A Technical Introduction: A Model-Based Approach to Latent Class Analysis With Distal Outcomes

Stephanie T. Lanza, Xianming Tan & Bethany C. Bray

The Pennsylvania State University

Technical Report Series
#11-116

College of Health and Human Development
The Pennsylvania State University

Running head: TECHNICAL INTRODUCTION: LCA WITH DISTAL OUTCOMES

A Technical Introduction: A Model-Based Approach to Latent Class Analysis With Distal Outcomes

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

[1]The Methodology Center, The Pennsylvania State University

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

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

† Tan is currently a Biostatistical Consultant at the Research Institute of the McGill University Health Center.

Abstract

A model-based approach is proposed to empirically derive and summarize the class-dependent density functions of distal outcomes with categorical, continuous, or count distributions. A Monte Carlo simulation study is conducted to compare the performance of the new technique to two commonly used classify-analyze techniques: maximum-probability assignment and multiple pseudo-class draws. Simulation results show that the model-based approach produces substantially less biased estimates of the effect compared to either classify-analyze technique, particularly when the association between the latent class variable and the distal outcome is strong. In addition, we show that only the model-based approach is consistent. Sample SAS syntax for implementing this approach using PROC LCA and a corresponding macro are provided.

A Technical Introduction: A Model-Based Approach to Latent Class Analysis With Distal Outcomes

When an observed predictor is used to predict latent class membership, the mathematical model is well understood. LCA with covariates has been described in detail in the literature (see Collins & Lanza, 2010; Lanza, Collins, Lemmon, & Schafer, 2007) and is summarized below. However, in the current study we are interested in an effect in the opposite direction, in which the predictor is latent and the outcome is manifest (i.e., predicting a distal outcome from latent class membership). To be more precise, we are interested in the conditional distribution of a distal outcome, Z, given a latent class variable, C. In this case, the problem is more difficult because the predictor (true subgroup membership) is unknown (see Figure 1; Lanza, Collins, Schafer, & Flaherty, 2005).

The two most common approaches to LCA with a distal outcome are the maximum-probability assignment rule (Nagin, 2005) and the multiple pseudo-class draws approach (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Wang, Brown, & Bandeen-Roche, 2005). Because these two classify-analyze approaches involve assigning (i.e., imputing) latent class membership and conducting the outcome analysis in separate steps, conclusions drawn about the effect of C on Z may be incorrect for several reasons. First, there is uncertainty related to class membership, which is not taken into account in the maximum-probability assignment rule. Second, and more importantly, all standard classify-analyze approaches impute the latent variable under a model that is not sufficiently general; this may result in attenuated estimates of the relation between C and Z.

We propose a new model-based approach to LCA with distal outcomes that is flexible in terms of the metric of Z and straightforward to implement. After a brief introduction to the latent class model, we introduce a model-based approach to LCA with a distal outcome and perform a simulation study to demonstrate its performance relative to classify-analyze approaches.

A Brief Review of the Latent Class Model

The latent class model, which is described in detail by Collins and Lanza (2010) and Lanza et al. (2007), can be summarized as follows. Suppose that there are K latent subgroups that must be inferred from $j = 1, ..., J$ observed variables, and that variable j has $r_j = 1, ..., R_j$ response categories. Let $x = (r_1, ..., r_J)$ represent the vector of a particular subject's responses to the J variables. Let C represent the latent variable with latent classes $c = 1, ..., K$. Finally, $I(x_j = r_j)$ is an indicator function that equals 1 when the response to variable $j = r_j$, and equals 0 otherwise. The probability of observing a particular response pattern is

$$\Pr\{X = x\} = \sum_{c=1}^{K} \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(x_j=r_j)}, \tag{1}$$

where γ_c represents the probability of membership in latent class c and $\rho_{j,r_j|c}^{I(x_j=r_j)}$ represents the probability of response r_j to item j given membership in latent class c.

This model can be extended to include covariates (i.e., predictors of latent class membership) using a logistic regression model in which the outcome is a categorical latent variable (see (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Collins & Lanza, 2010; Dayton & Macready, 1988)). Suppose that a covariate U is used to predict latent class membership. Then the latent class model can be expressed as

$$\Pr\{X = x|U = u\} = \sum_{c=1}^{K} \gamma_c(u) \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(x_j=r_j)}, \tag{2}$$

where $\gamma_c(u) = \Pr\{C = c|U = u\}$ is a standard baseline-category multinomial logistic model (e.g., Agresti, 2002).

With a single covariate U, $\gamma_c(u)$ can be expressed as

$$\gamma_c(u) = \Pr\{C = c|U = u\} = \frac{e^{\beta_{0c}+\beta_{1c}u}}{1 + \sum_{c'=1}^{K-1} e^{\beta_{0c'}+\beta_{1c'}u}} \tag{3}$$

for $c' = 1, ..., K - 1$ and reference class K.

Individuals' posterior probabilities of membership in each latent class can be obtained from the resultant LCA parameters by applying Bayes' theorem (e.g., Gelman, Carlin, Stern, & Rubin, 2003; Lanza et al., 2007):

$$\Pr\{C = c|U = u\} = \frac{\Pr\{C = c\}\Pr\{U = u|C = c\}}{\Pr\{U = u\}}. \tag{4}$$

A model with a particular number of latent classes can be selected using a bootstrap likelihood-ratio test (McLachlan & Peel, 2000; McLachlan, 1987), as well as information criteria such as AIC (Akaike, 1974), BIC (Schwartz, 1978), CAIC (Bozdogan, 1987), and a-BIC (Sclove, 1987). Multiple sets of random starting values should be used to assess the degree of certainty that the global maximum (as opposed to a local maximum) in the likelihood function has been identified. In addition, the ability to interpret the latent classes in a solution can help guide model selection.

Effect sizes in LCA. It is possible to calculate an effect size (Cohen, 1992) indicating the strength of association between a latent class variable C and a distal outcome Z. The effect size is calculated as follows:

- For a categorical outcome Z with m categories,

$$\omega = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{K}\frac{(P_{ij} - P_{0ij})^2}{P_{0ij}}},$$

where $P_{ij} = \Pr\{Z = i, C = j\} = \pi_j \Pr\{Z = i|C = j\}$, $P_{0ij} = \Pr\{Z = i\}$. We note that $\omega = 0$ if and only if $P_{ij} = \Pr\{Z = i, C = j\} = \Pr\{Z = i\}$. That is, $\omega = 0$ if and only if C and Z are independent.

- For a continuous or count outcome,

$$\omega = \sqrt{\sum_{c=1}^{K}\pi_c(\mu_c - \overline{\mu})^2},$$

where $\pi_c = \Pr\{C = c\}$, $\mu_c = E(Z|C = c)$, and $\overline{\mu} = E(Z) = \sum_{c=1}^{K}\pi_c E(Z|C = c) = \sum_{c=1}^{K}\pi_c\mu_c$.

The estimated effect size will vary depending on whether a model-based approach, maximum-probability assignment, or a multiple pseudo-class draws approach is used to estimate the effect. In addition, for a continuous distal outcome, while it is typical to use the conditional mean when calculating the effect size, when the distribution of $Z|C$ is skewed we may instead use the mode when calculating the effect size.

A Model-Based Approach to Predict a Distal Outcome from Latent Class Membership

Let us first restate the problem more precisely. We have multiple observed indicators X, a distal outcome Z, and a latent class variable C. We assume that (X, C) follows an LCA model with a fixed number of classes. Although C is not observable, we wish to estimate the conditional distribution of the distal outcome for each latent class $(Z|C)$. However, without certain assumptions regarding the joint distribution of (X, Z, C), the estimation of $Z|C$ is not possible. In general, the joint distribution of random variables is not identifiable from their marginal distributions alone (Casella & Berger, 1990).

An Important Assumption: Conditional Independence Between X and Z Given C

In order to be able to estimate the conditional distribution of Z given C, $f(Z|C)$, we propose making the assumption of conditional independence between X and Z given the latent class variable C. That is, we assume that $f(X, Z|C) = f(X|C)f(Z|C)$. Although there might be alternative assumptions which can also resolve the non-identifiability issue, we prefer this conditional independence assumption for its similarity to the local independence assumption underlying most LCA models (Collins & Lanza, 2010).

For completeness, the assumptions underlying the proposed model-based approach to LCA with distal outcomes can be explicitly listed as follows. First, we assume that in addition to the observed response indicator variables X and distal outcome Z, there exists a latent class variable C, and the marginal distribution of the latent class variable C is $\Pr\{C = c\} = \pi_c$ $(c = 1, 2, ..., K)$, with $0 < \pi_c < 1$ $(c = 1, 2, ..., K)$ and $\sum_{c=1}^{K} \pi_c = 1$. Second, we assume that the conditional

distribution of X given C is implied by the fundamental LCA model, defined above. Third, we assume that the conditional distribution of C given Z can be summarized by a logistic regression model:

$$\Pr\{C = c \mid Z = z\} = \frac{e^{\beta_{0c}+\beta_{1c}z}}{1 + \sum_{c'=1}^{K-1} e^{\beta_{0c'}+\beta_{1c'}z}}.$$

Assuming a logistic regression model for predicting C from a covariate is quite reasonable, and is standard practice in the LCA literature (e.g., Vermunt & Magidson, 2005) .

Modeling the Latent Class Variable and the Effect of C on Z Simultaneously

In LCA with a distal outcome, interest lies in the density $f\{Z = z|C = c\}$. We can determine the desired distribution of $Z|C$ by applying Bayes' Theorem:

$$f\{Z = z|C = c\} = \frac{f\{Z = z\} \times f\{C = c|Z = z\}}{f\{C = c\}}.$$

Given the assumptions above, $f\{C = c\}$ is determined by the LCA model and $f\{C = c|Z = z\}$ is determined by the LCA model with Z included as a covariate. The final piece of necessary information, $f\{Z = z\}$, is the marginal distribution of Z, which can be estimated using the empirical distribution of Z. In the following, we first present how to estimate $f\{Z = z|C = c\}$ when Z is a binary distal outcome, and then discuss extending this approach to categorical outcomes with more than two categories and to count outcomes; we then present an approach for estimating the conditional distribution of a continuous Z. No assumption about the particular distributional form of Z, such as Gaussian, is required.

Prediction of a binary/categorical/count distal outcome. When Z is binary, including Z as an additional indicator in the LCA model, including Z as a grouping variable in the LCA model, and incorporating Z into the LCA model as a covariate are mathematically equivalent. All of these approaches require the assumption of conditional independence between X and Z given C Roeder et al. (1999). We recommend the third approach of incorporating Z as a covariate because it can be readily extended to other types of distal outcomes without requiring distributional

assumptions of Z. Then, the density of concern, $f\{Z = z|C = c\}$, can be expressed as

$$\Pr\{Z = z|C = c\} = \frac{\Pr\{Z = z\}e^{\beta_{0c}+\beta_{1c}z}}{\Pr\{C = c\}(1 + \sum_{c'=1}^{K-1} e^{\beta_{0c'}+\beta_{1c'}z})}.$$

Using this approach, $\Pr\{Z = z\}$ is estimated from the empirical distribution of Z (i.e., from the proportions in the observed data); the estimates for $\{\beta_{0c}, \beta_{1c}; c = 1, 2, ..., K - 1\}$ are provided by the LCA with covariates model; and the marginal distribution $\Pr\{C = c\}$ can be obtained by multiplying $\Pr\{C = c \mid Z = z\}$ by the marginal distribution $\Pr\{Z = z\}$. If one uses PROC LCA (Lanza, Dziak, Huang, Xu, & Collins, 2011), $\Pr\{C = c\}$ is part of the default output even when the model includes a covariate. Thus, we can estimate $\Pr\{Z = z|C = c\}$ given these estimates for $\Pr\{Z = z\}$, $\Pr\{C = c\}$, and $\{\beta_{0c}, \beta_{1c}; c = 1, 2, ..., K - 1\}$.

An Excel calculator has recently been published online (Lanza & Rhoades, 2011b) so that analysts can implement this approach to LCA with a binary distal outcome in their work. The calculator uses as inputs the logistic regression coefficients (β_{1c}) and the known marginal probabilities of the binary distal outcome; the calculator then provides the probabilities of Z given C. This approach is demonstrated in the corresponding article by Lanza and Rhoades (2011a).

The arguments used for a binary distal outcome, described above, can be extended to a categorical outcome with more than two categories (i.e., $Z \in \{1, 2, 3..., m\}$ and $m \geq 2$), if we assume that

$$\Pr\{C = c \mid Z = i\} = \frac{e^{\beta_{0c}+\beta_{1ci}}}{1 + \sum_{c'=1}^{K-1} e^{\beta_{0c'}+\beta_{1c'i}}}, \quad \text{for } i = 2, 3, ..., m.$$

This is equivalent to using Z as a grouping variable; in this case $\Pr\{C = c \mid Z = i\}$ is a group-specific mixing proportion.

The model for LCA with a binary distal outcome can also be extended to a count type outcome with more than two categories (i.e., $Z \in \{0, 1, 2, 3, ...\}$), if we assume that

$$\Pr\{Z = z|C = c\} = \frac{\Pr\{Z = z\}e^{\beta_{0c}+\beta_{1c}z}}{\Pr\{C = c\}(1 + \sum_{c'=1}^{K-1} e^{\beta_{0c'}+\beta_{1c'}z})}, \quad z = 0, 1, 2, ... \ .$$

In this approach, $\Pr\{Z = z\}$ is also estimated from the empirical distribution of Z, instead of

assuming a certain conditional distribution for $Z|C$, such as a conditional Poisson distribution $Z|(C = c) \sim \text{Poisson}(\lambda_c)$.

Prediction of a continuous distal outcome. Obtaining the distribution of a continuous distal outcome given C is a more complicated case than that of a categorical Z. We propose extending the approach described above for a binary/categorical/count distal outcome to continuous outcomes. Similar to the binary/categorical/count case, using this approach we are able to obtain estimates for $\{\beta_{0c}, \beta_{1c}; c = 1, 2, ..., K-1\}$ from the LCA with covariates model. Then, to estimate $f\{Z = z|C = c\}$ we need to estimate $f\{Z = z\}$, and the marginal distribution $\Pr\{C = c\}$ can be obtained by multiplying $\Pr\{C = c \mid Z = z\}$ by the marginal distribution $f\{Z = z\}$. As mentioned above, this is part of the standard output of PROC LCA. We estimate the density of Z using kernel density estimates (Silverman, 1986) for continuous variables, which can be readily implemented using SAS PROC KDE (SAS Institute Inc., 2002-2004). The default bandwidth selection method in PROC KDE is based on the plug-in formula of Sheather and Jones, as suggested in Jones, Marron, and Sheather (1996). In sum, we propose a flexible, semi-parametric approach for modeling the effect of C on a continuous Z, in which we empirically estimate the distribution of Z. Using the conditional and marginal distributions we can obtain the mean (or mode) of Z for each latent class. Again, this approach does not require a specification of the conditional distribution of Z given C, such as a conditional normal distribution $Z|(C = c) \sim N(\mu_c, \sigma^2)$; instead, it uses the empirical distribution of Z.

Software. LCA, as well as the proposed model-based approach to LCA with a distal outcome, can be conducted in SAS. Syntax for conducting LCA with a distal outcome is included in the Appendix. The SAS procedure for conducting latent class analysis, PROC LCA (Lanza, Dziak, Huang, Xu, & Collins, 2011), and the new %LCA_distal macro (Tan, Lanza, & Wagner, 2011), are available for download at methodology.psu.edu.

In order to examine the properties of this model-based approach to LCA with a distal outcome, we now move to a simulation study. The impact of four factors on performance of this technique is examined for binary, count, and continuous outcomes. Performance of the proposed

model-based approach is compared to that of maximum-probability assignment and multiple pseudo-class draws.

A Comparison of Three Estimation Methods for LCA with a Distal Outcome

Design

In this simulation study, we examined the effect of four factors on the performance of the model-based approach, as well as the two classify-analyze approaches, to LCA with a distal outcome. The factors were the conditional distribution of the distal outcome, Z; the strength of the association between the latent class variable and the distal outcome (i.e., effect size); the quality of the LCA measurement model (i.e., the degree of association between the observed and latent variables, which in this case corresponds to the degree of separation between latent classes); and the sample size. Specifically, the levels of the factors considered were as follows.

Type of Z. Three types of the distal outcome were considered: binary, continuous, and count. (Categorical distal outcomes were not considered in the current simulation study.) In our simulation, we let $Z|C = c \sim \text{Binom}(p_c)$ for binary Z; $Z|C = c \sim N(\mu_c, 1)$ for continuous Z; and $Z|C = c \sim \text{Poisson}(\lambda_c)$ for count Z. We hypothesized that any attenuation observed when a model-based approach is not used would be present regardless of the distribution of Z.

Strength of the effect of C on Z. For each Z distribution listed above, four strengths of association between the latent class variable and the distal outcome were considered. These corresponded to no effect, weak effect, medium effect, and strong effect as defined by Cohen (1992). The corresponding population values of p_c (for binary Z), μ_c (for continuous Z), and λ_c (for count Z), are listed in the top, middle and lower panel of Table 1, respectively. We hypothesized that attenuation of the effect of C on Z would increase as the effect size increases, and that this attenuation would be much smaller for the model-based approach as compared to the two classify-analyze approaches.

LCA measurement model. Using the empirical example of latent classes of adolescent depression described in Lanza, Flaherty, & Collins (2003) as a basis, latent class models with eight binary indicators and five latent classes were considered. We specified latent class prevalences and measurement models that had a structure similar to that in the empirical study. For all models in this simulation study, the proportion of individuals in Classes 1 through 5 were specified to be 40%, 20%, 20%, 10%, and 10%, respectively. Two levels of measurement quality were considered: moderate, characterized by item-response probabilities equal to .8 or .2, and high, characterized by item-response probabilities equal to .9 or .1. Table 2 shows the set of item-response probabilities specified to achieve these two levels of measurement. We hypothesized that high measurement quality would reduce bias under any other combination of factors, regardless of estimation method.

Sample size. We considered sample sizes of 500 and 1000. Assessing performance for very small sample sizes was not a goal of this study; rather, we were interested in examining whether any benefits are achieved by increasing n from a moderate size to a large size. We hypothesized that there would be little difference in the results when comparing sample sizes of 500 and 1000.

The fully crossed factorial design consisted of 48 conditions. For each condition, we implemented three approaches for estimating the effect of C on Z: the proposed model-based approach, the maximum-probability assignment approach, and the multiple (in this case, 20) pseudo-class draws approach. For each condition, we replicated the analysis 1000 times and summarized the simulation outputs to assess how each factor affected performance of the three approaches.

Procedure

The following Monte Carlo procedure was used in each of the 48 simulation design cells.

Step 1: Generation of LCA data. Given the specified LCA model (i.e., latent class prevalences and item-response probabilities) and the specified strength of association between C and Z, to generate one random observation, we first generated a latent class variable C from a

multinomial distribution specified by the latent class prevalences (i.e., mixing proportions); we then generated item responses based on the item-response probabilities (i.e., ρ parameters) for that cell, and then generated the distal outcome Z based on the C-Z model for that cell.

Step 2: LCA model fitting. For each replicate data set, two different LCA models were fit. The first model included no distal outcome Z (for the maximum-probability assignment and pseudo-class draws approaches), and the second model included the distal outcome Z as a covariate (for the model-based approach). We used 100 sets of random starting values for the LCA model that did not include Z in order to avoid local maxima and for an examination of model identification. The parameter estimates from the model that did not include Z were used as starting values for the LCA model with Z as a covariate.

Step 3: Calculation of Z given C for each approach. Given the LCA results derived in Step 2, along with the random sample, the estimation of the effect of C on Z was conducted for each approach. For the model-based approach we employed the procedure described above, which relies on the β, γ and ρ parameters from the LCA model with Z included as a covariate. For maximum-probability assignment and multiple pseudo-class draws, we first inferred the latent classes C for each observation using the corresponding approaches (described above), and then in a subsequent model we estimated the effect of C on Z. For the pseudo-class draws approach, this final step was repeated 20 times and results were combined across draws.

Step 4: Summary of results. The goal of this step was to summarize results across the 1000 replicate data sets in order to draw comparisons between the three methods of estimation. For each approach, we first compared the estimated effect of C on Z to the true effect, shown in Table 1, and then summarized the results across replications to obtain the bias and root mean squared error (RMSE) for each parameter estimate. This step required that we address the issue that the ordering of the latent classes is random across the 1000 replicates. To impose a standard order on the latent classes, we wrote a SAS macro to take the LCA estimates and true LCA model parameters as inputs, then reordered the latent classes based on distance calculations comparing

the estimated LCA parameters and the true LCA model parameters.

Results

Tables 3, 4, and 5 show simulation results for the binary, continuous, and count outcomes, respectively. Within each table, we present results for $n = 500$ in the top panel and for $n = 1000$ in the bottom panel. Moderate measurement quality is shown on the left side, and high measurement quality on the right side. For each effect size (zero, small, medium, large), we present results based on the three analytic approaches: the proposed model-based method (Model), maximum-probability assignment (Assign), and multiple pseudo-class draws (P-C). Each cell reflects the bias (i.e., mean estimated value minus true value) in the estimate of Z given C. For example, Table 3 shows that for moderate measurement quality, $n = 1000$, and large effect size, the bias in the estimated proportion of individuals in each latent class with a 1 on the binary outcome was 0.003, -0.002, -0.016, -0.061, and -0.010 for Latent Classes 1, 2, 3, 4, and 5, respectively. Recall from Table 1 that the true proportions for this cell were 0.006, 0.153, 0.300, 0.447, and 0.594. Negative values of bias indicate that the class-specific prevalence of the outcome is underestimated. For the same set of conditions, the bias was from 2 to 10 times larger for the maximum-probability assignment (0.035, 0.022, -0.105, -0.115, -0.091) and the multiple pseudo-class draws (0.042, 0.023, -0.112, -0.130, -0.120) approaches.

Several general patterns emerged across results for the binary, continuous, and count distal outcomes. First, as expected, when the effect size was set to zero, all three methods performed equally well, in that bias was less than 0.01 for each latent class regardless of sample size, measurement quality, or method. Second, because the prevalence of Latent Class 1 was considerably larger than that of other latent classes (0.4; see Table 2), bias was consistently smaller for this latent class. This was expected because, all other factors held constant, there is more information available related to larger latent classes, making estimation more accurate. Similarly, the bias was consistently larger for the smaller latent classes (Latent Classes 4 and 5) because there was less information available for estimation. Third, as expected, as the strength of the association between the latent class variable and the distal outcome strengthened, the

potential for bias increased, and – importantly – the benefits of using a model-based approach became more significant. Fourth, when the methods performed differentially, the model-based approach consistently performed better than the two classify-analyze approaches. In every case, the impact of using either maximum-probability assignment or multiple pseudo-class draws was manifested by an attenuation of the effect of C on Z. That is, the more negative biases seen in the two classify-analyze approaches confirmed our hypothesis that these methods would result in underestimation of the distal outcome for the latent classes that are furthest from the mean on Z.

A somewhat surprising finding was that maximum-probability assignment worked at least as well as the multiple pseudo-class draws technique in terms of bias/attenuation of the effect of C on Z. This suggests that, in the long run, this simple classify-analyze approach is preferable to the pseudo-class draws approach. However, the variability in the estimates across the 1000 replicates for the maximum-probability assignment approach was higher than that for the multiple pseudo-class draws approach (not shown). Therefore, in empirical studies the pseudo-class draws approach may be more reliable than maximum-probability assignment. Regardless of this fact, however, the model-based approach introduced here performed substantially better than either of the standard classify-analyze techniques.

One final important finding is that, in addition to the model-based approach being less biased in the long run, this new method was shown to be consistent. That is, as n increased, bias was reduced. However, sample size had essentially no effect on performance of the maximum-probability assignment or pseudo-class draws methods; neither classify-analyze strategy appeared to be consistent.

In sum, improving measurement quality (i.e., moving from item-response probabilities of .2 and .8 to probabilities of .1 and .9) had a substantial impact for all methods, such that bias was reduced consistently by more than half for all methods. As discussed above, as the effect size between C and Z increased, the potential for bias increased. With larger effect sizes, attenuation increased much more in the two classify-analyze approaches than it did in the model-based approach. All of these patterns emerged consistently for all types (binary, continuous, and count)

of distal outcome. Thus, the model-based approach proposed here outperformed maximum-probability assignment and multiple pseudo-class draws under every condition.

We next move to an empirical demonstration of the model-based approach to LCA with a distal outcome. The motivating example involves latent classes of depression in adolescence. Three distal outcomes are included for demonstration purposes: a binary outcome (regular smoking), a continuous outcome (grades), and a count outcome (delinquency).

Conclusions

By applying Bayes' theorem, we can capture information from a model that is well-understood (LCA with covariates) and transform it into information that addresses this exact research question. This is the foundation for the flexible model-based approach proposed here. The critical pieces of information come from two sources. First, a latent class model is specified with the distal outcome as a covariate in order to obtain the logistic regression coefficients reflecting their association. Second, the class-conditional marginal density of Z is estimated, for example using a kernel density estimation approach. The SAS macro %LCA_distal (Tan, Lanza, & Wagner, 2011), introduced here for estimating LCA with distal outcomes that are categorical, continuous, or count variables, automates this approach.

A Monte Carlo simulation study was conducted to compare the performance of this new approach to two classify-analyze approaches: maximum-probability assignment and multiple pseudo-class draws. Simulation results show that the model-based approach produces substantially less biased estimates of the effect compared to either classify-analyze technique, particularly when the association between the latent class variable and the distal outcome is strong. Although the RMSE was larger for the model-based approach in the case of no or small effect size, as the strength of the effect of C on Z increased the relative performance reversed, such that the model-based approach had smaller RMSE. Taken together, when a moderate to strong relation exists between the latent class variable and the distal outcome, we recommend the model-based approach because of its lower bias and lower RMSE. In addition, we show that only the model-based approach exhibits the property of consistency (i.e., its performance improves as

n increases).

In addition, we made several hypotheses regarding the factors examined in the simulation study. We expected the performance of the model-based approach to be superior to that of both classify-analyze approaches, regardless of the metric of the distal outcome (categorical, continuous, and count). This was consistently supported in the simulation study. Our hypothesis that the attenuation of effects would increase as the effect size increased was confirmed. In addition, improving measurement quality resulted in better performance (i.e., less bias) for the model-based approach and for both classify-analyze approaches. As expected, we observed no improvement in the performance of either classify-analyze approach as sample size increased. For the model-based approach, however, performance did improve as sample size increased, suggesting that this method is statistically consistent.

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.

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

Casella, G., & Berger, R. L. (1990). *Statistical inference*. Belmont, CA: Duxbury.

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.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173-178.

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

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*, 401-407.

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., Collins, L. M., Schafer, J. L., & Flaherty, B. P. (2005). Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods*, *10*, 84-100.

Lanza, S. T., Dziak, J. J., Huang, L., Xu, S., & Collins, L. M. (2011). *Proc LCA & Proc LTA users' guide* (Version 1.2.7). University Park, PA: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu.

Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2003). Latent class and latent transition analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2, research methods in psychology* (p. 663-685). Hoboken, NJ : Wiley.

Lanza, S. T., & Rhoades, B. L. (2011a). 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.

Lanza, S. T., & Rhoades, B. L. (2011b). *LCA outcome probability calculator* (Version 1.0). University Park, PA : The Methodology Center, Penn State. Retrieved from The Methodology Center: http://methodology.psu.edu/ra/lcalta/calculator.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *36*(3), 318-324.

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

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

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.

SAS Institute Inc. (2002-2004). *SAS 9.1.3 help and documentation.* Cary, NC : SAS Institute Inc.

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.

Silverman, B. W. (1986). *Density estimation.* New York, NY: Chapman & Hall.

Tan, X., Lanza, S. T., & Wagner, A. T. (2011). *LCA distal SAS macro users guide* (Version 1.1.0). University Park, PA: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu.

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.

Appendix

SAS Syntax

This appendix provides SAS syntax for implementing the model-based approach to estimating the probability of regular smoking in Grade 12 (Z) conditional on a depression latent class variable (C) comprised of five classes and indicated by eight binary items.

```
*Estimate latent class model with binary distal outcome Z included as covariate;
proc lca data=outcomes start=baseline_start outparam=estimates_cigZ;
  nclass 5;
  items w1fs3 w1fs6 w1fs13 w1fs16 w1fs9 w1fs19 w1fs14 w1fs17;
  categories 2 2 2 2 2 2 2 2;
  covariates cig_t2;
  reference 4;
run;

*Execute macro to obtain distribution of Z given C;
%LCA_distal(input_data = outcomes,       /*input random sample*/
  param = estimates_cigZ,                /*dataset generated by outparam in PROC LCA*/
  distal = cig_t2,                       /*distal outcome variable*/
  metric = 1,                            /*1=binary, 2=continuous, 3=count, 4=categorical*/
  output_dataset_name= Cig_results       /*output results*/
);
```

Author Note

This project 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. The authors thank Aaron T. Wagner for helpful comments on an earlier draft of this technical report.

Table 1
Patterns of $Z|C$: Specified true values for the distal outcome given latent class membership in the simulation study

	Latent Class					
	1	2	3	4	5	Effect Size
			Binary Z			
$\Pr\{Z\|C\}$	0.300	0.300	0.300	0.300	0.300	$= 0.0$
$\Pr\{Z\|C\}$	0.234	0.267	0.300	0.333	0.366	≈ 0.1
$\Pr\{Z\|C\}$	0.110	0.205	0.300	0.395	0.490	≈ 0.3
$\Pr\{Z\|C\}$	0.006	0.153	0.300	0.447	0.594	≈ 0.5
			Continuous Z (Conditional Normal)			
$E\{Z\|C\}$	0.00	0.00	0.00	0.00	0.00	$= 0.0$
$E\{Z\|C\}$	-0.14	-0.07	0.00	0.07	0.14	≈ 0.1
$E\{Z\|C\}$	-0.38	-0.19	0.00	0.19	0.38	≈ 0.3
$E\{Z\|C\}$	-0.64	-0.32	0.00	0.32	0.64	≈ 0.5
			Count Z (Conditional Poisson)			
$E\{Z\|C\}$	0.80	0.80	0.80	0.80	0.80	$= 0.0$
$E\{Z\|C\}$	0.66	0.73	0.80	0.87	0.94	≈ 0.1
$E\{Z\|C\}$	0.42	0.61	0.80	0.99	1.18	≈ 0.3
$E\{Z\|C\}$	0.16	0.48	0.80	1.12	1.44	≈ 0.5

Table 2
Patterns of item-response probabilities: Two conditions for item-response probabilities specified in the simulation study

	Latent Class				
	1	2	3	4	5
LC Membership Probabilities	0.4	0.2	0.2	0.1	0.1
Moderate Measurement Quality					
Could note shake blues	0.2	0.8	0.2	0.8	0.8
Felt depressed	0.2	0.8	0.2	0.8	0.8
Felt lonely	0.2	0.8	0.2	0.8	0.8
Felt sad	0.2	0.8	0.2	0.2	0.8
People unfriendly	0.2	0.2	0.8	0.2	0.8
Disliked by people	0.2	0.2	0.8	0.8	0.8
Life was failure	0.2	0.2	0.2	0.2	0.8
Life not worth living	0.2	0.2	0.2	0.2	0.8
High Measurement Quality					
Could not shake blues	0.1	0.9	0.1	0.9	0.9
Felt depressed	0.1	0.9	0.1	0.9	0.9
Felt lonely	0.1	0.9	0.1	0.9	0.9
Felt sad	0.1	0.9	0.1	0.9	0.9
People unfriendly	0.1	0.1	0.9	0.9	0.9
Disliked by people	0.1	0.1	0.9	0.9	0.9
Life was failure	0.1	0.1	0.1	0.1	0.9
Life not worth living	0.1	0.1	0.1	0.1	0.9

Table 3

Simulation results for LCA with a binary distal outcome: Bias in proportion with $Z = 1$ given latent class membership

		Moderate Measurement					Strong Measurement				
		Latent Class					Latent Class				
Method	ES	1	2	3	4	5	1	2	3	4	5
						$n = 500$					
Model	Zero	-0.003	-0.000	0.009	0.005	0.006	-0.000	0.002	0.002	0.002	-0.002
Assign	Zero	-0.002	0.001	0.002	0.002	0.004	-0.000	0.002	0.003	0.002	-0.002
P-C	Zero	-0.002	0.000	0.003	0.002	0.004	-0.000	0.002	0.002	0.002	-0.001
Model	Sm	-0.000	0.001	-0.017	-0.017	-0.009	-0.000	-0.001	-0.012	-0.009	-0.003
Assign	Sm	0.008	0.007	-0.025	-0.031	-0.025	0.003	0.002	-0.017	-0.014	-0.007
P-C	Sm	0.010	0.008	-0.027	-0.034	-0.030	0.003	0.002	-0.018	-0.015	-0.009
Model	Med	0.001	0.005	-0.033	-0.072	-0.023	0.000	0.001	-0.030	-0.028	-0.005
Assign	Med	0.023	0.020	-0.066	-0.097	-0.078	0.009	0.006	-0.050	-0.041	-0.018
P-C	Med	0.028	0.020	-0.071	-0.104	-0.091	0.010	0.006	-0.053	-0.044	-0.022
Model	Lg	0.010	0.003	-0.037	-0.115	-0.029	0.001	-0.001	-0.034	-0.027	-0.008
Assign	Lg	0.039	0.028	-0.103	-0.150	-0.113	0.012	0.008	-0.074	-0.056	-0.026
P-C	Lg	0.046	0.030	-0.111	-0.159	-0.135	0.013	0.009	-0.080	-0.061	-0.032
						$n = 1000$					
Model	Zero	-0.000	-0.001	0.003	-0.002	0.001	0.001	-0.000	0.001	0.003	0.001
Assign	Zero	0.000	-0.000	0.002	-0.003	0.001	0.002	-0.000	0.000	0.002	0.001
P-C	Zero	0.000	-0.000	0.002	-0.002	0.000	0.001	0.000	0.001	0.002	0.001
Model	Sm	-0.003	0.001	-0.013	-0.011	-0.004	0.000	-0.002	-0.014	-0.009	0.000
Assign	Sm	0.008	0.007	-0.023	-0.023	-0.021	0.003	0.001	-0.019	-0.014	-0.004
P-C	Sm	0.009	0.006	-0.025	-0.027	-0.028	0.003	0.001	-0.020	-0.015	-0.006
Model	Med	-0.001	-0.001	-0.030	-0.040	-0.008	0.001	-0.002	-0.020	-0.019	-0.003
Assign	Med	0.023	0.015	-0.070	-0.077	-0.061	0.009	0.006	-0.048	-0.037	-0.016
P-C	Med	0.028	0.015	-0.075	-0.086	-0.079	0.010	0.006	-0.051	-0.041	-0.021
Model	Lg	0.003	-0.002	-0.016	-0.061	-0.010	0.000	0.001	-0.018	-0.022	-0.005
Assign	Lg	0.035	0.022	-0.105	-0.115	-0.091	0.012	0.011	-0.075	-0.056	-0.024
P-C	Lg	0.042	0.023	-0.112	-0.130	-0.120	0.013	0.010	-0.081	-0.062	-0.032

Note: Model = model-based approach; Assign = maximum-probability assignment rule; P-C = multiple pseudo-class draws.

Table 4
Simulation results for LCA with a continuous distal outcome: Bias in mean given latent class membership

		Moderate Measurement					Strong Measurement				
		Latent Class					Latent Class				
Method	ES	1	2	3	4	5	1	2	3	4	5
						$n = 500$					
Model	Zero	0.004	-0.002	0.002	-0.001	0.002	-0.000	-0.001	-0.001	-0.007	-0.006
Assign	Zero	0.003	-0.001	0.001	-0.000	0.002	0.000	-0.002	-0.002	-0.006	-0.005
P-C	Zero	0.002	-0.000	0.001	-0.002	-0.001	-0.000	-0.001	-0.002	-0.006	-0.005
Model	Sm	-0.005	0.005	-0.034	-0.052	-0.010	-0.004	-0.001	-0.022	-0.018	0.010
Assign	Sm	0.018	0.019	-0.046	-0.073	-0.057	0.006	0.004	-0.035	-0.031	-0.009
P-C	Sm	0.022	0.018	-0.052	-0.078	-0.065	0.006	0.004	-0.037	-0.034	-0.011
Model	Med	-0.016	0.005	-0.056	-0.145	-0.018	-0.015	-0.004	-0.059	-0.042	0.012
Assign	Med	0.044	0.038	-0.131	-0.199	-0.147	0.012	0.009	-0.098	-0.082	-0.040
P-C	Med	0.054	0.040	-0.138	-0.212	-0.176	0.014	0.009	-0.105	-0.087	-0.047
Model	Lg	-0.015	0.012	-0.111	-0.218	-0.043	-0.014	0.001	-0.067	-0.045	0.029
Assign	Lg	0.082	0.061	-0.224	-0.314	-0.251	0.028	0.024	-0.165	-0.127	-0.053
P-C	Lg	0.097	0.065	-0.240	-0.334	-0.298	0.032	0.023	-0.177	-0.140	-0.066
						$n = 1000$					
Model	Zero	0.000	0.002	0.001	0.003	-0.006	-0.001	0.005	-0.001	-0.000	0.004
Assign	Zero	-0.000	0.001	0.001	0.001	-0.003	-0.001	0.004	0.000	-0.000	0.003
P-C	Zero	-0.000	0.001	-0.000	0.003	-0.003	-0.001	0.004	-0.000	-0.000	0.004
Model	Sm	-0.006	-0.001	-0.028	-0.040	0.002	-0.008	-0.003	-0.019	-0.013	0.005
Assign	Sm	0.018	0.011	-0.048	-0.061	-0.043	0.003	0.004	-0.032	-0.028	-0.013
P-C	Sm	0.022	0.012	-0.052	-0.066	-0.058	0.003	0.003	-0.035	-0.031	-0.017
Model	Med	-0.019	-0.001	-0.065	-0.076	0.003	-0.014	-0.004	-0.037	-0.021	0.017
Assign	Med	0.044	0.034	-0.135	-0.155	-0.117	0.015	0.013	-0.096	-0.072	-0.034
P-C	Med	0.054	0.035	-0.146	-0.172	-0.154	0.016	0.012	-0.104	-0.080	-0.043
Model	Lg	-0.027	-0.002	-0.067	-0.136	0.017	-0.016	0.002	-0.033	-0.031	0.022
Assign	Lg	0.074	0.053	-0.225	-0.266	-0.190	0.027	0.027	-0.156	-0.128	-0.057
P-C	Lg	0.090	0.053	-0.243	-0.292	-0.254	0.029	0.025	-0.170	-0.138	-0.075

Note: Model = model-based approach; Assign = maximum-probability assignment rule; P-C = multiple pseudo-class draws.

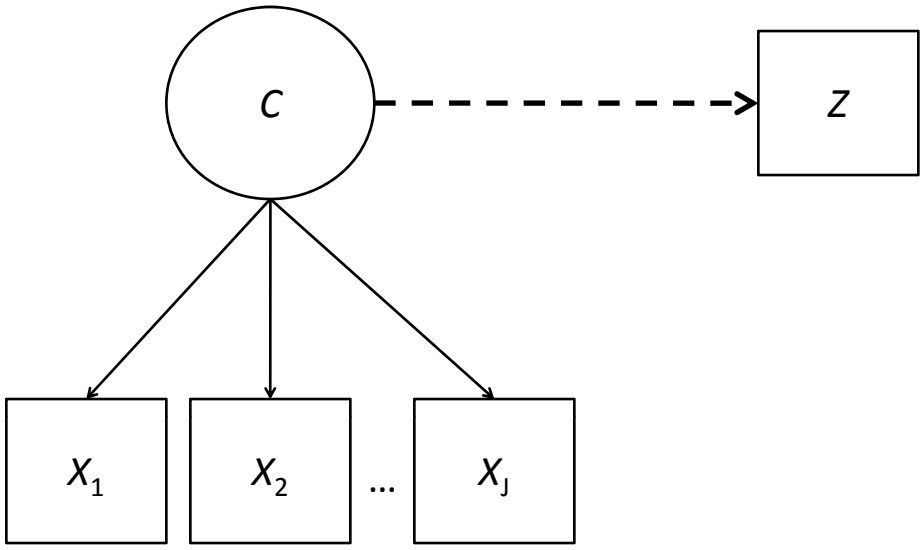
Table 5
Simulation results for LCA with a count distal outcome: Bias in mean count given latent class membership

		Moderate Measurement					Strong Measurement				
		Latent Class					Latent Class				
Method	ES	1	2	3	4	5	1	2	3	4	5
						$n = 500$					
Model	Zero	-0.005	0.003	0.015	0.004	0.014	-0.000	-0.000	0.002	-0.004	-0.003
Assign	Zero	-0.002	0.003	0.007	0.001	0.007	0.000	-0.001	0.001	-0.003	-0.003
P-C	Zero	-0.001	0.002	0.006	0.001	0.008	-0.000	-0.001	0.001	-0.002	-0.003
Model	Sm	0.003	-0.001	-0.035	-0.062	-0.025	-0.002	0.001	-0.029	-0.020	-0.008
Assign	Sm	0.019	0.012	-0.051	-0.074	-0.064	0.005	0.007	-0.039	-0.027	-0.018
P-C	Sm	0.022	0.011	-0.055	-0.076	-0.073	0.006	0.006	-0.042	-0.029	-0.021
Model	Med	0.001	-0.001	-0.065	-0.133	-0.051	0.001	-0.000	-0.052	-0.060	-0.002
Assign	Med	0.047	0.032	-0.132	-0.189	-0.155	0.018	0.012	-0.093	-0.086	-0.030
P-C	Med	0.056	0.035	-0.142	-0.201	-0.182	0.020	0.012	-0.099	-0.093	-0.036
Model	Lg	0.004	0.010	-0.072	-0.217	-0.060	0.000	-0.001	-0.060	-0.064	-0.002
Assign	Lg	0.083	0.062	-0.222	-0.323	-0.241	0.027	0.019	-0.167	-0.124	-0.046
P-C	Lg	0.099	0.065	-0.235	-0.345	-0.288	0.029	0.019	-0.180	-0.135	-0.059
						$n = 1000$					
Model	Zero	0.005	-0.001	-0.004	0.000	-0.003	-0.001	-0.002	0.001	0.002	-0.001
Assign	Zero	0.003	-0.001	-0.003	-0.002	-0.002	-0.000	-0.001	0.000	0.002	0.000
P-C	Zero	0.002	-0.000	-0.002	-0.001	-0.001	-0.000	-0.002	0.001	0.002	-0.000
Model	Sm	-0.005	0.001	-0.030	-0.039	0.003	-0.002	0.001	-0.019	-0.023	0.003
Assign	Sm	0.015	0.012	-0.051	-0.056	-0.039	0.005	0.007	-0.033	-0.031	-0.008
P-C	Sm	0.019	0.013	-0.055	-0.062	-0.052	0.006	0.006	-0.036	-0.034	-0.012
Model	Med	-0.007	-0.001	-0.048	-0.080	-0.011	-0.002	-0.003	-0.038	-0.038	-0.002
Assign	Med	0.044	0.030	-0.131	-0.145	-0.118	0.016	0.012	-0.093	-0.077	-0.029
P-C	Med	0.054	0.032	-0.141	-0.166	-0.153	0.018	0.011	-0.102	-0.084	-0.040
Model	Lg	-0.001	0.003	-0.030	-0.133	-0.029	-0.001	-0.002	-0.029	-0.044	-0.007
Assign	Lg	0.075	0.049	-0.219	-0.252	-0.201	0.026	0.020	-0.158	-0.128	-0.050
P-C	Lg	0.091	0.050	-0.237	-0.285	-0.264	0.028	0.018	-0.171	-0.139	-0.068

Note: Model = model-based approach; Assign = maximum-probability assignment rule; P-C = multiple pseudo-class draws.

Figure Captions

Figure 1. Graphical representation of the latent class model with a distal outcome. C refers to the latent class variable, $X_1, X_2, ...X_J$ refer to manifest indicators of C, and Z refers to the distal outcome.


C
Z
X1
X2
...
XJ