inclusive classify-analyze approaches for LCA. First, measurement quality is directly linked to posterior probability estimates; as measurement quality increases, measurement error decreases, and posterior probabilities

move closer to 0 and 1, indicating greater confidence in class assignment. Second, if it is true that effects estimated using non-inclusive approaches are attenuated, as hypothesized, this attenuation will be more pronounced for stronger relations between latent class membership and the distal outcome. Thus, strength of the association between the latent class variable and distal outcome (i.e., effect size) was examined. Third, sample size affects estimation of both the latent class model and the relation between the latent classes and distal outcome; larger samples provide more information for parameter estimation. The simulation was designed to correspond to the empirical demonstration. Five latent classes, corresponding to Low Risk, Peer Risk, Economic Risk, Household & Peer Risk, and Multi-Risk, were measured using six binary indicators. Latent class membership proportions (i.e., γ parameters) were held constant across all conditions: .40 for Low Risk, .20 for Peer Risk, .20 for Economic Risk, .10 for Household & Peer Risk, and .10 for Multi-Risk (see Table 4).

Measurement Quality. Four sets of ρ parameters that represent different levels of measurement quality in the item-response probabilities were considered: real life, high quality, medium quality, and low quality. True values for the ρ parameters for each latent class for each measurement quality condition are shown in Table 4. The 'real life' condition was based on the item-response probability estimates presented in the empirical demonstration (see Table 2), whereas the remaining conditions were designed to reflect high, medium, and low measurement quality while maintaining interpretation.

Effect Sizes. Five Cohen's w effect sizes that represent different strengths of the relation between the latent classes and distal outcome were considered: real life (.42), large effect (.50), medium effect (.30), small effect (.10), and no effect (.00). For each effect size, true values for the proportion of individuals binge drinking in the past year conditional on latent class membership are shown in Table 5. The overall proportion of binge drinking differed somewhat across different effect sizes, ranging from .19 to .30. The 'real life' condition was selected to mimic that of the empirical demonstration, whereas the

remaining conditions were designed to reflect decreasing strength in the relation between the latent classes and distal outcome.

Sample Sizes. Two sample sizes were considered: large ($n = 800$), which roughly equaled that of the empirical demonstration, and small ($n = 400$).

Process

A single cell of the simulation represents one combination of measurement quality, effect size, and sample size conditions. The following Monte Carlo procedure was used in each of the 40 cells of the simulation.

Data Generation. Given (a) the latent class model specified by the latent class membership probabilities and item-response probabilities, (b) the strength of the association between latent classes and distal outcome, and (c) the sample size, random observations were generated by (1) generating a latent class variable from a multinomial distribution specified by the latent class membership proportions, (2) generating item responses based on the item-response probabilities, and (3) generating outcomes based on the multinomial logistic regression model linking latent class membership and the distal outcome. Random observations were generated to create 1000 replicate datasets.

Classification Step. In order to generate posterior probabilities using an inclusive approach and a non-inclusive approach, two LCAs were conducted on each replicate dataset. The inclusive LCA included the distal outcome as a covariate; the non-inclusive LCA did not include the distal outcome as a covariate. To ensure model identification in the non-inclusive LCAs, 100 random sets of starting values were used; parameter estimates from the maximum-likelihood solution were used as starting values for the inclusive LCAs. Then, a maximum-probability assignment rule was used to infer class membership based on posterior probabilities from the non-inclusive LCAs, and again based on posterior probabilities from the inclusive LCAs. Finally, multiple pseudo-class draws (e.g., 1, 20, or 40) were used to infer class membership based on posterior probabilities from the

non-inclusive LCAs, and again based on posterior probabilities from the inclusive LCAs.

Analysis Step. Using each of the eight approaches, the effect of latent class membership on the distal outcome was estimated by calculating the proportion of observations with the distal outcome conditional on latent class membership. The estimated relation between latent class membership and the distal outcome was compared to the true values shown in Table 5.

Results

Inclusive Versus Non-Inclusive Approaches. In order to summarize the results concisely, results for the Household & Peer Risk latent class are discussed. This latent class was small and had the highest prevalence of binge drinking; thus, this set of results is ideal for studying the performance of the eight approaches. Simulation results for this latent class are shown in Table 6 for the small sample size and in Table 7 for the large sample size. Each cell contains the bias (i.e., mean estimated value minus true value) and root mean square error (RMSE; i.e., $\sqrt{\text{bias}^2 + \text{stderr}^2}$ where 'stderr' is the standard error of the 1000 estimated values) for the estimate of the effect of Household & Peer Risk latent class membership on binge drinking (i.e., proportion of adolescents reporting binge drinking). For example, Table 6 shows that for high measurement quality, large effect size, and small sample size, the bias in the estimated proportion of adolescents in the Household & Peer Risk latent class reporting binge drinking was -.119 and -.034 for non-inclusive and inclusive maximum-probability assignment, respectively, and was -.134 and -.045 for non-inclusive and inclusive 20 pseudo-class draws, respectively. In other words, the proportion of adolescents binge drinking was underestimated using the standard approach.

For maximum-probability assignment and all pseudo-class draws approaches, the bias was smaller for the inclusive approaches than the non-inclusive approaches. For example, Table 7 shows that with medium measurement quality, large sample size, and maximum-probability assignment, the inclusive approach resulted in biases of -.007, -.021,

-.010, and -.001 for large, medium, small, and no effect, respectively, compared to biases of -.253, -.136, -.052, and -.002 using the non-inclusive approach. As expected, relying on a standard (i.e., non-inclusive) approach produced estimates of the effect that were substantially attenuated, such that the estimated proportion of adolescents binge drinking conditional on latent class membership was severely underestimated for Household & Peer Risk. The results across all five latent classes showed this pattern consistently[1]. In sum, across all conditions and all latent classes, when a non-inclusive approach was used, the estimated proportions of binge drinking were closer to the overall proportion than the true class-specific proportions (i.e., the association was attenuated); that is, binge drinking was under-reported for the Peer Risk, Household & Peer Risk, and Multi-Risk latent classes, and over-reported for the Low Risk and Economic Risk latent classes.

Maximum-Probability Assignment Versus Pseudo-Class Draws. For both non-inclusive and inclusive approaches, the bias was smaller for maximum-probability assignment compared to multiple pseudo-class draws[2]. For example, Table 7 shows that with high measurement quality, large effect size, and large sample size, the bias for non-inclusive maximum-probability assignment was -.115 and was -.133, -.132, and -.132 for non-inclusive pseudo-class draws with 1, 20, and 40 draws, respectively. Similarly, the biases for the corresponding inclusive approaches were -.012, -.033, -.032, and -.033.

When a non-inclusive approach was used, maximum-probability assignment consistently had smaller RMSEs compared to the pseudo-class draws approaches for large and medium effect sizes, but consistently had larger RMSEs for small and no effect sizes. In contrast, when an inclusive approach was used, the pseudo-class draws approaches consistently had smaller RMSEs compared to maximum-probability assignment, particularly as measurement quality or effect size decreased. For example, with low

[1]Additional results for the other four latent classes are not shown, but are available upon request to the corresponding author.

[2]The one exception is for the inclusive approach with real life measurement quality and a large sample size; the bias for the multiple pseudo-class draws approaches is smaller than that of maximum-probability assignment for all effect sizes.

measurement quality and small sample size, the RMSE for the inclusive maximum-probability assignment approach (see Table 6) was .358, .272, .206, and .184 for large, medium, small, and no effect, respectively, compared to RMSEs of .359, .250, .173, and .154 using the inclusive 20 pseudo-class draws approach. Interestingly, when a non-inclusive approach was used, this tendency was not evident. In addition, comparing 20 and 40 pseudo-class draws for both inclusive and non-inclusive approaches suggests that using 20 draws performed as well as using 40 draws, both in terms of bias and RMSE.

Factors Affecting Performance. As measurement quality increased, all approaches were less biased, but inclusive approaches still performed better than non-inclusive approaches. For example, for a real life effect size, large sample size, and the inclusive 20 pseudo-class draws approach (see Table 7), bias decreased from -.181 to -.065 when measurement quality increased from low to medium, and further decreased from -.065 to -.018 when measurement quality increased from medium to high. A similar pattern was seen for the non-inclusive 20 pseudo-class draws approach (-.308 to -.200 to -.077).

As effect size increased (i.e., strength of the relation between latent class membership and distal outcome increased), attenuation of the estimated effect increased, and was particularly pronounced when non-inclusive approaches were used. For example, for medium measurement quality, large sample size, and the non-inclusive 20 pseudo-class draws approach (see Table 7), bias increased from -.060 to -.156 to -.287 as the effect size increased from small to medium to large. In comparison, for the inclusive approach, it increased from -.021 to -.047 to -.061 as the effect sized increased from small to medium to large.

As sample size increased, all approaches were less biased, but inclusive approaches still performed better than non-inclusive approaches. For example, increasing the sample size from small to large decreased the bias from -.044 to -.003 for real life measurement quality, real life effect size, and the inclusive 20 pseudo-class draws approach; for the non-inclusive approach, the bias decreased from -.157 to -.130 (see Tables 6 and 7).

Importantly, increasing measurement quality appeared to decrease the bias more than doubling the sample size. Overall, even with inclusive approaches, results were quite biased when sample size was small, effect size was large, and measurement quality was low. In contrast, with inclusive approaches, bias was nearly eliminated with medium or high measurement quality and a large sample size, or high measurement quality and a small sample size.

Discussion

Motivated by an examination of the effect of risk exposure latent class membership on binge drinking, this study demonstrated the importance of using an inclusive approach when implementing classify-analyze for LCA. Based on results of the Monte Carlo study, we recommend that applied scientists employ an inclusive LCA for obtaining the posterior probabilities on which classify-analyze approaches are based. Interestingly, our study showed no improvement using multiple pseudo-class draws over maximum-probability assignment when a standard (i.e., non-inclusive) approach was taken; maximum-probability assignment was both less biased and less variable unless the effect was small or non-existent, at which point it was still less biased but slightly more variable. Using an inclusive approach over a non-inclusive one resulted in large improvements in performance, both in terms of bias and variability; using maximum-probability assignment with an inclusive approach was the least biased overall, but using 20 or more pseudo-class draws with an inclusive approach was considerably less variable, particularly with lower levels of measurement quality.

This study showed that a classify-analyze approach for relating latent class membership to other variables of interest can be used to obtain unbiased estimates, provided that an inclusive LCA is used to obtain the posterior probabilities. This is consistent with the literature on multiple imputation for missing data (e.g., Collins et al., 2001) that emphasizes the importance of imputing data under a model that is at least as

general as the subsequent analysis model. We reiterate that this approach is recommended when a classify-analyze approach for LCA is necessary. For example, the model for LCA with covariates is well-understood (e.g., Lanza & Collins, 2008), so addressing questions about predictors of latent class membership need not be addressed using classify-analyze approaches.

Comparison of Approaches

Three general conclusions can be drawn when comparing the different classify-analyze approaches. First, an inclusive approach was less biased than a non-inclusive approach. As expected, when a non-inclusive approach was used, the effect of latent class membership on the distal outcome was attenuated. When an inclusive approach was used, effect estimates were unbiased with sufficiently strong measurement quality or sufficiently large sample size. Second, inclusive and non-inclusive maximum-probability assignment actually outperformed inclusive and non-inclusive multiple pseudo-class draws in terms of bias, but inclusive multiple pseudo-class draws with a sufficiently large number of draws had a smaller RMSE. This suggests that although maximum-probability assignment is less biased overall, multiple pseudo-class draws is less variable across individual datasets when an inclusive approach is used; thus, we recommend this approach for substantive applications. Third, using 40 pseudo-class draws compared to 20 pseudo-class draws did not result in reduced bias or RMSE. Therefore, the standard practice of using 20 pseudo-class draws appears to be sufficient.

Comparison of Latent Class Model Features

The design of the simulation allowed an evaluation of the influence of different features of the latent class model, including measurement quality, effect size, and sample size. As the measurement quality of the latent class model increased, both inclusive and non-inclusive approaches were less biased although inclusive approaches consistently performed better. As the effect of latent class membership on the distal outcome became

stronger, the bias increased for non-inclusive approaches and inclusive approaches; however, inclusive approaches performed considerably better and were unbiased with sufficiently strong measurement quality or sufficiently large sample size. As sample size increased, both inclusive and non-inclusive approaches were less biased although inclusive approaches consistently performed better. Importantly, stronger measurement quality appeared to improve performance markedly more than increasing the sample size.

Recommendations

In general, estimates of the effect of latent class membership on the distal outcome were most biased using a non-inclusive approach with small sample size, low measurement quality, and large effect size. Conversely, results appeared to be unbiased with an inclusive approach with medium or high measurement quality and a large sample size, and with high measurement quality and a small sample size. Combined with previous work on LCA (e.g., Lanza et al., 2007), the results of the current study suggest a series of steps when using classify-analyze approaches for LCA: (1) determine the optimal number of latent classes by fitting and comparing models without covariates included in the model; (2) after the optimal model has been selected and interpreted, re-fit the latent class model with the other variables of interest included as covariates to produce posterior probabilities using an inclusive approach; (3) use multiple pseudo-class draws with at least 20 draws for the 'classification' step of the classify-analyze approach; (4) treat class membership as known to perform the desired analysis for each dataset generated by the multiple pseudo-class draws for the 'analysis' step of the classify-analyze approach; (5) combine results across datasets (i.e., across pseudo-class draws) for the final results.

Software

All modern statistical software packages for LCA, including PROC LCA (Lanza et al., 2011), M*plus* (Muthén & Muthén, 1998-2007), and Latent GOLD (Vermunt & Magidson, 2005), can be used to calculate posterior probabilities of latent class

membership using the non-inclusive approach. Further, other variables of interest can be included as covariates in any of these programs, so that an inclusive classify-analyze approach can be used. Then, posterior probabilities can be used in any statistical software package (e.g., SAS, SPSS, STATA) to assign individuals to latent classes (a) once based on a maximum-probability assignment rule, or (b) multiple times using multiple pseudo-class draws. Sample SAS code for doing this appears in the Appendix.

Classification Error

All classify-analyze approaches to assigning individuals to latent classes for subsequent analysis are based on posterior probabilities derived from the latent class measurement model. Estimates of the relation between latent class membership and other variables of interest are unbiased using an inclusive approach because the posterior probabilities generated from an inclusive LCA provide more accurate classifications than do those from a non-inclusive LCA. Several methods for evaluating classification error (i.e., classification uncertainty) have been proposed in the context of latent class model evaluation. For example, Goodman (2007) discusses two criteria to assess when an assignment procedure is adequate. The criteria can be used to estimate the proportion of incorrect assignments when a maximum-probability assignment rule and single pseudo-class draw are used. In addition, Vermunt and Magidson (2002) use the Goodman-Kruskal lambda and Goodman-Kruskal tau to assess assignment in addition to the proportion of incorrect assignments; they also use several indices that combine information about model fit and classification errors, including entropy R^2, classification likelihood, and approximate weight of evidence (Celeux & Soromenho, 1996). Finally, in the context of classes of developmental trajectories, it has been suggested that when the mean posterior probability of class membership for individuals assigned to each class exceeds .70, hypothesis tests of differences across classes may be unaffected (Roeder, Lynch, & Nagin, 1999; White, Nagin, Replogle, & Stouthamer-Loeber, 2004; Nagin, 2005).

However, in our empirical demonstration, the mean posterior probabilities of membership for individuals assigned to each class ranged from .84 (Household & Peer Risk) to .93 (Economic Risk) for the non-inclusive maximum-probability assignment approach, and this approach showed considerable bias in the simulation study. In comparison, the mean posterior probabilities ranged from .82 (Household & Peer Risk) to .92 (Economic Risk) for the inclusive maximum-probability assignment approach, and this approach performed very well in the simulation study. In sum, regardless of whether classifications can be considered 'adequate' or 'satisfactory' based on these criteria, we have demonstrated that relations between latent class membership classification and other variables will be attenuated if these variables are not included in the model used to generate the posterior probabilities upon which the classifications are based.

Limitations

There are two primary limitations to the current simulation study. First, as with any simulation study, conclusions about the results are limited to the set of conditions that were examined. For example, the current study did not examine different distributions of latent class membership proportions. Thus, our findings cannot be generalized to the case where, for example, latent classes are of equal size. Future simulation studies should consider how relative class size may interact with other factors in determining the performance of inclusive versus non-inclusive approaches.

Second, this study only examined the relatively simple case of using classify-analyze to predict a single distal outcome from latent class membership. Research questions posing complex relations, such as a latent class variable as a moderator or mediator, would require that multiple variables be included as covariates in the classification step. Extrapolating from the results here and the known impact of imputing data under a classification model that is more restrictive than the analysis model (e.g., Schafer, 1997; Collins et al., 2001), we believe that failure to include all additional variables of interest, along with relevant

interactions, in the estimation of the posterior probabilities will result in significant attenuation of the effects in the subsequent analysis step. However, the magnitude of attenuation in a variety of more complex scenarios needs to be investigated in future studies.

Conclusions

As research questions regarding the role of latent class membership in developmental processes become more complex, it is increasingly difficult to address all questions within the context of the latent class model itself. Addressing questions within the context of the latent class model itself is desirable because it allows measurement error to be estimated and removed from parameter estimates; this approach does not require the classification of individuals to latent classes. When a question cannot be addressed in this context, however, a classify-analyze approach is required. The current study showed that standard, non-inclusive classify-analyze approaches for LCA that do not include other variables of interest in the model during the classification step produce substantially attenuated effect estimates in the analysis step. We demonstrated a straightforward solution, that of fitting an inclusive LCA to derive posterior probabilities, which can be readily adopted by scientists to reduce or eliminate bias in the associations between a latent class variable and other variables of interest. This approach opens the door to broaden modeling approaches when a latent class variable is embedded in a complex model.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

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.

Biemer, P. P., & Wiesen, C. (2002). Measurement error evaluation of self-reported drug use: A latent class analysis of the US National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *165*(1), 97-119.

Billy, J. O. G., Wenzlow, A. T., & Grady, W. R. (1998). *The National Longitudinal Study of Adolescent Health public use contextual database* [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Bozdogan, H. (1987). Model selection and Akaike Information Criterion (AIC): The general theory and its analytical extension. *Psychometrika*, *52*, 345-370.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*, 195-212.

Chung, H., Flaherty, B. P., & Schafer, J. L. (2006). Latent class logistic regression: Application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society, Series A*, *169*(Part 4), 723-743.

Clark, S. L., & Muthén, B. O. (2009). *Relating latent class anlaysis results to variables not included in the analysis*. Manuscript submitted for publication. Available for download at http://www.statmodel.com/papers.shtml.

Clogg, C. C. (1995). Latent class models: Recent developments and prospects for the future. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311-359). New York, NY: Plenum Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral and health sciences*. Hoboken, NJ: John Wiley & Sons, Inc.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330-351.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. New York, NY: Taylor & Francis.

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I – A modified latent structure approach. *American Journal of Sociology*, *79*, 1179-1259.

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231.

Goodman, L. A. (2007). On the assignment of individuals to latent classes. *Sociological Methodology*, *37*(1), 1-22.

Hardigan, P. C., & Sangasubana, N. (2010). A latent class analysis of job satisfaction and turnover among practicing pharmacists. *Research in Social and Administrative Pharmacy*, *6*, 32-38.

Harris, K. M. (2009). *The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994-1996; Wave III, 2001-2002; Wave IV, 2007-2009* [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Harris, K. M., Halpern, C. T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., & Udry, J. R. (2009). *The National Longitudinal Study of Adolescent Health: Research design* [WWW document]. URL: http://www.cpc.unc.edu/projects/addhealth/design. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

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

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

Lanza, S. T., Dziak, J. J., Huang, L., Xu, S., & Collins, L. M. (2011). *Proc LCA & Proc LTA users' guide* (Version 1.2.6). University Park, PA: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu.

Lanza, S. T., & Rhoades, B. L. (2011). Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*. Advanced online publication. doi: 10.1007/s11121-011-0201-1.

Loken, E. (2004). Using latent class analysis to model temperament types. *Multivariate Behavioral Research*, *39*(4), 625-652.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley and Sons, Inc.

Muthén, L. K., & Muthén, B. O. (1998-2007). *M*plus *user's guide* (Fifth ed.). Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nylund, K., Bellmore, A., Nishina, A., & Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say? *Child Development*, *78*(6), 1706-1722.

Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 69-100). New York, NY: Springer.

Reboussin, B. A., Song, E.-Y., Shrestha, A., Lohman, K. K., & Wolfson, M. (2006). A latent class analysis of underage problem drinking: Evidence from a community sample of 16-20 year olds. *Drug and Alcohol Dependence*, *83*(3), 199-209.

Reinke, W. M., Herman, K. C., Petras, H., & Ialongo, N. S. (2008). Empirically derived subtypes of child academic and behavior problems: Co-occurrence and distal outcomes. *Journal of Abnormal Child Psychology*, *36*, 759-770.

Roberts, T. J., & Ward, S. E. (2011). Using latent transition analysis in nursing research to explore change over time. *Nursing Research*, *60*(1), 73-79.

Roeder, K., Lynch, K., & Nagin, D. S. (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association*, *94*, 766-776.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey research*. New York, NY: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

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

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333-343.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89-106). New York, NY: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 users' guide.* Belmont, MA: Statistical Innovations.

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.

White, H. R., Nagin, D. S., Replogle, E., & Stouthamer-Loeber, M. (2004). Racial differences in trajectories of cigarette use. *Drug and Alcohol Dependence*, *76*, 219-227.

Appendix

Syntax for SAS Software

The data analysis was generated using SAS V9 software. Non-inclusive and inclusive LCAs were conducted with PROC LCA (Lanza et al., 2011); PROC LCA and the corresponding users' guide are available for free download at `methodology.psu.edu/downloads`. Annotated code to assign individuals to latent classes and perform the analysis relating latent class membership to the distal outcome is available below.

Step 1: Retain Posterior Probabilities from Non-inclusive LCA

```
PROC LCA DATA=ca.mbrlca outparam=lca_start_ni outpost=lca_post_ni;
  TITLE1 'Five Risk Classes, Non-inclusive';
  ID newaid bio_sex yr2_binge_gv;
  NCLASS 5;
  ITEMS HH_poverty single peer_cig peer_alc unemp below_pov;
  CATEGORIES 2 2 2 2 2 2;
  SEED 452035948;
RUN;
```

Step 2: Retain Posterior Probabilities from Inclusive LCA

```
PROC LCA DATA=ca.mbrlca outparam=lca_start_i outpost=lca_post_i;
  TITLE1 'Five Risk Classes, Inclusive';
  ID newaid bio_sex yr2_binge_gv;
  NCLASS 5;
  ITEMS HH_poverty single peer_cig peer_alc unemp below_pov;
  CATEGORIES 2 2 2 2 2 2;
  SEED 452035948;
  COVARIATES yr2_binge_gv;
  REFERENCE 3;
RUN;
```

Step 3: Assign Individuals to Latent Classes

```
*This code is for non-inclusive approaches because it uses posterior;
*probabilities from the non-inclusive LCA;
*For inclusive approaches use the posterior probabilities from the;
*inclusive LCA instead;

DATA ni_assigned;
```

```
SET lca_post_ni;

*First, maximum-probability assignment rule;
  IF MAX(OF postlc1,postlc2,postlc3,postlc4,postlc5)=postlc1
    THEN classify=1; ELSE
  IF MAX(OF postlc1,postlc2,postlc3,postlc4,postlc5)=postlc2
    THEN classify=2; ELSE
  IF MAX(OF postlc1,postlc2,postlc3,postlc4,postlc5)=postlc3
    THEN classify=3; ELSE
  IF MAX(OF postlc1,postlc2,postlc3,postlc4,postlc5)=postlc4
    THEN classify=4; ELSE
  IF MAX(OF postlc1,postlc2,postlc3,postlc4,postlc5)=postlc5
    THEN classify=5;

*Second, 20 pseudo-class draws;
  ARRAY m20impute(20) m20imp1-m20imp20;
  CALL STREAMINIT(841525);
  DO i = 1 to 20;
    m20impute[i]=RAND('table',postlc1,postlc2,postlc3,postlc4,postlc5);
    END;
  OUTPUT;
*Note that to use a different number of draws (e.g., 1 or 40), change;
*the dimension of the array and the 'do' loop;

RUN;
```

Step 4: Analyze Using Class Assignment Treated as Known

```
*First, maximum-probability assignment rule;
PROC FREQ DATA = ni_assigned;
  TABLES classify*yr2_binge;
RUN;

*Second, 20 pseudo-class draws;
%MACRO imputations(var,data);
  PROC FREQ DATA = ni_assigned;
    TABLES &var*yr2_binge;
    ODS OUTPUT CROSSTABFREQS = &data;
  RUN;

  DATA freqsyes&data (KEEP = classnum rp&data code);
  SET &data;
    IF yr2_binge = 1 AND &var IN (1,2,3,4,5);
    RENAME &var=classnum;
    RENAME rowpercent=rp&data;
    code=_n_;
```

```
  RUN;
%MEND;

%imputations(m20imp1,d1);
%imputations(m20imp2,d2);
%imputations(m20imp3,d3);
.
.
.
%imputations(m20imp18,d18);
%imputations(m20imp19,d19);
%imputations(m20imp20,d20);

DATA freqsyes;
MERGE
     freqsyesd1  freqsyesd2  freqsyesd3  freqsyesd4  freqsyesd5
     freqsyesd6  freqsyesd7  freqsyesd8  freqsyesd9  freqsyesd10
     freqsyesd11 freqsyesd12 freqsyesd13 freqsyesd14 freqsyesd15
     freqsyesd16 freqsyesd17 freqsyesd18 freqsyesd19 freqsyesd20;
BY classnum code;
averagerp = MEAN( rpd1, rpd2, rpd3, rpd4, rpd5, rpd6, rpd7, rpd8, rpd9,
                  rpd10,rpd11,rpd12,rpd13,rpd14,rpd15,rpd16,rpd17,rpd18,
                  rpd19,rpd20);
RUN;

PROC PRINT DATA=freqsyes;
  VAR averagerp;
  BY classnum;

*Note that to use a different number of draws (e.g., 1 or 40), change;
*the number of times macro is invoked, and expand the 'DATA freqsyes';
*statements to include all 'freqsyesd' datasets and 'rpd' variables;

RUN;
```

Table 1
Model Fit Information for LCAs With 1-6 Latent Classes

Classes	*df*	AIC	BIC	CAIC	a-BIC	BLRT	Entropy R^2	G^2	Solution %
1	57	699.20	727.63	733.63	708.58	—	1.00	687.20	100
2	50	365.40	426.99	439.99	385.71	$< .001$	.71	339.40	69.5
3	43	179.78	274.54	294.54	211.03	$< .001$	.78	139.78	100
4	36	140.50	268.43	295.43	182.68	$< .001$	.86	86.50	19.1
5	**29**	**112.83**	**273.93**	**307.93**	**165.95**	$<$ **.001**	**.80**	**44.83**	**25.8**
6	22	113.90	308.16	349.16	177.96	.185	.81	31.90	19.7

Note: Dashes indicate criterion was not applicable to the model. Bold font indicates selected model. Solution % refers to the percentage of times the maximum-likelihood solution was selected out of 1000 random sets of starting values.

Table 2
Parameter Estimates for Five-Class LCA

		Low Risk	Peer Risk	Econ Risk	H&P Risk	Multi-Risk
Latent Class Membership Proportions		.41	.22	.19	.13	.04
Indicator	*Overall Proportion*	*Item-Response Proportions*				
HH Below Poverty	.37	.24	.00	.68	1.0	.47
HH Single-Parent	.29	.15	.14	.49	.58	.52
Peer Cigarette Use	.38	.00	.88	.15	.89	1.0
Peer Alcohol Use	.42	.16	.77	.21	.77	1.0
NH Unemployment	.23	.06	.06	.68	.19	1.0
NH Below Poverty	.24	.01	.03	.81	.22	.97

Note: Econ=Economic; H&P=Household & Peer; HH=Household; NH=Neighborhood.

Table 3
Parameter Estimates for the Relation Between Risk Exposure and Binge Drinking

	Approach		Low Risk	Peer Risk	Econ Risk	H&P Risk	Multi-Risk
1	Maximum-Probability	Non-inclusive	.16	.39	.18	.38	.44
2	Maximum-Probability	Inclusive	.11	.42	.12	.60	.36
3	Pseudo-Class 01 Draw	Non-inclusive	.16	.35	.16	.44	.35
4	Pseudo-Class 20 Draws	Non-inclusive	.16	.37	.17	.39	.41
5	Pseudo-Class 40 Draws	Non-inclusive	.15	.37	.17	.39	.42
6	Pseudo-Class 01 Draw	Inclusive	.10	.41	.11	.67	.35
7	Pseudo-Class 20 Draws	Inclusive	.11	.41	.12	.62	.36
8	Pseudo-Class 40 Draws	Inclusive	.11	.41	.12	.61	.36
			Overall Proportion = .25				

Note: Econ=Economic; H&P=Household & Peer; Maximum-Probability=Assignment based on maximum posterior probability; Non-Inclusive=Outcome was not included in LCA to generate posterior probabilities on which approach was based; Inclusive=Outcome was included in LCA to generate posterior probabilities on which approach was based. Table entries represent the probabilities of past-year binge drinking conditional on latent class membership.

Table 4
Patterns of γ and ρ Parameters for Simulation Measurement Quality Conditions

Indicator	Low Risk	Peer Risk	Econ Risk	H&P Risk	Multi-Risk
	γ Parameters				
	.4	.2	.2	.1	.1
Real Life Measurement Quality ρ Parameters					
HH Below Poverty	.2	.1	.7	.9	.5
HH Single-Parent	.2	.1	.5	.6	.5
Peer Cigarette Use	.1	.9	.1	.9	.9
Peer Alcohol Use	.2	.8	.2	.8	.9
NH Unemployment	.1	.1	.7	.2	.9
NH Below Poverty	.1	.1	.8	.2	.9
High Measurement Quality ρ Parameters					
HH Below Poverty	.1	.1	.9	.9	.9
HH Single-Parent	.1	.1	.9	.9	.9
Peer Cigarette Use	.1	.9	.1	.9	.9
Peer Alcohol Use	.1	.9	.1	.9	.9
NH Unemployment	.1	.1	.9	.1	.9
NH Below Poverty	.1	.1	.9	.1	.9
Medium Measurement Quality ρ Parameters					
HH Below Poverty	.2	.2	.8	.7	.7
HH Single-Parent	.2	.2	.8	.7	.7
Peer Cigarette Use	.2	.8	.2	.7	.7
Peer Alcohol Use	.2	.8	.2	.7	.7
NH Unemployment	.2	.2	.8	.3	.7
NH Below Poverty	.2	.2	.8	.3	.7
Low Measurement Quality ρ Parameters					
HH Below Poverty	.3	.3	.7	.7	.7
HH Single-Parent	.3	.3	.7	.7	.7
Peer Cigarette Use	.3	.7	.3	.7	.7
Peer Alcohol Use	.3	.7	.3	.7	.7
NH Unemployment	.3	.3	.7	.3	.7
NH Below Poverty	.3	.3	.7	.3	.7

Note: Econ=Economic; H&P=Household & Peer; HH=Household; NH=Neighborhood. Table entries for the ρ parameters represent the probabilities of exposure to the risk factor conditional on latent class membership.

Table 5
Patterns of Relation Between Latent Classes and Outcome for Simulation Effect Sizes

Label	Effect Size	Low Risk	Peer Risk	Econ Risk	H&P Risk	Multi-Risk	Overall Proportion
Real Life	.42	.10	.40	.10	.60	.40	.24
Large Effect	.50	.05	.30	.05	.67	.30	.19
Medium Effect	.30	.12	.30	.12	.48	.30	.21
Small Effect	.10	.23	.30	.23	.37	.30	.27
No Effect	.00	.30	.30	.30	.30	.30	.30

Note: Econ=Economic; H&P=Household & Peer. Table entries represent the probabilities of past-year binge drinking conditional on latent class membership.

Table 6

Simulation Results for Relation Between Latent Classes and Outcome for Latent Class Four (Household & Peer Risk): Small Sample Size ($n = 400$)

		Bias (RMSE)							
		1	2	3	4	5	6	7	8
ρ	ES	MP-NI	MP-I	PC01-NI	PC20-NI	PC40-NI	PC01-I	PC20-I	PC40-I
RL	RL	-.140 (.184)	-.028 (.193)	-.160 (.197)	-.157 (.192)	-.157 (.192)	-.046 (.186)	-.044 (.182)	-.045 (.182)
RL	Lg	-.192 (.233)	-.008 (.207)	-.219 (.256)	-.218 (.250)	-.218 (.250)	-.041 (.202)	-.038 (.197)	-.038 (.197)
RL	Med	-.107 (.151)	-.010 (.176)	-.120 (.160)	-.121 (.156)	-.120 (.155)	-.026 (.166)	-.027 (.162)	-.026 (.162)
RL	Sm	-.045 (.095)	-.013 (.146)	-.051 (.099)	-.051 (.091)	-.051 (.091)	-.017 (.142)	-.017 (.134)	-.017 (.134)
RL	Zero	.000 (.077)	.002 (.132)	.000 (.078)	.000 (.068)	.000 (.068)	.003 (.130)	.002 (.123)	.002 (.123)
High	RL	-.078 (.110)	-.026 (.102)	-.087 (.118)	-.087 (.115)	-.086 (.115)	-.035 (.105)	-.034 (.103)	-.034 (.103)
High	Lg	-.119 (.145)	-.034 (.110)	-.135 (.160)	-.134 (.157)	-.134 (.156)	-.046 (.117)	-.045 (.113)	-.045 (.113)
High	Med	-.054 (.093)	-.010 (.101)	-.063 (.099)	-.063 (.096)	-.063 (.096)	-.018 (.098)	-.018 (.097)	-.018 (.097)
High	Sm	-.020 (.075)	-.004 (.087)	-.023 (.077)	-.024 (.073)	-.024 (.073)	-.007 (.088)	-.008 (.084)	-.008 (.084)
High	Zero	-.002 (.072)	-.001 (.090)	-.002 (.071)	-.003 (.068)	-.003 (.068)	-.002 (.090)	-.002 (.087)	-.002 (.087)
Med	RL	-.208 (.244)	-.085 (.241)	-.224 (.253)	-.223 (.250)	-.223 (.250)	-.110 (.235)	-.108 (.232)	-.108 (.232)
Med	Lg	-.294 (.330)	-.099 (.273)	-.319 (.348)	-.317 (.345)	-.317 (.344)	-.140 (.273)	-.136 (.270)	-.136 (.270)
Med	Med	-.157 (.188)	-.062 (.202)	-.172 (.198)	-.171 (.194)	-.171 (.194)	-.084 (.192)	-.082 (.189)	-.082 (.189)
Med	Sm	-.064 (.098)	-.034 (.146)	-.069 (.101)	-.069 (.095)	-.069 (.095)	-.042 (.135)	-.042 (.133)	-.042 (.133)
Med	Zero	.000 (.069)	.001 (.133)	.001 (.073)	-.000 (.060)	-.000 (.060)	.001 (.123)	-.001 (.117)	-.000 (.117)
Low	RL	-.303 (.323)	-.155 (.314)	-.310 (.329)	-.309 (.325)	-.309 (.324)	-.197 (.303)	-.195 (.300)	-.195 (.300)
Low	Lg	-.415 (.432)	-.201 (.358)	-.426 (.441)	-.426 (.438)	-.425 (.438)	-.258 (.362)	-.254 (.359)	-.254 (.359)
Low	Med	-.226 (.242)	-.116 (.272)	-.233 (.248)	-.233 (.244)	-.233 (.244)	-.149 (.254)	-.148 (.250)	-.148 (.250)
Low	Sm	-.085 (.115)	-.035 (.206)	-.086 (.118)	-.089 (.109)	-.089 (.109)	-.050 (.179)	-.052 (.173)	-.052 (.174)
Low	Zero	-.004 (.077)	.006 (.184)	-.002 (.079)	-.004 (.065)	-.003 (.065)	.003 (.158)	.002 (.154)	.002 (.154)

Note: RMSE=Root Mean Square Error; ES=Effect Size; RL=Real Life (Measurement Quality or Effect Size); High=High Measurement Quality; Med=Medium (Measurement Quality or Effect Size); Low=Low Measurement Quality; Lg=Large Effect Size; Sm=Small Effect Size; Zero=No Effect; MP-NI=Maximum-Probability Assignment Non-Inclusive; MP-I=Maximum-Probability Assignment Inclusive; PC01-NI=Pseudo-Class 01 Draws Non-Inclusive; PC20-NI=Pseudo-Class 20 Draws Non-Inclusive; PC40-NI=Pseudo-Class 40 Draws Non-Inclusive; PC01-I=Pseudo-Class 01 Draws Inclusive; PC20-I=Pseudo-Class 20 Draws Inclusive; PC40-I=Pseudo-Class 40 Draws Inclusive.

Table 7

Simulation Results for Relation Between Latent Classes and Outcome for Latent Class Four (Household & Peer Risk): Large Sample Size (n = 800)

		Bias (RMSE)							
		1	2	3	4	5	6	7	8
ρ	ES	MP-NI	MP-I	PC01-NI	PC20-NI	PC40-NI	PC01-I	PC20-I	PC40-I
RL	RL	-.106 (.131)	.020 (.130)	-.130 (.152)	-.130 (.148)	-.130 (.149)	-.005 (.117)	-.003 (.112)	-.004 (.112)
RL	Lg	-.158 (.179)	.042 (.135)	-.196 (.214)	-.193 (.209)	-.194 (.210)	-.004 (.118)	-.002 (.112)	-.002 (.112)
RL	Med	-.083 (.108)	.022 (.127)	-.103 (.123)	-.102 (.120)	-.103 (.120)	-.002 (.109)	-.001 (.106)	-.001 (.106)
RL	Sm	-.035 (.068)	.003 (.101)	-.040 (.072)	-.042 (.066)	-.042 (.066)	-.003 (.094)	-.003 (.091)	-.003 (.091)
RL	Zero	.002 (.058)	.004 (.097)	.002 (.057)	.001 (.049)	.001 (.048)	.003 (.091)	.003 (.087)	.003 (.086)
High	RL	-.065 (.088)	-.005 (.075)	-0.078 (.098)	-.077 (.095)	-.077 (.095)	-.018 (.076)	-.018 (.073)	-.018 (.073
High	Lg	-.115 (.132)	-.012 (.087)	-0.133 (.148)	-.132 (.146)	-.132 (.146)	-.033 (.089)	-.032 (.086)	-.033 (.086
High	Med	-.059 (.081)	-.009 (.075)	-0.069 (.090)	-.069 (.087)	-.069 (.087)	-.021 (.075)	-.021 (.072)	-.022 (.072
High	Sm	-.019 (.055)	.000 (.068)	-0.023 (.055)	-.023 (.053)	-.023 (.053)	-.005 (.064)	-.005 (.063)	-.005 (.062
High	Zero	.000 (.047)	-.001 (.061)	0.001 (.046)	.000 (.043)	.000 (.043)	.001 (.058)	.000 (.056)	.000 (.056)
Med	RL	-.172 (.197)	-.035 (.184)	-.202 (.221)	-.200 (.217)	-.200 (.217)	-.068 (.176)	-.065 (.174)	-.066 (.174)
Med	Lg	-.253 (.278)	-.007 (.204)	-.288 (.306)	-.287 (.304)	-.287 (.304)	-.064 (.192)	-.061 (.190)	-.061 (.190)
Med	Med	-.136 (.156)	-.021 (.163)	-.158 (.173)	-.156 (.169)	-.156 (.169)	-.050 (.150)	-.047 (.147)	-.048 (.147)
Med	Sm	-.052 (.075)	-.010 (.115)	-.060 (.080)	-.060 (.075)	-.060 (.075)	-.020 (.103)	-.021 (.100)	-.021 (.100)
Med	Zero	-.002 (.050)	-.001 (.099)	-.001 (.050)	-.002 (.042)	-.002 (.041)	-.000 (.090)	-.002 (.086)	-.002 (.086)
Low	RL	-.297 (.314)	-.121 (.312)	-.310 (.324)	-.308 (.320)	-.308 (.320)	-.185 (.296)	-0.181 (.292)	-.181 (.292)
Low	Lg	-.404 (.418)	-.113 (.325)	-.422 (.433)	-.421 (.430)	-.421 (.430)	-.201 (.317)	-0.199 (.316)	-.199 (.316)
Low	Med	-.228 (.242)	-.090 (.275)	-.236 (.247)	-.236 (.245)	-.236 (.245)	-.138 (.246)	-0.136 (.243)	-.136 (.243)
Low	Sm	-.085 (.106)	-.034 (.203)	-.089 (.106)	-.090 (.102)	-.090 (.101)	-.058 (.157)	-0.058 (.154)	-.058 (.154)
Low	Zero	-.000 (.060)	.016 (.188)	-.002 (.059)	-.001 (.044)	-.001 (.044)	.004 (.134)	0.005 (.131)	.005 (.131)

Note: RMSE=Root Mean Square Error; ES=Effect Size; RL=Real Life (Measurement Quality or Effect Size); High=High Measurement Quality; Med=Medium (Measurement Quality or Effect Size); Low=Low Measurement Quality; Lg=Large Effect Size; Sm=Small Effect Size; Zero=No Effect; MP-NI=Maximum-Probability Assignment Non-Inclusive; MP-I=Maximum-Probability Assignment Inclusive; PC01-NI=Pseudo-Class 01 Draws Non-Inclusive; PC20-NI=Pseudo-Class 20 Draws Non-Inclusive; PC40-NI=Pseudo-Class 40 Draws Non-Inclusive; PC01-I=Pseudo-Class 01 Draws Inclusive; PC20-I=Pseudo-Class 20 Draws Inclusive; PC40-I=Pseudo-Class 40 Draws Inclusive.