Running Head: Relating LCA Results

Relating Latent Class Analysis Results to Variables not Included in the Analysis

Shaunna L. Clark & Bengt Muthén
University of California, Los Angeles

The research of the first author was sponsored by the Center for Prevention and Early Intervention (P30 MH066247) and jointly funded by the National Institutes of Mental Health and Drug Abuse. The research of the second author was supported by grant K02 AA 00230 from National Institute on Alcohol Abuse and Alcoholism, by grant 1 R21 AA10948-01A1 from the NIAAA, by NIMH under grant No. MH40859, and by grant P30 MH066247 from the NIDA and the NIMH.

Abstract

An important interest in mixture modeling is the investigation of what types of individuals belong to each latent class by relating classes to covariates, concurrent outcomes and distal outcomes, also known as auxiliary variables. This article presents results from real data examples and simulations to show how various factors, such as the degree to which people are classified correctly into latent classes and sample size, can impact the estimates and standard errors of auxiliary variable effects and testing mean equality across classes. Based on the results of the examples and simulations, suggestions are made about how to select auxiliary variables for a latent class analysis.

Relating Latent Class Analysis Results to Variables not Included in the Analysis

Introduction

Mixture modeling in the form of latent class analysis and growth mixture modeling has become an important tool for researchers (for an overview see Muthén, 2008). Mixtures help model unobserved heterogeneity in a population by identifying different latent classes of individuals based on their observed response pattern. An important interest in mixture modeling is the investigation of what types of individuals belong to each class by relating classes to covariates, concurrent outcomes and distal outcomes, also known as auxiliary variables. This paper will compare techniques for relating latent classes to auxiliary variables.

As a first step in investigating the relationship between latent classes and auxiliary variables, many researchers utilize mean comparisons tests, such as *t*-tests, ANOVAs, or chi-square tests, to get an idea of whether or not a relationship is present. In order to conduct these tests, the first step is to estimate the mixture model based only on latent class indicators, obtaining each individual's most likely class membership, with assignment into classes being based on the highest probability of being in a given class. Using these assigned class memberships, the mean comparison tests can then be performed.

Furthermore, regression models are used to explore the relationship between latent classes and auxiliary variables. There are four commonly used regression approaches:

- Most likely class regression: Regression of most likely class membership on the covariates,
- Probability regression: Regression of an individual's logit-transformed posterior probability to be in a given class on the covariates,

- Probability-weighted regression: Regression that is weighted by an individual's posterior probability to be in a given class,
- Single-step regression: Including the covariates in the analysis while forming the latent classes.

In both the mean- and regression-oriented approaches, a problem with using most likely class membership is that it is treated as an exact, observed variable. The problem with treating class membership as exact can be easily illustrated. Suppose a 2-class model and take two individuals, one with a probability of 1.0 for belonging to Class 1 and 0.0 for Class 2 and the other with a probability of 0.51 for belonging to Class 1 and 0.49 for Class 2. Both individuals would be assigned and treated as members of Class 1 in the subsequent analyses. But the analyses does not take into account that the two individuals have different probabilities of being in the same class and instead are treated as if they both have a probability of 1.0 of being in Class 1. This will distort estimates because individuals are forced into their most likely latent classes. The standard errors will also be incorrect because the analysis does not take into account the uncertainty of the classification but treats it as an observed variable. This poses a problem because incorrect standard errors can lead to erroneous conclusions about the significance of an effect.

As in the most likely class regression, the first step of the probability and probability-weighted regressions is the estimation of the latent class model based only on the latent class indicators. In the second step, instead of having assigned class membership as the outcome, the probability regression uses an individual's logit- transformed posterior class probability as the outcome. For the probability weighted regression, the regression of class membership on the covariates is weighted by each individual's posterior probability. Using the probabilities of being in a given class may give less bias to regression coefficients but is still problematic because the probabilities are also estimates and an analysis will not take into account the error associated with those estimates. So, the standard errors of a regression between the posterior probabilities and an auxiliary variable will be incorrect.

In the single-step approach, the problem of incorrect estimates and standard errors is circumvented because the analysis allows individuals to be fractional members of all classes and the latent class variable is not treated as observed. However, such an approach may be cumbersome when many auxiliary variables are involved because of the increased computation time associated with the inclusion of more auxiliary variables. Furthermore, a researcher may not always want auxiliary variables to influence the determination of class membership because the inclusion of auxiliary variables can potentially change the substantive interpretation of the latent classes.

A fifth approach, which has recently been put forward, is pseudo-class draws (Asparouhouv & Muthén, 2007; Wang et al., 2005). Here, several random draws are made from each individual's posterior probability distribution to determine an individual's class membership. Based on these draws mean tests and regression estimates can be computed.

This paper will explore the quality of estimates and standard errors incurred when researchers use the five regression approaches introduced above. Additionally, this study will investigate how using most likely class membership in mean comparison testing can potentially distort the test statistic and its interpretation. Using Monte Carlo simulations, this study will examine how various factors, such as the degree to which people are classified correctly into latent classes, can impact the estimates and standard errors of auxiliary variables and testing mean equality across classes. Based on the results of the real data examples and simulations, suggestions will be made about how to select covariates for an analysis.

The first section of this paper introduces the latent class analysis model and describes the approaches for examining the relationship between the latent classes and auxiliary variables. The next section provides two real data examples to demonstrate the problem of treating class membership as an observed variable and also to show how incorrect the estimates and standard errors can be when including many auxiliary variables. The third section describes the simulation study and its results to confirm the results of the real data examples as well as to show the extent of the problem. The final section, presents highlighted results, suggests under what conditions it is appropriate to use the methods examined, as well as suggesting a process by which to select auxiliary variables for an analysis.

Background

Latent Class Analysis Model. The latent class analysis (LCA) model, introduced by Lazarfeld and Henry (1968), is used to identify subgroups, or classes, of a study population. A diagram of an example of a latent class analysis model is shown in Figure 1a. There are two major concepts depicted in Figure 1a, the latent class itself and the observed outcomes or items that define the class. These can be seen in Figure 1 as the c, and u_1-u_r, respectively. The boxes, u_1 to u_r, represent the observed response items or outcomes. The outcomes in an LCA model can be categorical or continuous, though this paper will specifically focus on dichotomous, categorical items. The circle with the letter c in the middle is the unordered, categorical latent class variable with K classes. The arrows pointing from the latent class variable to the boxes above indicate that those items are measuring the latent class variable. This means that class membership is based on the observed response pattern of items. An important assumption, called the conditional or local independence assumption, implies that the correlation among the observed outcomes is explained by the latent class variable, c. Because of this, there is no residual correlation between the items.

For an LCA model with categorical outcomes, there are two types of model parameters: conditional item probabilities and class probabilities. The conditional item probabilities are specific to a given class and provide information about the probability that an individual in that class will endorse that item. The class probabilities specify the relative size of each class, or the proportion of the population that is in a particular class.

The LCA model with *r* observed binary items, *u*, has a categorical latent variable *c* with K classes ($c = k$; $k = 1, 2, . . ., \mathrm{K}$). The marginal item probability for item $u_{\mathrm{j}} = 1$ is

$$P(u_j = 1) = \sum_{k=1}^{K} P(c = k) P(u_j = 1 \mid c = k).$$

Assuming conditional independence, the joint probability of all the *r* observed items is

$$P(u_1, u_2, ..., u_r) = \sum_{k=1}^{K} P(c = k) P(u_1 \mid c = k) P(u_2 \mid c = k) ... P(u_r \mid c = k).$$

A product of LCA is the estimated class probabilities for each individual, called posterior probabilities, analogous to factor scores in a factor analysis (Muthén 2001). These are estimates of

$$P(c = k \mid u_1, u_2, ..., u_r) = \frac{P(c = k) P(u_1 \mid c = k) P(u_2 \mid c = k) ... P(u_r \mid c = k)}{P(u_1, u_2, ..., u_r)}.$$

Note that each individual is allowed fractional class membership and may have non-zero values for many classes. It is from these probabilities that class membership is assigned. An individual is assigned to be a member of a class based on their highest probability of being in a given class, even though an individual may have several classes to which they are a partial member. Once assigned to a class, an individual is assumed to be a part of that class 100%. Therefore, the fractional membership in multiple classes disappears when using most likely class membership as an observed variable in subsequent analyses.

One way to examine how well individuals have been classified is to look at the entropy of the latent class model. One way to compute entropy is to use the following formula:

$$E_k = 1 - \frac{\sum_i \sum_k (-\hat{p}_{ik} \ln \hat{p}_{ik})}{n \ln K},$$

where $\hat{p}_{ik}$ denotes the estimated posterior probability for individual *i* in class *k*. Entropy is measured on a zero to one scale with a value of one indicating the individuals are perfectly classified into latent classes. Higher values of entropy indicate better classification of individuals.

Mean Comparison Testing. One way to investigate the relationship between the latent classes and auxiliary variables is to conduct a mean difference test across the classes. This is done by first estimating the latent class model and then classifying people into latent classes based on their highest posterior probability of being in a given class. These memberships are then used as the groups in the mean comparisons tests, though which test to use depends on the type of covariate.

If the covariate is continuous, a one-way analysis of variance (ANOVA) is appropriate because the test requires a continuous outcome and can handle having two or more latent classes. When there are only two class means to compare, one could also choose to use a *t*-test but, with only two classes, the *t*-test and the ANOVA are equivalent. One assumption of ANOVA is that being a member of one class has no effect on whether an individual is a member of a different class. This assumption is incorrect when using latent classes as the group in an ANOVA because if an individual is not a member of class one, then they have to be a member of class two. Even though the analyses do not meet the assumptions of this test, it is still commonly used by researchers when comparing the mean of an outcome across latent classes. So, it is important to investigate the impact of using this test.

If the covariate is categorical, a chi-square test of equal proportions will test whether the frequency of the covariate is same across all the class. The chi-square test can be computed as:

$$\mathrm{Q_P} = \sum_{j=1}^{K} \frac{(f_j - e_j)^2}{e_j},$$

where f_j denotes the frequency of the covariate in class *j* (or the number of observations in class *j*) for $j = 1, 2, \ldots, K$. Since the null hypothesis specifies equal proportions of the total sample size for each class, the expected frequency for each class equals the total sample size divided by the number of classes, or:

$$e_j = n / K \text{ for } j = 1, 2, \ldots, K.$$

Most Likely Class Regression. In this approach, most likely class membership is related to the auxiliary variables. First, the LCA is conducted based on the latent class indicators and individuals are assigned to their most likely class. In the second step, the membership assignments from the previous step are used as an observed variable in the regression. In the covariate case, the assigned membership, *m*, which can take on values from 1 up to K, is regressed onto the covariate using a multinomial logistic regression where :

$$\mathrm{P}(m_\mathrm{i} = k \mid x_\mathrm{i}) = \frac{\mathrm{e}^{\alpha_k + \beta_k x_\mathrm{i}}}{\sum_{s=1}^{\mathrm{K}} \mathrm{e}^{\alpha_s + \beta_s x_\mathrm{i}}},$$

where $\alpha_K = 0$, $\beta_K = 0$ so that $e^{\alpha_K + \beta_K x_i} = 1$, implying that the log odds of comparing Class *k* to the last Class *K* is

$$\log\,[\mathrm{P}(m_{\mathrm{i}} = k \mid x_{\mathrm{i}}) \,/\, \mathrm{P}(m_{\mathrm{i}} = K \mid x_{\mathrm{i}}\,)] = \alpha_k + \beta_k\, x_i\,.$$

In the case of a distal outcome, the distal outcome is regressed on the assigned membership and the type of regression used will depend on the type of variable the distal outcome is.

Probability regression and probability- weighted regression. One way around using the assigned class membership, is to use class probabilities. Two ways of incorporating class probabilities will be discussed in this study. The first method is the probability regression in which an individual's posterior class probability is converted to the logistic scale and then regressed on the covariates using a linear regression. The probabilities are converted because on a probability scale the values can only range from 0 to 1, which is not suitable for linear regression. On a logistic scale, the converted probabilities can have any value. In the second approach, the regression is weighted by each individual's posterior probability of being in a given class.

Using the probabilities of being in a given class may give less bias to regression coefficients but is still problematic because the probabilities are also estimates and an analysis will not take into account the error associated with those estimates. So, the standard errors of a regression between the posterior probabilities and an auxiliary variable will be incorrect. This poses a problem because incorrect standard errors can lead to erroneous conclusions about the significance of an effect.

Single Step Approach. In order to see who is a member of each class, covariates can be added to the LCA model. Figure 1b shows the diagram of this model. As with the plain LCA model, there is a latent class variable, *c*, and the items that measure the latent class variable. What has been added is the box to the left of the latent class variable, with the *x* in the middle, which represents the covariate. There is an arrow starting from the covariate box and ending at the latent class variable. This indicates that the latent class variable is being regressed on the covariate *x.* More specifically, this regression is a multinomial logistic regression because the outcome, the latent class variable *c*, is categorical with potentially more than two categories.

LCA with covariates has been considered by Bandeen-Roche et. al. (1997), Dayton and Macready (1988), Formann (1992), and Heijden et. al. (1996). This modeling considers a covariate *x*, where the probability that individual *i*, falls in class *k* of the latent class variable *c* is expressed through multinomial logistic regression as

$$\mathrm{P}(\mathrm{c}_{\mathrm{i}} = k \,|\, x_{\mathrm{i}}) = \frac{\mathrm{e}^{\alpha_k + \gamma_{\mathrm{k}} x_{\mathrm{i}}}}{\sum_{s=1}^{\mathrm{K}} \mathrm{e}^{\alpha_{\mathrm{s}} + \gamma_{\mathrm{s}} x_{\mathrm{i}}}},$$

where $\alpha_K = 0$, $\gamma_K = 0$ so that $e^{\alpha_K + \gamma_K x_i} = 1$, implying that the log odds of comparing class *k* to the last class *K* is

$$\log\,[\mathrm{P}(c_{\mathrm{i}} = k \,|\, x_{\mathrm{i}}) \,/\, \mathrm{P}(c_{\mathrm{i}} = K \,|\, x_{\mathrm{i}}\,)] = \alpha_k + \gamma_k\, x_{\mathrm{i}}\,.$$

Muthén and Muthén (2000) gave an example of LCA with covariates applied to antisocial behavior classes related to age, gender, and ethnicity.

Similar to covariates, distal outcomes can be added to the LCA model to see how class membership predicts the distal outcome. Here, the latent class variable is an exogenous instead of endogenous variable. This is one way that substantive researchers use to investigate the predictive validity of the latent classes. Figure 1c shows an LCA model with a distal outcome. Again, there is the traditional latent class model, but the addition is the box to the right of the latent class variable with a *y* in the middle. This box represents the distal outcome. There is also an arrow pointing from the latent class variable to the distal outcome box indicating that the distal outcome is being regressed on the latent class variable. This regression can be linear, logistic, or another type of regression depending on the form of the distal outcome.

Pseudo-class draws. Instead of using assigned class membership, another option is to use pseudo-class draws. When doing latent class analysis, every individual has a posterior class distribution or, stated differently, every individual has a posterior probability of being in each class. The distribution of these probabilities is multinomial because there are potentially more than two classes. The draws are made by taking random samples from this multinomial distribution. By having multiple random samples, individuals are given a chance to flip into neighboring classes, which gives a sense of the variation associated with the distribution. The pseudo-class draws are similar to multiple imputation in missing data analysis, except in this case, the latent classes are what is missing. Given the pseudo-class draws, class specific means, variances, mean equality tests and regressions can be computed. For a more technical treatment of pseudo-class draws, see Asparohouv & Muthén (2007) and Wang et al. (2005).

Previous Work. While many researchers have identified the problem of using class membership as an observed variable in an analysis, none has shown how problematic it can be. Hagenaars (1993) points out that because there can be high misclassification of individuals into classes, using most likely class assignment can be problematic because those individuals that are assigned to the wrong class will skew the true relationship between the classes and the external variables. Clogg (1995) also points out the dangers of using this strategy and gave it a name, the *classify-analyze* strategy. Nagin and Tremblay (2001) show how after individuals are assigned to classes that comparisons can be made across the classes on an outcome of interest by inspecting the means of the outcome. They continue on to point out that this technique is problematic because there is no valid basis for computing the standard errors and therefore, confidence intervals or hypothesis tests cannot be computed (Roeder et. al. 1999). Heijden et al. (1996) offer that one way to avoid the problems of the *classify-analyze* strategy is to estimate class membership and the relationship to external variables all in one step. Heijden also points out that there are three advantages to conducting the analyses in one step. First, is that by doing the analyses in one step the classification issue is avoided. Second, it is better to work with one model and its model fit than to worry about model fit for two separate models, the latent class analysis and the subsequent regression. Finally, it is possible to investigate models that have zero degrees of freedom or are unidentified in ordinary latent class analysis by using covariates in the analysis.

This study shows how much of a problem it is to use class membership as an observed variable by demonstrating how distorted the estimates and standard errors of a covariate or distal outcome effect can be. It also shows results for a new approach for relating latent classes to auxiliary variables, namely using pseudo-class draws.

Real Data Example

In this section two real data examples will be discussed to explore how the problem of treating assigned class membership as an observed variable is compounded when there are multiple auxiliary variables in an analysis. The first example examines antisocial behavior in 16-23 year olds. The second example investigates aggressive behavior of first grade children in the classroom.

Antisocial Behavior Example. The data for this example are the Antisocial Behavior (ASB) data which were taken from the National Longitudinal Survey of Youth (NLSY) that is sponsored by the Bureau of Labor Statistics. These data are made available to the public by

Ohio State University. The data were obtained as a multistage probability sample with an over-sampling of African Americans, Hispanics, and economically disadvantaged non-black and non-Hispanics.

The ASB data include 17 antisocial behavior items that were collected in 1980 when respondents were between the ages of 16 and 23. The ASB items assessed the frequency of various behaviors during the past year. A sample of 7,326 respondents has complete data on the antisocial behavior items and the covariates used in this example. For covariates, gender, age and ethnicity will be considered. Gender and ethnicity are dichotomous variables with ethnicity being split into two separate variables, Black and Hispanic, which are referenced to the predominant ethic group in the sample, White.

The data were first analyzed by conducting the LCA and inclusion of the covariates in one single step. The next phase was to conduct the other strategies: regressing most likely class membership on the covariates, regressing the class probabilities on the covariates, using the class probabilities as weights in a regression, or pseudo-class regression. Also, once class membership was established, the mean comparison tests were conducted. The results of these analyses can be seen in Table 1 for the mean comparisons and Table 2 for the regression. In Table 2 there are four classes in the analysis which means that when the latent classes are regressed on the covariates a multinomial logistic regression is used. Because of this, the results are in terms of logits and the regressions are in reference to the last class which has been set to be the largest class.

Before discussing the results, it is important to understand the class structure of this dataset so it can be understood how using different methods of incorporating covariates can affect the overall interpretation. Figure 2 shows a class profile plot for these data with four classes. The first class, and the smallest class in size, is considered to be a high antisocial behavior class because the probability of endorsing almost all of the items is high. The second class can be thought of as the person offense class because the items that have a high probability of endorsement are items related to aggression against another person such as fighting. The third class can be thought of as the drug class because only items relating to drugs are highly endorsed. The four class and largest in size can be thought of as the normative class because almost all items have a low probability of endorsement, with marijuana use being the only exception. The entropy is 0.74.

The results in Table 1 suggest that the pseudo-class Wald chi-square and equal proportion chi-square yield similar results. The absolute value of the chi-square statistics are similar across both types of tests, with the equal proportions chi-square test yielding higher values. Additionally, the significance of the tests is equivalent across all the covariates, though this may be due to chi-square being sensitive to the large sample size of this data set. Even though the values and significance of the tests are similar across the two methods, the potential for problems becomes apparent. Differences between the tests are likely to arise when the tests are close to the borderline of being significant. When this occurs it is likely that the equal proportions chi-square will lead one to conclude that there are significant differences across the classes because it has higher test values, and the pseudo-class Wald chi-square will lead one to conclude that there are no differences across the classes

Table 2 shows that how covariates are incorporated into an analysis can affect the interpretation of the relationship between the covariate and the latent class variable. Using the assigned class membership, the class probability-weighted regression or pseudo-class draws yields smaller regression estimates and standard errors than conducting the analyses all in one step. Psuedo-class regression estimates are the furthest away from the estimates obtained when including the covariates in the analysis, followed by probability-weighted regression. Most likely class membership estimates and standard errors come closest to the results from including the covariates in the analysis.

In Table 2, the results for the probability regression are presented on a continuous dependent variable scale while the rest of the table is presented on a logistic scale. Because the results in the table are not directly comparable, the comparison between the methods will focus on the sign and significance of the regression estimates. In most of the regressions, the sign and significance of the regression estimate are the same across all methods. For the regressions of the Drug class on gender and the Person offense class on Hispanic, the sign of the probability regression estimate differs from the other methods. Additionally, for the Person offense class on Hispanic regression, the estimate is significant for the probability regression, but it is insignificant for the other methods.

One interesting point of difference between the five approaches occurs for the case of the regressions of the High class on Black and Hispanic. In these two cases using most likely class membership, probability regression, probability-weighted regression, or pseudo-class regression would lead to the conclusion that the estimate is significant since the ratio of the estimate to the standard error is greater than 1.96. But, including covariates while forming the latent classes leads one to conclude that the estimate is insignificant because the ratio of the estimate to the standard error is less than 1.96. These two cases highlight the problem associated with underestimated standard errors. Because the top four methods in the table have lower standard errors, they are more significant than the last method because in the estimate to standard error ratio they have a smaller denominator. This also suggests that the standard error difference among the methods is larger than the difference in the estimates.

This example has shown that using assigned class membership regression, the class probability-weighted regression, or pseudo-class draw regression underestimates the regression effects of the classes on the covariates. The probability regression sometimes showed different sign and significance of the regression estimates when compared with the other method. The mean comparisons had different values for each test, but the overall significance was the same which may be in part do to chi-square's sensitivity to large sample sizes.

Aggression Example. The data for this example come from a randomized preventive field trial conducted in Baltimore public schools (Dolan et. al. 1993, Ialongo et al. 1999). The study applied a universal intervention aimed at reducing aggressive-disruptive behavior during first and second grade to improve reading and reduce aggression with outcomes assessed through middle school and beyond (Kellam et al. 1994). The outcome variables of interest are teacher ratings of each child's aggressive behavior in classroom. The ratings were made using the Teacher's Observation of Classroom Adaptation Revised (TOCA-R) scaling instrument, with 10 items. For simplicity, the items were dichotomized with a value of one representing presence of the aggressive behavior and zero representing no symptom presence. Information was also collected on concurrent and distal outcomes, including school removal and juvenile court records.

This example will focus on analyzing the pre-intervention data of Cohort 1 when the children entered into the intervention study at the beginning of first grade. In total, there were 1,174 children participating in this study. For covariates, gender, ethnicity, and treatment condition will be considered. All three covariates are dichotomous with gender indicating whether an individual is male, ethnicity indicating whether a child is African American, as opposed to Caucasian, and the treatment variable indicating whether a child's classroom was assigned to the treatment or control condition. Since this example focuses on a pre-intervention time point and classrooms were assigned to treatment or condition randomly, the relationship between the latent class variable and the treatment condition indicator should be non-significant.

This analysis was executed similarly to the ASB example. The data was first analyzed by conducting the LCA and inclusion of the covariates in one single step. The next phase was to conduct the other strategies: regressing most likely class membership on the covariates, regressing the logistic class probabilities on the covariates, conducting a class probability-weighted regression, or pseudo-class regression. Also, once class membership was established, the mean comparison tests were conducted. As in the previous example, the probability regression results are on a different scale and comparison of this method to the others will focus on the sign and significance of the regression estimate. The results of these analyses can be seen in Table 3 for the mean comparisons and Table 4 for the regression results.

A three class solution was found to fit these data. The profiles for the TOCA data can be found in Figure 3. The first class has a high average probability of endorsing all of the items, so this class is named the high aggression class. The second class has moderate endorsement for those items dealing with verbal aggressions and low endorsement for the other items so this class is called the verbal aggression class. The final class has low endorsement of all items and comprised the largest percentage of the sample and so was named the normative class. The entropy is 0.79.

Comparing Table 3 to the results seen in the ASB example, a similar pattern emerges. When comparing the pseudo-class chi-square results to the equal proportions chi-square, the values of the statistics are similar in size with the equal proportions chi-square having a higher value. For the gender and race covariates, the significance of the test is equivalent across the two statistics and both tests would lead one to conclude that the effects of gender and race are significantly different across the classes. For the treatment indicator, both tests indicate that there are no significant differences across the classes because the p-value associated with both statistics is greater than 0.05.

The regression results presented in Table 4 show a similar pattern to what was seen in the ASB data example. A comparison of the most likely class membership regression, probability-weighted regression, and pseudo-class regression shows that all of the methods yield similar estimates and the standard errors for the regression between the classes and the covariate. But, when compared with including covariates during the formation of the latent classes, the other three methods tend to produces slightly smaller estimates and standard errors. In all cases, except for the regression of the Verbal Aggression class on Black, the significance of the effect is the same across the methods. For the regression of the Verbal Aggression class on Black, the pseudo-class regression is not significant, but relatively close to being significant, while the other three methods are significant.

When comparing the sign and significance of the probability regression estimate to the other methods, Table 7 shows that there are differences. For the regressions of the Verbal class on gender and Black, the estimate from the probability regression is not significant, but the estimate is significant for the other methods. The regression of High class on treatment condition shows the opposite pattern, where the estimate is significant for the probability regression but not for the other methods. Any regression with treatment condition as a covariate was not expected to be significant because the data being analyzed are a pre-intervention time point. A significant effect of any class on treatment condition would suggest that randomization did not work. For the regression Verbal class on gender, the probability estimate has a negative sign while the other methods have a positive one.

Comparing the two different data examples, there are consistencies in the performances of the methods investigated. For the mean comparisons, in both data sets, the statistics yielded similar values and led to similar conclusions about the significance of the covariate of interest. For the regression results, in the TOCA and ASB data, most likely class membership regression, probability-weighted regression, and pseudo-class regression performed similarly, but in comparison to including covariates while forming the latent classes the other methods tended to underestimate the effect and the standard errors. The probability regressions showed a tendency to have different signs and significance of the regression coefficients when compared to the other methods. A question to ask is under what conditions will these differences that are seen in the regression results arise? Potential explanations for the differences between the methods are how well individuals are being classified into their latent classes, sample size, and the size of the covariate effect. One way to investigate this question is to do a Monte Carlo simulation study where the settings of the analysis, such as sample size, entropy, and the size of the covariate effect, can be controlled to fully explore potential reasons for differences in the methods.

Monte Carlo Simulation Study

One of the main questions of this study is how to select important covariates for an analysis. This will be investigated by comparing the effectiveness of the mean comparison and regression methods described above. One way to conduct these investigations is to do a Monte Carlo simulation study where the covariate effect, classification of people into latent classes, and model population parameters are known and how to treat latent class membership is explored under a variety of different conditions.

In this study, the sample size was chosen to be 250 and 1,000. The number of replications of data generation was chosen to be 500 for all the models since this assures sufficient reliability of the summary information.

For the model population, or the model from which the data is generated, there are two possible sources of variation that will be examined in this paper: 1) degree to which individuals are correctly classified into latent classes, and 2) the strength of the relationship between the latent classes and the covariate. These will be discussed in detail below.

Data Generation Models. The overall plan of the simulation was to generate the data using a latent class model with a covariate effect. Specifically, data was generated using the model shown in Figure 1b. Ten items were used in the analyses. All of the items in the analysis were chosen to be binary so that the conditional item probabilities, the probability of endorsing an item given membership in a specific class, could be used to distinguish among the classes. Also, binary items are common when studying behavioral disorders because they can indicate either presence or absence of a symptom.

When generating latent class models, it is important to specify the number of classes in each model and what the class profiles look like. The number of classes for each of the latent class variables was chosen to be two for simplicity. Another key feature of latent class models is the class profile or item endorsement profile. One way to visually inspect the class profile is to plot the items on the x-axis and the conditional item probabilities are on the y-axis. In this study, a complex profile was used in which the two class profile cross. In Figure 1d, which shows an example of a complex profile, class one has high endorsement of the first five items and low endorsement of the last five. Class two has the opposite pattern of class one, with low endorsement of the first five items and high endorsement of the last five items.

One hypothesis is that as classification of people into latent classes becomes worse (as entropy becomes close to zero) treating the latent class variable as observed further distorts the estimate of the covariate or distal effect and the standard error of that parameter. In order to explore this hypothesis, different entropy settings were used in this study: perfect, high, medium and low entropy. For the perfect entropy case, the average entropy value was 1.0 meaning that people were perfectly classified into latent classes. The high entropy had an average value of 0.80 across the replications, which says that 80% of the time individuals were correctly classified in latent classes. The medium entropy case had an average value of 0.60 and the low entropy case had an average value of 0.40.

Also, the strength of the covariate or distal outcome effect was studied using two different settings. In Figure 1b, the covariate effect can be seen as the path emanating from the covariate *x* and ending at the latent class variable *c*. The first setting was to set this path to 0.5 indicating that there is a strong, positive relationship. The final setting assumes that there is no relationship at all and so the path was set to zero.

Overall, there were eight data generation models specified. There were four different entropy settings (perfect, almost perfect, high and low) and two different covariate effects giving eight data generation models.

Data Analysis Models. The data was generated using the model specifications that were described above. The data were then analyzed using each of the five different regression approaches described and once most likely class membership information was obtained, mean comparison tests were conducted. The simulation and analysis of the data was carried out using the Monte Carlo facilities in Mplus version 5.0 (Muthén & Muthén, 2007). Mplus input scripts for each method are presented in the Appendix for the case where the sample size is 1,000 and the entropy is 0.8.

Two techniques were used to evaluate the mean differences across classes on the outcome of interest. The first uses pseudo-class draws to conduct a pseudo-class chi-square test. The other technique is a *t*-test which was conducted by classifying individuals into their most likely latent class and then comparing mean differences in each class using an independent samples *t*-test. In order to be able to compare the results of the pseudo-class chi-square test and the most likely class membership *t*-test, the value of the *t*-test was squared in Table 5 in order to make the *t* value asymptotically equivalent to χ^2. In the real data examples, a chi-square test of equal proportions was compared to the pseudo-class chi-square test instead of a *t*-test because in both real data examples there were more than two classes and the covariates were categorical. The presence of only two classes and a continuous covariate in the simulation required that a *t*-test be used.

As mentioned previously, five different regression approaches were used to analyze the data and obtain estimates and standard errors of the covariate or distal outcome effect: most likely class membership regression, probability regression, probability-weighted regression, pseudo-class regression, and single step regression.

Simulation Results

This section summarizes results from the simulation study. First the results of the mean comparison tests will be examined. Next, the comparison of the five approaches used to incorporate covariates into a latent class analysis will be conducted.

Mean Comparison Simulation Results. In some instances, it is not possible to include the covariates when conducting the latent class analyses. When this is the case, one way to examine differences in covariates is to compare the means across classes. As mentioned previously, two techniques were used to investigate which is the best method for testing mean equality: pseudo-class chi-square test and an most likely class membership *t*-test. The results for these comparisons are presented in Table 5.

In Table 5, the results of the means comparisons are broken down by the size of the relationship between the latent classes and the covariate. The top half of the table presents results for when there is no relationship between the latent classes and the covariate. Because there is no relationship, it is expected that across all settings the tests will not be significantly different from zero. The p-value of the average statistic, across all settings, is greater that 0.05 which indicates that the test is not significant, which is what was expected. The bottom two rows of this case give an idea of the type I error, or when a difference is observed when in truth there is no difference, associated with these tests because these rows give the proportion of replications that have a significant p-value. Given that the significance level has been set to 5%, it is expected that if the test is working properly, 5% of the time the tests will find a significant difference when there is none. In the bottom row of the table, the pseudo-class chi-square has the proportion of replications giving significant p-values close to zero for nearly all the settings examined. This indicates that the test is not performing as it should because a 5% error rate is not recovered. For the most likely class membership *t*-statistic, the proportion has values from 0.036 up to 0.052, which is close to the expected value of 0.05. This indicates that the test is performing close to what it should be.

The bottom half of the table is the largest relationship with the class on covariate effect being 0.5. This setting shows similar results to what was seen in the real data examples with the average value of the statistics being close in size, but with the most likely class membership *t*-value average being larger than the pseudo-class chi-square. The difference between the averages of the two statistics increases in size as both the entropy and the sample size increases. The p-value associated with the average statistic is slightly smaller for the most likely class membership *t*-statistic than the pseudo-class chi-square. More importantly, however, is that the decision of whether or not the covariate is significantly different across the classes does not change across the two methods. Both p-values lead to the conclusion that the covariate is significantly different across the classes. Even though the average p-values for both statistics indicate that the covariate is significantly different across the classes, comparing the number of replications which have significant p-values (p < 0.5) does show a difference. For the smaller sample sizes, the most likely class membership *t*-statistic produces a larger number of significant replications. When the sample size is equal to 1,000 the number of replications with significant p-values increases when compared with the smaller sample size of 250. Additionally, as the entropy increases, the number of significant replication approaches 500 or 100% of the replications showing significant p-values. One explanation for these results in the higher sample size is that the pseudo-class chi-square and the most likely class membership *t*-statistic, which, when squared, is asymptotically equivalent to a chi-square test, are picking up on a chi-square test's sensitivity to large sample sizes and will always conclude that covariate is significantly different across the classes.

The simulation study has shown that the pseudo-class chi-square and the most likely class membership *t*-value yield similar results for the magnitude of their values and the significance of the covariate effect. This is similar to what was seen in the real data examples.

Regression Simulation Results. The comparison of the approaches examined in this paper was executed by simulating data under a variety of conditions and then analyzing the simulated data using all five of the regression approaches: most likely class membership regression, probability regression, probability-weighted regression, pseudo-class regression, and single step regression. When examining the results it is important to examine the power and mean square error (MSE), in addition to the estimates of the coefficient. The power, or the probability of rejecting the null hypothesis when it is false, provides an estimate of whether there is enough information in the data to detect a covariate effect. The MSE quantifies the amount by which an estimate differs from the true value of the regression coefficient. The results of these analyses are displayed in Tables 6 and 7.

Table 6 displays the results for when the relationship between the latent classes and the covariates is specified to be 0, or no relationship. The table rows are broken down into five sections, one for each of the methods used to incorporate covariates. The top of the table shows results for when most likely class membership is regressed on the covariates. Following that, there are sections for when the class probabilities are regressed on the covariates, when posterior class probabilities are used as weights in the regression of class membership on the covariates, pseudo-class regression, and when covariates are included while the latent classes are formed. The single step regression was placed at the bottom of the table for easier comparisons since this was the model from which the data was generated and the other models will be compared to.

Moving from left to right in the table, the entropy, or the degree to which individuals are correctly classified into their latent classes, increases.

Starting with the two rightmost columns of the table, where individuals are perfectly classified, a comparison of the five methods shows that for this entropy setting the methods are equivalent. When there is perfect entropy, every individual has a probability of 1.0 of being in their latent class, which means that there is no error involved when classifications are made into most likely latent class. With the pseudo-class draws, all of the draws are equivalent to one another since there would be no class switching because individuals are perfectly classified.

Comparing the estimates obtained by the five methods, all of them are able to recover the true effect of zero, but when comparing the standard errors of the methods a similar pattern to what was seen in the real data examples emerges. When comparing the standard errors to including the covariates in the model, the four other methods underestimate the standard errors. Standard errors associated with using class probabilities as weights are the smallest overall.

One way to further compare among the methods is to examine the mean square error (MSE), which quantifies the amount by which an estimate differs from the true value of the regression coefficient. An MSE value of zero indicates that there is no difference between the estimated regression coefficient and its true value. Thus, when comparing the MSE of two models, the model with the smallest MSE is interpreted as being the best model for explaining the variability in the observations. In Table 6, in all settings, the MSE of each of the techniques is close to zero. When comparing among the techniques, for the two smallest entropies, pseudo-class regression and probability-weighted regression have the smallest MSEs. When the entropy increases to 0.80, probability regression has the smallest MSEs. But, the difference between the probability regression MSEs and the next smallest MSEs, from pseudo-class regression, is only 0.005.

One way of examining how well the parameters and their standard errors are being estimated in a simulation study is to look at the coverage. Coverage is defined as the proportion of replications for which the 95% confidence interval contains the population parameter value (Muthén & Muthén, 2008). The closer the coverage is to 1 the better the estimates of the parameters and their standard errors. Coverage is high across the methods and in all settings, with the lowest value in the entire table being .938. Probability-weighted regression has slightly smaller coverage values than the other methods, but they are still high.

In Table 6, the last row in each section of the table examines whether the correct alpha level is recovered. Because the population value for this table is equal to zero, the values in this row are estimates of Type I error, or the probability of rejecting the null hypothesis when it is false. If a method is performing correctly, then a value of 0.05 is expected based on the 5% significance level that was chosen for this test. For the most likely class membership, probability regression, class probability-weighted regression and including the covariates in the model, the values in these cells are all close to 0.05 which shows that these methods are performing as expected. The pseudo-class regression cells have values close to zero across all settings indicating that the method is not performing as expected.

In Table 7, where the relationship between the classes and covariates is 0.5, the equivalency of the methods that was seen when there is no relationship breaks down as the entropy decreases. The most likely class membership, probability, probability-weighted, and pseudo-class draw estimates are all smaller than the true value of 0.5, with the pseudo-class regression and probability-weighted regression having the smallest values across all settings. As the entropy settings decrease, the estimates for all four methods get further away from the true value. Including the covariates in the models tends to slightly overestimate the true value across the settings with the biggest difference occurring for the small sample size and lowest entropy setting.

While not as big of a problem as having incorrect parameter estimates, there are also differences in the standard errors that each method is obtaining. As the entropy decreases, the standard errors for the most likely class membership, probability, class probability weighted regression, and pseudo-class regression are underestimated when compared with the ones obtained by including the covariates while forming the latent classes. Having underestimated standard errors is problematic because it could lead a researcher to conclude that an effect is significant when it may not be. The standard errors for the pseudo-class regression, while not exactly the same as the ones obtained when including the covariates during the formation of the latent classes, are relatively close but still underestimated while the standard errors for the probability-weighted regression are the most underestimated.

In Table 7, the MSEs of the each of the methods is still relatively close to zero. When comparing among the methods, for the highest entropy settings, the methods has similar MSEs with probability regression having the highest MSEs and including covariates in the model having the smallest. For the lower entropy settings, including covariates in the models tended to have the smallest MSEs.

For the two highest entropy settings, the coverage is quite good with most methods and settings having coverage of at least 0.80, except in the probability case where the coverage was 0.50 or lower in the high entropy setting. As the entropy decreases, the coverage stays around 0.95 for the case where covariates are included while forming the latent classes, which indicates that the parameters and standard errors are being estimated well for this method. For the two lower entropy settings, the methods using class probabilities have the lowest coverage of all the methods. The pseudo-class regression and most likely class membership have similar coverage across the settings, but it is still low in the two lowest entropy settings.

The last row in each section of Table 7, displays the proportion of the replications for which the null hypothesis that a parameter is equal to zero is rejected at the 0.05 level (two-tailed test with a critical value of 1.96). For parameters with population values different from zero, this value is an estimate of power with respect to a single parameter, that is, the probability of rejecting the null hypothesis when it is false. A proportion of 0.80 or greater is considered to be good power to reject that the parameter is zero. For the two highest entropy settings, all of the methods have a high proportion of replications for which the null hypothesis is rejected, except for pseudo-class regression with a sample size of 250 and an entropy setting of 0.80. This exception hints at what is seen in the rest of table, which is that pseudo-class regression has lower power across most of the settings when compared with the other four methods. This is especially noticeable in the smaller sample sizes and lower entropies in first few columns. But, as the entropies and sample sizes become smaller the coverage does decrease across all methods. In the first column, with the small sample size and small entropy, none of the methods have a high power to reject the null hypothesis. Across all settings, probability-weighted regression has the highest power.

The simulation study shed light on the hypotheses. For the mean comparison simulations, the results show that the squared *t*-test produced higher test statistic values than the pseudo-class Wald chi-square test but, the significance of the tests was equivalent. For the regression simulation approaches, the results show that for approaches that do not include covariates during the formation of the latent classes, the estimates and standard errors are different when compared to including the covariates. Further comparisons among the regression methods and recommendations for which method to use in practice will be made in the discussion section below.

Discussion and Conclusion

The purpose of this study was to investigate how the different methods for treating latent class variables can impact the relationship between the latent classes and auxiliary variables. This was investigated by exploring two different real data examples that showed similar results when latent class membership is used as the dependent variable in a regression analysis. This issue was further investigated by utilizing Monte Carlo simulation techniques to explore how changing different settings in an analysis and varying the technique by which latent classes are incorporated can impact the estimates and standard errors of a regression analysis between latent classes and auxiliary variables.

Given the results of the real data examples and the simulation study, the question arises of which method or methods are best to use to incorporate auxiliary variables in an analysis? Table 6 shows that when there is no relationship between the covariates, all five of the regression approaches examined are able to recover the true effect of zero and that the effect is not significant. When there is no effect, it does not matter which approach is used, but most researchers will not know ahead of time whether or not a covariate will be significant. When there is a relationship between the covariate and latent class the question posed becomes more difficult to answer. Across all of the settings, including the covariate while forming the latent classes performed the best. The method was able to recover the true effect and had high coverage and power in most settings. But this method could be problematic to use in many real data analyses because the inclusion of many covariates will significantly increase computation time and the covariates may potentially influence the formation and interpretation of the latent classes.

If including the covariates while forming the latent classes is not an option because of the reasons stated above, one alternative is to use the most likely class membership, but only if the entropy is 0.80 or greater. When the entropy was high, most likely class membership was the best performing method in terms of recovering the true value used in the simulation study, and had relatively good coverage and power in the settings examined. But, when using most likely class membership, researchers need to be aware that this method does have the potential for underestimating the standard errors of the parameter. Therefore, when deciding on the significance of auxiliary variables, a more stringent criterion than the 5% level for deciding on significance should be employed. If the entropy is lower than 0.60, it is unclear as to which method is best to use. None of the methods were able to recover the true effect and all had various problems with underestimated standard errors, coverage and power.

One issue that arises with latent class analysis, and other mixture models, is how to decide on which covariates should be included in an analysis when there are a large number of covariates to select from. Based on the results of this study, a strategy for how to select covariates is suggested below. The first step is to conduct the latent class analysis without any covariates in the model in order to understand the substantive interpretation of the latent classes. In the second step, a pseudo-class Wald chi-square test is conducted to examine whether a covariate has some impact on the latent classes in terms of the mean differences. Next, a pseudo-class regression should be conducted to further winnow down the pool of potential covariates. A pseudo-class regression was chosen for this step since this method had the highest power to detect covariate effects of all the methods examined. Since the pseudo-class regression had biased estimates, as a final step, an analysis with the covariates included while forming the latent classes should be conducted in order to obtain unbiased estimates of the regression coefficient.

A major limitation of all simulation studies is the failure of any study to cover all possibilities that are seen in real data analysis. In the present study, the simulations only considered the impact of one continuous covariate, but in many real data analyses, many covariates, of possibly different types, would be included. The simulation study, however, was not meant to cover all possible situations that researchers may encounter, but instead to point out that researchers need to be conscientious of the magnitude of the problem associated when using these different methods. Future work in this area should investigate how the addition of multiple covariates and different types and combinations of covariates impact the estimates and standard errors obtained by these methods.

This paper has considered the impact on estimates and standard errors when class membership is used as the outcome in a regression analysis. But the question arises of what happens when class membership is used as a predictor in a regression of a distal outcome on class membership. Petras and Masyn (2009) considered this issue in their study looking at how to incorporate covariates and distal outcomes in growth mixture modeling. In their study, the authors found that treating class membership as an observed variable leads to different conclusions when compared with incorporating the distal outcome while forming the latent classes. Future work in this area should examine under what conditions it is appropriate to treat class membership as an observed variable when it is used as a predictor of a distal outcome.

One aspect that has not been discussed is the comparability of the regression models. With most likely class membership, probability regression, probability-weighted regression and pseudo-class draws the covariate is included after the latent class analysis has already been conducted. In the single step method, the covariate was included at the same time as the latent class analysis was conducted. By including the covariate with the latent class analysis, there is a potential for a direct effect between the covariates and the latent class indicators. This direct effect can potentially impact the relationship between the covariate and the classes because the covariate effect is partially filtered through the relationship between the covariate and the latent class indicators. In the simulation study, this was not problematic because the models were specified so that there was no direct effect between the covariate and the latent class indicators. But the potential for direct effects does arise in the real data examples. Nylund and Masyn (2008) are currently investigating the impact of mis-specifying direct effects in latent class analyses.

Despite the limitations discussed, this study is not without its strengths. While other authors have discussed the problems with using most likely class membership, this is one of the first studies to look at how problematic it can be when used in regression and mean comparisons. This study takes other authors' work one step further by also comparing the results using most likely class membership regression to those obtained from other methods: probability-weighted regression, probability, pseudo-class draws, and including the covariate while the latent classes are formed. This is the first study to make suggestions about when it is appropriate to use these techniques in practice.

References

Asparouhouv, T. & Muthén, B.O. (2007). *Wald test of mean equality for potential latent class predictors in mixture modeling*. From http://www.statmodel.com/download/MeanTest1.pdf

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, 1375-1386.

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.311-352). New York: Plenum.

Dayton, C.M., & Macready, G.B. (1998). Concomitant variable latent class analysis. *Journal of the American Statistical Association*, 83, 173-178.

Dolan, L., Kellam, S.G., Brown, C.H., Werthamer-Larsson, L., Rebok, G.W., Mayer, L.S., Laudolff, J., Turkkan, J.S., Ford, C., and Wheeler, L. (1993). The short-term impact of two classroom based preventive intervention trials on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology* 14, 317-345.

Formann, A.K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87, 476-486.

Hagenaars, J.A. (1993). *Loglinear models with latent variables*. London: Sage.

van der Heijden, P., Dessens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm. *Journal of Educational and Behavioral Statistical, 31(3), 215-229.*

Ialongo, L.N., Werthamer, S., Kellam, S.K., Brown, C.H., Wang, S., and Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology,* 27, 599-641.

Kellam, S.G., Rebok, G.W., Ialongo, N., and Mayer, L.S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry*, 25, 359-382.

Lazarfeld, P., & Henry, N. (1968). *Latent Structure Analysis*. New York: Houghton Mifflin.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley.

Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In Hancock, G. R., & Samuelsen, K. M. (Eds.), Advances in latent variable mixture models, pp. 1-24. Charlotte, NC: Information Age Publishing, Inc.

Muthén, L., & Muthén, B. (1998-2008). *Mplus User's Guide*. Fifth Edition. Los Angeles, CA: Muthen & Muthen.

Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism*: *Clinical and Experimental Research*, 24, 882-891.

Nagin, D.S., & Tremblay, R.E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods*, 6(1), 18-34.

Nylund, K.L. & Masyn, K. (2008). Covariates and Mixture Modeling: Results of a Simulation Study Exploring the Impact of Misspecified Covariate Effects. Paper presented at the annual meeting of the Society for Prevention Research.

Petras, H., & Masyn, K. (2009). General growth mixture analysis with antecedents and consequences of change. Forthcoming in Piquero, A., & Weisburd, D. (Eds.), *Handbook of Quantitative Criminology*.

Roeder K., Lynch, K.G., & Nagin D.S. (1999). Modeling uncertainty in latent class membership: a case study in criminology. *Journal of the American Statistical Association*, 94 (47), 766-7776.

Wang C.P., Brown, C.H., Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, 100 (3), 1054-1076.

Appendix: Example M*plus* code for the case in Table 7 where entropy is high, and the sample size is large.

Input 1: Monte Carlo input to generate data under a latent class model with a covariate and analyze the generated data using the same model.

```
 Title:    Monte Carlo Settings
           High entropy = 0.8, Sample size = 1,000,
           Class on covariate = 0.5
 Montecarlo:
     Names are u1-u10 x; !Assigning names to generated variables
     Generate = u1-u10(1);
           ! Generating 10 categorical items with one threshold
     Categorical = u1-u10; ! Specifying which items are categorical
     Genclasses = c(2); ! Generating 2 classes named c
     Classes = c1(2); ! Analyzing data with 2 classes called c1
     Nobservations = 1000; !Sample size of each data set
     Seed = 86142; ! Specifying seed to be used for random draws
     Nrep = 500; !Number of data replications
     Repsave = ALL; !Specifies that all replications will be saved
     Save = sim*.dat;
           ! Naming the files to which the data will be saved. The
           ! asterisk is replaced by the replication number.
 Analysis: Type = Mixture;
           !Specifies that a mixture model will be used
           Starts = 0; !Specifies the number of random starts
 Model Population:
     !Specifying the model from which the data is generated
     %Overall%
     [x@0];  ! Setting covariate mean to zero
```

```
      x@1;  ! Setting covariate variance to zero
      [c#1*0];
      c#1 on x*0.5; !Regression of class on covariate
      %c#1% !Specifying the class specific item thresholds
      [u1$1*1 u2$1*1 u3$1*1 u4$1*1 u5$1*1 u6$1*-1 u7$1*-1 u8$1*-1 u9$1*-1
      u10$1*-1];
      %c#2%
      [u1$1*-1 u2$1*-1 u3$1*-1 u4$1*-1 u5$1*-1 u6$1*1 u7$1*1 u8$1*1 u9$1*1
      u10$1*1];
  Model: !Specifying the model by which the data is analyzed
      %Overall%
      [c1#1*0];
      c1#1 on x*0.5;
      %c1#1%
      [u1$1*1 u2$1*1 u3$1*1 u4$1*1 u5$1*1 u6$1*-1 u7$1*-1 u8$1*-1 u9$1*-1
      u10$1*-1];
      %c1#2%
      [u1$1*-1 u2$1*-1 u3$1*-1 u4$1*-1 u5$1*-1 u6$1*1 u7$1*1 u8$1*1 u9$1*1
      u10$1*1];
Output: Tech9;
      !Provides information about the convergence of each replication
```

Input 2: Analyzing data generated in Input 1 with a latent class model in order to be able to save class probabilities and most likely class membership.

```
Variable:
      Names are u1-u10 x c0;
      Usevariables are u1-u10;
            ! Specifies which variables are to be used in the analysis
      Categorical = u1-u10;
      Classes = c1(2);
      Auxiliary = x;
            ! Identifies auxiliary variables and includes them when
            ! saving data
Data: File is sim.dat;
Analysis:
      Type = Mixture;
      Starts = 0;
Model:
      %Overall%
      [c1#1*0];
      %c1#1%
      [u1$1*1 u2$1*1 u3$1*1 u4$1*1 u5$1*1 u6$1*-1 u7$1*-1 u8$1*-1 u9$1*-1
      u10$1*-1];
      %c1#2%
      [u1$1*-1 u2$1*-1 u3$1*-1 u4$1*-1 u5$1*-1 u6$1*1 u7$1*1 u8$1*1 u9$1*1
      u10$1*1];
Output: Tech9;
Savedata: File is simpp.dat; !Name of file to save data to
      Save = Cprob;
            !Saving class probabilities and membership information
```

Input 3: Taking most likely class membership saved from Input 2 and regressing class membership on the covariate x.

```
Variable:
      Names are u1-u10 x p1 p2 c0;
      Usevariables are c0 x;
      Nominal = c0; ! Specify which dependent variable is nominal


Data: File is simpplist.dat;
            !Name of file that contains a list of names of the !generated
            data
      Type = Montecarlo;
            !Specifies that data are multiple data sets generated using !the
            Monte Carlo option
Analysis: Algorithm = Integration;
      !Specifies that numerical integration should be used
Model:
      c0#1 on x*0.5; ! Regressing class membership on the covariate
```

Input 4: Taking most likely class probabilities saved from Input 2, converting them to the logit scale and regressing them on the covariate x.

```
Variable:
      Names are u1-u10 x p1 p2 c0;
      Usevariables are x logit;
      !Uses logit variable created in Define statement below
Data: File is simpplist.dat;
      Type = Montecarlo;
Define:
      If(p1 lt .00001)then p1 =.00001;
      If(p1 gt .9999)then p1 = .9999;
      !Recoding p1 so that values of 0 and 1 will be able to be converted to
      !the logit scale
      logit = log(p1/(1-p1));
      !Creating a new variable called logit which converts p1 to logit  scale
```

```
Model:
          logit on x*0.5; !Regressing logit on covariate
```

Input 5: Taking most likely class probabilities saved from Input 2 and using the probabilities as weights in a regression.

```
Variable:
      Names are u1-u10 x p1 p2 c0;
      Usevariables are u1-u10 x p1 p2;
      Categorical = u1-u10;
      Classes = c1(2);
      Training = p1 p2 (probabilities);
            !Specifies which variables contain information about class
            !membership and that the information is the posterior
            !probabilities
Data: File is simpplist.dat;
      Type = Montecarlo;
Analysis: Type = Mixture;
      Starts = 0;
Model:
      %Overall%
      [c1#1*0];
       c1#1 on x*0.5;
      %c1#1%
      [u1$1*1 u2$1*1 u3$1*1 u4$1*1 u5$1*1 u6$1*-1 u7$1*-1 u8$1*-1 u9$1*-1
      u10$1*-1];
      %c1#2%
      [u1$1*-1 u2$1*-1 u3$1*-1 u4$1*-1 u5$1*-1 u6$1*1 u7$1*1 u8$1*1 u9$1*1
      u10$1*1];
```

Input 6: Taking data generated from Input 2 and analyzing using a latent class model with pseudo-class draws.

```
Variable:
      Names are u1-u10 x c0;
      Usevariables are u1-u10;
      Categorical = u1-u10;
      Classes = c1(2);
      Auxiliary = x(r);
            ! Specifies the auxiliary variable and asks for psuedo-!class
            regression
Data: File is simlist.dat;
Analysis: Type = Mixture;
      Starts = 0;
Model:
      %Overall%
      [c1#1*0];
      %c1#1%
      [u1$1*1 u2$1*1 u3$1*1 u4$1*1 u5$1*1 u6$1*-1 u7$1*-1 u8$1*-1 u9$1*-1
      u10$1*-1];
       %c1#2%
      [u1$1*-1 u2$1*-1 u3$1*-1 u4$1*-1 u5$1*-1 u6$1*1 u7$1*1 u8$1*1 u9$1*1
      u10$1*1];
```

Input7: Taking data generated using Input 2 and analyzing with a latent class model with the covariates included.

```
Variable:
      Names are u1-u10 x c0;
      Usevariables are u1-u10 x;
      Categorical = u1-u10;
      Classes = c1(2);
Data: File is simlist.dat;
      Type = Montecarlo;
Analysis:
      Type = Mixture;
      Starts = 0;
Model:
      %Overall%
      [c1#1*0];
      c1#1 on x*0.5;
      %c1#1%
      [u1$1*1 u2$1*1 u3$1*1 u4$1*1 u5$1*1 u6$1*-1 u7$1*-1 u8$1*-1 u9$1*-1
      u10$1*-1];
      %c1#2%
      [u1$1*-1 u2$1*-1 u3$1*-1 u4$1*-1 u5$1*-1 u6$1*1 u7$1*1 u8$1*1 u9$1*1
      u10$1*1];
```

Input 8: Taking data saved in Input 2 and analyzing with a latent class model with pseudo class draws used to form Wald chi-square tests for mean comparisons.

```
Variable:
      Names are u1-u10 x c0;
      Usevariables = u1-u10;
      Categorical = u1-u10;
      Classes = c1(2);
      Auxiliary = x(e);
            ! Specifies the auxiliary variable and asks for psuedo-!class
            Wald chi-square test
```

```
Data: File = simlist.dat;
Analysis: Type = Mixture;
      Starts = 0;
Model:
      %Overall%
      %c1#1%
      [u1$1*1 u2$1*1 u3$1*1 u4$1*1 u5$1*1 u6$1*-1 u7$1*-1 u8$1*-1 u9$1*-1
      u10$1*-1];
      %c1#2%
      [u1$1*-1 u2$1*-1 u3$1*-1 u4$1*-1 u5$1*-1 u6$1*1 u7$1*1 u8$1*1 u9$1*1
      u10$1*1];
```

Figures and Tables

Table One. Antisocial Behavior Mean Comparison Results

Mean Comparison Table

	Sex	Black	Hispanic	Age
Pseudo-Class Chi-Square	287.6 (0)	112.9 (0)	8.49 (0.04)	101.2 (0)
Equal Proportions Chi-Square	627.4 (0)	122.7 (0)	14.83 (0)	255.5 (0)

Table Two. Antisocial Behavior Regression Results.

	Gender			Age			Black			Hispanic		
	High	Person	Drug	High	Person	Drug	High	Person	Drug	High	Person	Drug
Most Likely Class Regression												
Est.	2.18	0.96	0.25	-0.16	-0.16	0.06	-0.22	0.36	-0.72	-0.32	-0.02	-0.53
S.E.	0.11	0.06	0.07	0.02	0.02	0.02	0.1	0.07	0.08	0.12	0.08	0.09
Est./S.E.	19.4	15.9	3.78	-8.02	-11.1	3.85	-2.17	5.35	-8.66	-2.65	-0.24	-5.75
p-value	0	0	0	0	0	0	0.03	0	0	0.008	0.81	0
Probability Regression												
Est.	2.77	0.35	-0.45	-0.19	-0.18	0.21	-0.53	1.19	-0.91	-0.37	0.42	-0.67
S.E.	0.117	0.084	0.087	0.027	0.019	0.02	0.134	0.096	0.1	0.158	0.11	0.12
Est./S.E.	23.65	4.22	-5.14	-7.15	-9.48	10.39	-3.96	12.48	-9.11	-2.36	3.66	-5.77
p-value	0	0	0	0	0	0	0	0	0	0.018	0	0
Probability-Weighted Regression												
Est.	2.04	0.92	0.24	-0.15	-0.15	0.06	-0.23	0.36	-0.61	-0.31	-0.02	-0.47
S.E.	0.1	0.06	0.06	0.012	0.013	0.013	0.096	0.06	0.07	0.11	0.07	0.08
Est./S.E.	20.9	16.9	3.97	-8.28	-11.3	4.35	-2.41	5.98	-8.69	-2.78	-0.25	-5.81
p-value	0	0	0	0	0	0	0.02	0	0	0.005	0.8	0
Pseudo-Class Regression												
Est.	1.99	0.86	0.29	-0.14	-0.13	0.05	-0.28	0.33	-0.6	-0.32	-0.01	-0.44
S.E.	0.12	0.07	0.07	0.02	0.02	0.02	0.11	0.08	0.09	0.13	0.09	0.09
Est./S.E.	17.3	11.8	3.97	-6.19	-7.63	3.12	-2.45	4.31	-6.79	-2.51	-0.11	-4.47
p-value	0	0	0	0	0	0.002	0.014	0	0	0.012	0.92	0
Covariates in Model												
Est.	2.65	1.42	0.22	-0.22	-0.26	0.08	0.22	0.84	-0.96	-0.19	0.21	-0.69
S.E.	0.16	0.09	0.1	0.03	0.02	0.21	0.15	0.12	0.14	0.134	0.12	0.12
Est./S.E.	16.1	15.5	2.14	-8.36	-10.8	3.61	0.16	6.94	-7.01	-1.46	1.77	-5.89
p-value	0	0	0.03	0	0	0	0.88	0	0	0.145	0.08	0

Table Three. Aggression Mean Comparison Results

	Race	Gender	Treatment
Pseudo-Class Chi-Square	8.37 (0.02)	42.57 (0)	0.33 (0.85)
Equal Proportions Chi-Square	13.95(0.001)	45.2 (0)	2.31 (0.32)

Table Four. Aggression Regression Results

	Gender		Black		Treatment	
	High	Verbal	High	Verbal	High	Verbal
Most Likely Class Membership as Outcome						
Est.	1.343	0.347	0.713	0.47	-0.005	0.192
S.E.	0.199	0.14	0.206	0.154	0.186	0.139
Est./S.E.	6.743	2.478	3.457	3.057	-0.025	1.382
p-value	0	0.013	0.001	0.002	0.98	0.167
Probability Regression						
Est.	2.719	-0.265	0.884	0.291	-0.789	0.063
S.E.	0.356	0.23	0.38	0.246	0.359	0.232
Est./S.E.	7.64	-1.15	2.33	1.184	-2.19	0.273
p-value	0	0.248	0.02	0.236	0.028	0.785
Probability - Weighted Regression						
Est.	1.228	0.339	0.621	0.366	-0.15	0.055
S.E.	0.17	0.126	0.179	0.135	0.167	0.126
Est./S.E.	7.208	2.697	3.469	2.706	-0.897	0.434
p-value	0	0.007	0.001	0.007	0.37	0.664
Pseudo-Class Regression						
Est.	1.199	0.335	0.56	0.313	-0.137	0.035
S.E.	0.2	0.157	0.21	0.167	0.191	0.156
Est./S.E.	5.995	2.142	2.667	1.874	-0.718	0.225
p-value	0	0.032	0.008	0.061	0.473	0.822
Covariates in Model						
Est.	1.52	0.38	0.716	0.442	-0.154	0.096
S.E.	0.235	0.175	0.229	0.184	0.198	0.173
Est./S.E.	6.47	2.174	3.129	2.404	-0.779	0.554
p-value	0	0.03	0.002	0.016	0.44	0.58

Table Five. Monte Carlo Covariate Mean Comparison

Entropy	0.4				0.6				0.8				1.0			
Sample Size	250		1,000		250		1,000		250		1,000		250		1,000	
	Pc	Mlcm	Pc	Mlcm	Pc	Mlcm	Pc	Mlcm	Pc	Mlcm	Pc	Mlcm	Pc	Mlcm	Pc	Mlcm
Class on Covariate = 0																
Value	0.353	0.063	0.312	0.001	0.498	0.001	0.52	0.002	0.724	0.007	0.704	0.003	0.902	0.006	0.896	0.003
S.D.	0.532	0.994	0.411	1.02	0.676	0.967	0.698	0.981	1	0.993	0.982	0.981	0.994	0.995	0.953	0.961
P-value of average stat	0.552	0.083	0.576	0.972	0.48	0.976	0.471	0.964	0.395	0.936	0.401	0.956	0.382	0.941	0.395	0.954
Average *p*-value	0.669	0.285	0.674	0.282	0.596	0.287	0.603	0.281	0.544	0.282	0.551	0.282	0.516	0.281	0.524	0.286
Number rep $p < 0.05$	1	21	0	26	3	21	1	18	13	23	12	20	18	18	14	17
Proportion rep $p < 0.05$	0.002	0.042	0	0.052	0.006	0.042	0.002	0.036	0.026	0.046	0.024	0.04	0.036	0.036	0.028	0.034
Class on Covariate = 0.5																
Value	1.72	2.96	6.39	17.8	5.25	8.01	18.9	32.9	9.52	11.4	35.6	45.1	3.8	15.4	58.4	60.1
S.D.	1.55	1.05	2.94	1.04	3.31	1.03	6.62	1.02	5.37	1.04	10.2	1.04	16.2	1.06	16.2	1.01
P-value of average stat	0.19	0.085	0.011	0	0.022	0.005	0	0	0.002	0.001	0.007	0	0	0	0	0
Average *p*-value	0.31	0.136	0.03	0.004	0.081	0.038	0	0	0.027	0.017	0	0	0.007	0.006	0	0
Number rep $p < 0.05$	98	195	410	493	302	380	499	500	437	453	500	500	473	486	500	500
Proportion rep $p < 0.05$	0.196	0.39	0.82	0.986	0.604	0.76	0.998	1	0.874	0.906	1	1	0.946	0.972	1	1

Note: Pc – Psuedo-class chi-square; Mlcm – Most likely class membership squared *t*-test.

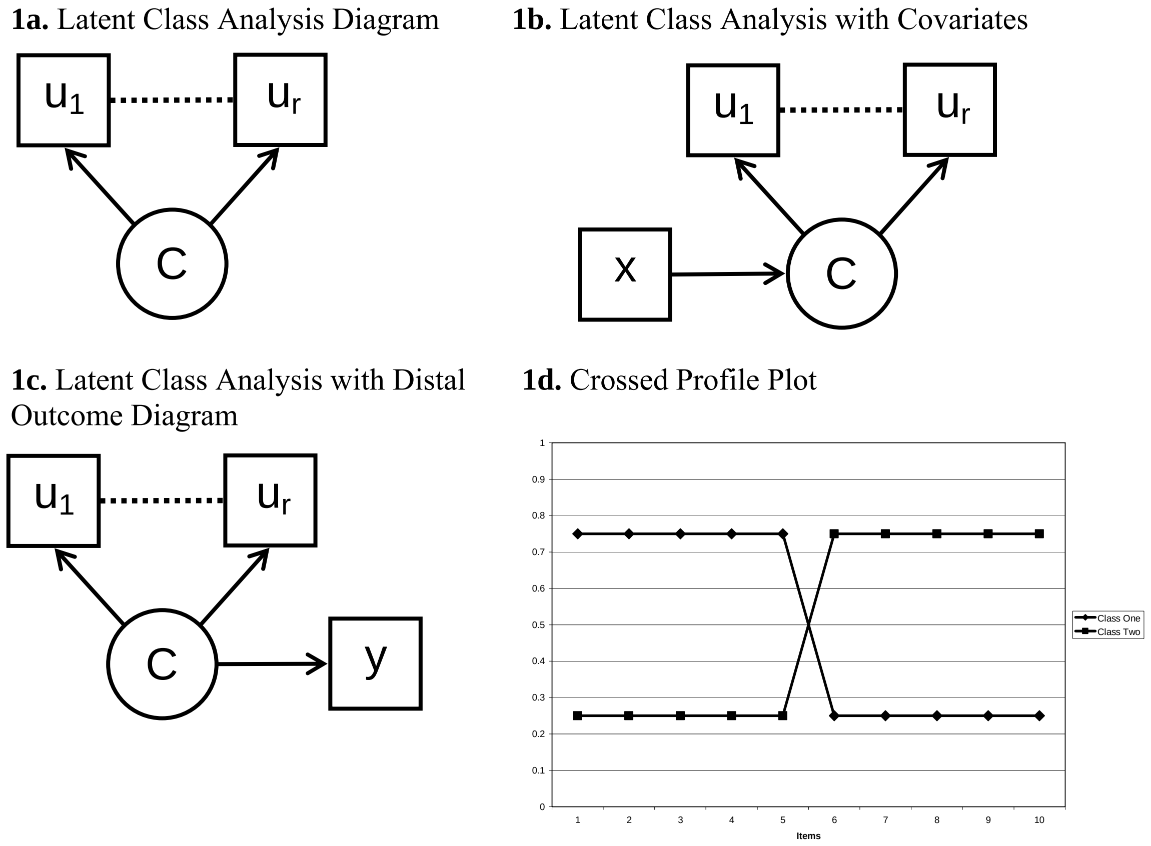
Table Six. Monte Carlo Covariate Regression Results: Logistic Regression Coefficient= 0

Entropy	0.4		0.6		0.8		1.0	
Sample Size	250	1,000	250	1000	250	1000	250	1000
Most Likely Class Membership Regression								
Est.	0.003	0.002	0.004	0.003	0.011	0.003	0.01	0.004
S.D.	0.143	0.066	0.126	0.063	0.128	0.063	0.128	0.061
S.E.	0.141	0.065	0.129	0.064	0.128	0.064	0.128	0.063
MSE	0.02	0.004	0.016	0.004	0.017	0.004	0.017	0.004
Coverage	0.954	0.938	0.954	0.955	0.954	0.95	0.958	0.958
Est./S.E.	0.021	0.031	0.031	0.047	0.086	0.047	0.076	0.0635
Prop Est/ SE > 1.96	0.046	0.062	0.046	0.045	0.046	0.05	0.042	0.042
Probability Regression								
Est.	0.001	0.005	0.009	0.011	0.005	0.002	0.01	0.004
S.D.	0.192	0.071	0.147	0.195	0.094	0.047	0.128	0.061
S.E.	0.183	0.069	0.141	0.189	0.092	0.046	0.128	0.063
MSE	0.033	0.005	0.022	0.005	0.009	0.002	0.017	0.004
Coverage	0.946	0.948	0.96	0.95	0.944	0.952	0.958	0.958
Est./S.E.	0.005	0.072	0.064	0.058	0.054	0.043	0.076	0.0635
Prop Est/ SE > 1.96	0.054	0.064	0.046	0.052	0.058	0.052	0.042	0.042
Probability-Weighted Regression								
Est.	0.0004	0.003	0.005	0.001	0.008	0.002	0.01	0.004
S.D.	0.105	0.043	0.106	0.052	0.118	0.057	0.128	0.0612
S.E.	0.1	0.042	0.106	0.052	0.117	0.058	0.128	0.0634
MSE	0.011	0.002	0.011	0.003	0.014	0.003	0.017	0.004
Coverage	0.946	0.932	0.956	0.95	0.95	0.95	0.958	0.958
Est./S.E.	0.004	0.071	0.047	0.019	0.068	0.034	0.076	0.0635
Prop Est/ SE > 1.96	0.054	0.068	0.044	0.05	0.05	0.05	0.042	0.042
Pseudo-Class Regression								
Est.	0.0004	0.003	0.005	0.001	0.008	0.002	0.01	0.004
S.D.	0.105	0.043	0.106	0.052	0.118	0.057	0.128	0.061
S.E.	0.167	0.081	0.149	0.074	0.139	0.069	0.128	0.063
MSE	0.011	0.002	0.011	0.003	0.014	0.003	0.017	0.004
Coverage	0.999	0.999	0.994	0.998	0.982	0.978	0.958	0.958
Est./S.E.	0.002	0.037	0.034	0.014	0.058	0.029	0.076	0.0635
Prop Est/ SE > 1.96	0	0	0.006	0.002	0.018	0.022	0.042	0.042
Covariates in Model								
Est.	0.01	0.007	0.008	0.003	0.009	0.002	0.01	0.004
S.D.	0.283	0.106	0.164	0.079	0.144	0.069	0.128	0.061
S.E.	0.301	0.107	0.172	0.079	0.144	0.07	0.128	0.063
MSE	0.079	0.011	0.027	0.006	0.021	0.005	0.017	0.004
Coverage	0.952	0.948	0.964	0.952	0.952	0.952	0.958	0.958
Est./S.E.	0.033	0.065	0.047	0.038	0.063	0.029	0.076	0.0635
Prop Est/ SE > 1.96	0.048	0.052	0.036	0.048	0.048	0.048	0.042	0.042

Table Seven. Monte Carlo Covariate Regression Results: Logistic Regression Coefficient = 0.5

Entropy	0.4		0.6		0.8		1.0	
Sample Size	250	1,000	250	1000	250	1000	250	1000
Most Likely Class Membership Regression								
Est.	0.245	0.274	0.371	0.371	0.445	0.436	0.518	0.506
S.D.	0.148	0.069	0.138	0.068	0.142	0.07	0.148	0.069
S.E.	0.143	0.066	0.135	0.067	0.137	0.067	0.139	0.069
MSE	0.088	0.056	0.036	0.021	0.023	0.009	0.022	0.005
Coverage	0.54	0.078	0.812	0.496	0.918	0.828	0.952	0.956
Est./S.E.	1.71	4.15	2.74	5.53	3.25	6.51	3.72	7.33
Prop Est/ SE > 1.96	0.402	0.997	0.774	1	0.906	1	0.976	1
Probability Regression								
Est.	0.263	0.237	0.303	0.287	0.307	0.302	0.518	0.506
S.D.	0.143	0.053	0.108	0.05	0.097	0.047	0.148	0.069
S.E.	0.133	0.05	0.102	0.049	0.095	0.047	0.139	0.069
MSE	0.076	0.072	0.135	0.048	0.047	0.043	0.022	0.005
Coverage	0.51	0.004	0	0.018	0.498	0	0.952	0.956
Est./S.E.	1.97	4.74	2.97	5.86	3.23	10.3	3.72	7.33
Prop Est/ SE > 1.96	0.52	0.996	0.84	1	0.892	1	0.976	1
Probability-Weighted Regression								
Est.	0.197	0.2	0.326	0.321	0.418	0.409	0.518	0.506
S.D.	0.106	0.045	0.112	0.055	0.13	0.062	0.148	0.069
S.E.	0.099	0.042	0.108	0.053	0.123	0.06	0.139	0.069
MSE	0.103	0.092	0.043	0.035	0.023	0.012	0.022	0.005
Coverage	0.156	0	0.616	0.09	0.87	0.664	0.952	0.956
Est./S.E.	1.99	4.76	3.02	6.06	3.39	6.81	3.72	7.33
Prop Est/ SE > 1.96	0.518	0.998	0.86	1	0.932	1	0.976	1
Pseudo-Class Regression								
Est.	0.197	0.2	0.326	0.321	0.418	0.409	0.518	0.506
S.D.	0.106	0.045	0.112	0.055	0.131	0.062	0.148	0.069
S.E.	0.176	0.083	0.156	0.077	0.148	0.074	0.139	0.069
MSE	0.103	0.093	0.043	0.035	0.024	0.012	0.022	0.005
Coverage	0.58	0.2	0.856	0.304	0.934	0.792	0.952	0.956
Est./S.E.	1.12	2.41	2.09	4.17	2.82	5.53	3.72	7.33
Prop Est/ SE > 1.96	0.186	0.81	0.578	0.998	0.864	1	0.976	1
Covariates in Model								
Est.	0.536	0.516	0.521	0.506	0.519	0.506	0.518	0.506
S.D.	0.316	0.122	0.19	0.09	0.166	0.079	0.148	0.069
S.E.	0.346	0.121	0.19	0.088	0.159	0.077	0.139	0.069
MSE	0.101	0.015	0.037	0.008	0.028	0.006	0.022	0.005
Coverage	0.97	0.954	0.96	0.94	0.952	0.944	0.952	0.956
Est./S.E.	1.55	4.26	2.74	5.75	3.26	6.57	3.72	7.33
Prop Est/ SE > 1.96	0.362	0.996	0.81	1	0.926	1	0.976	1

Figure One



1a. Latent Class Analysis Diagram

1b. Latent Class Analysis with Covariates

1c. Latent Class Analysis with Distal Outcome Diagram

1d. Crossed Profile Plot

Figure Two. Antisocial Behavior data Profile Plot

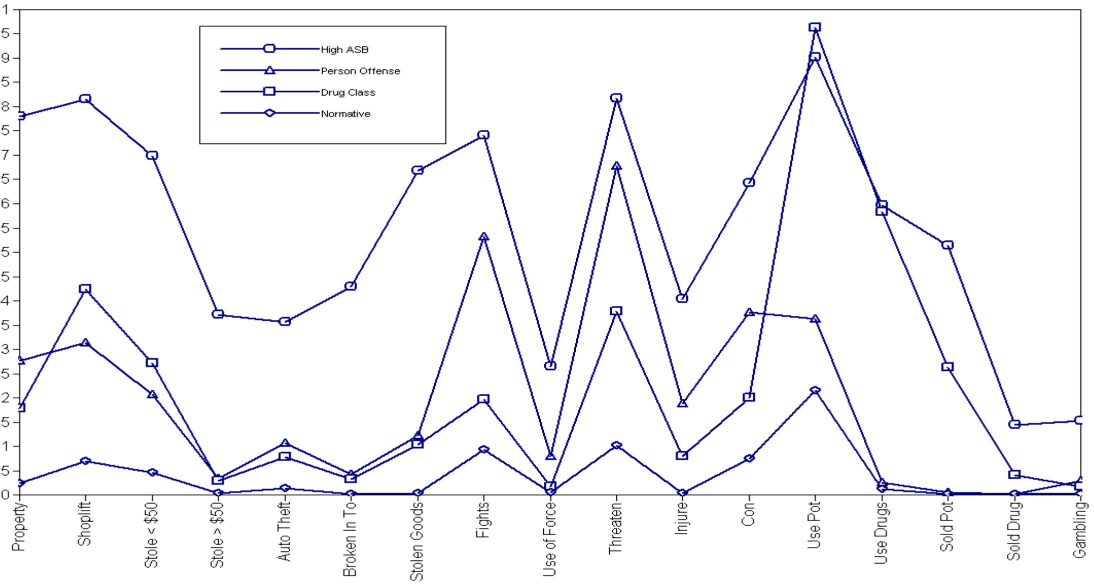
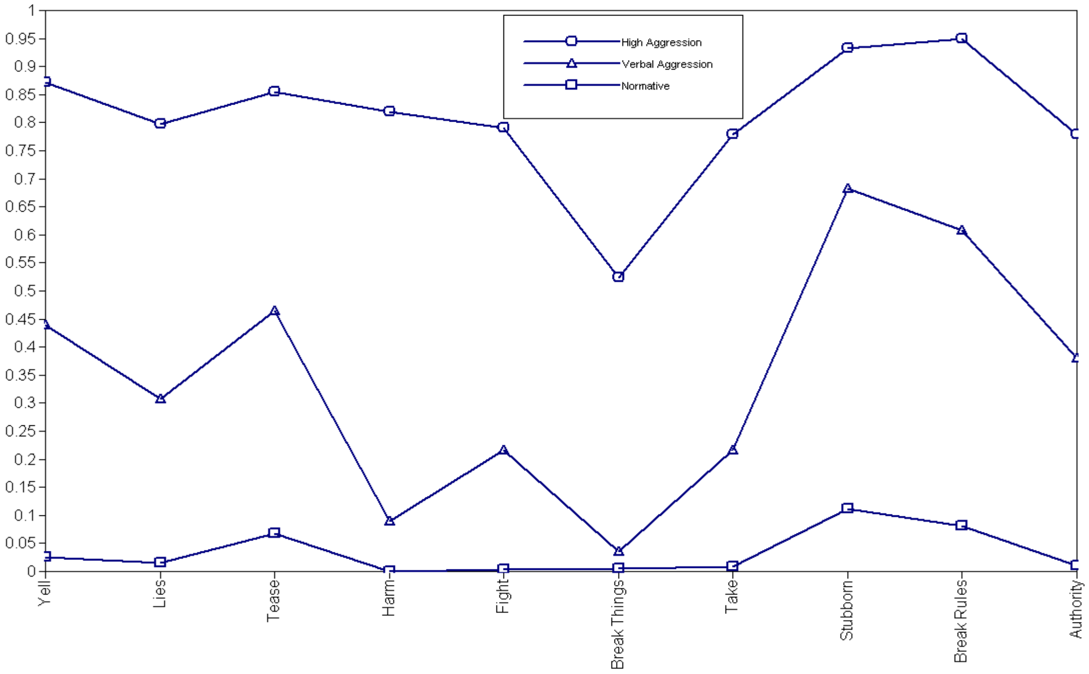


Figure Three. TOCA data Profile Plot


High Aggression
Verbal Aggression
Normative
Yell
Lies
Tease
Harm
Fight
Break Things
Take
Stubborn
Break Rules
Authority