Latent Class Analysis: An Alternative Perspective on Subgroup Analysis in Prevention and Treatment

Stephanie T. Lanza and
The Methodology Center, The Pennsylvania State University, 204 E. Calder Way Suite 400, State College, PA 16801, USA, SLanza@psu.edu

Brittany L. Rhoades
Prevention Research Center, The Pennsylvania State University, 206 Towers Building, University Park, PA 16802, USA

Abstract

The overall goal of this study is to introduce latent class analysis (LCA) as an alternative approach to latent subgroup analysis. Traditionally, subgroup analysis aims to determine whether individuals respond differently to a treatment based on one or more measured characteristics. LCA provides a way to identify a small set of underlying subgroups characterized by multiple dimensions which could, in turn, be used to examine differential treatment effects. This approach can help to address methodological challenges that arise in subgroup analysis, including a high Type I error rate, low statistical power, and limitations in examining higher-order interactions. An empirical example draws on *N*=1,900 adolescents from the National Longitudinal Survey of Adolescent Health. Six characteristics (household poverty, single-parent status, peer cigarette use, peer alcohol use, neighborhood unemployment, and neighborhood poverty) are used to identify five latent subgroups: Low Risk, Peer Risk, Economic Risk, Household & Peer Risk, and Multi-Contextual Risk. Two approaches for examining differential treatment effects are demonstrated using a simulated outcome: 1) a classify-analyze approach and, 2) a model-based approach based on a reparameterization of the LCA with covariates model. Such approaches can facilitate targeting future intervention resources to subgroups that promise to show the maximum treatment response.

In social, behavioral, and health research, prevention and intervention programs are often administered to populations without consideration of individual characteristics that might predict treatment response. Recently, however, there has been growing interest in individualizing treatments in order to administer the right program to the right individuals, thereby maximizing treatment effectiveness. For example, the Fast Track program (Conduct Problems Prevention Research Group 1992) was designed to periodically evaluate the needs of participants so that the particular needs of each youth could be addressed. Adaptive treatment strategies such as this rely on the successful identification of individual characteristics, sometimes called tailoring variables, which predict treatment response (Collins et al. 2004).

Traditionally, a treatment effect modifier is examined by incorporating it in a multiple regression model as a moderator. For example, Elkin et al. (1995) demonstrated that initial severity of depressive symptoms and impairment of functioning significantly moderated the response to four different treatments for major depressive disorder. Aiken and West (1991) provide a detailed presentation of estimating and interpreting statistical moderators. Substantive examples of this approach abound in recent literature, revealing individual and contextual characteristics that significantly moderate treatment effects (e.g., Baucom et al. 2009; Ondersma et al. 2009). More recently, extensions of this model such as moderated mediation have also been explored (e.g., Fairchild and MacKinnon 2009; MacKinnon 2009).

Although this traditional approach to subgroup analysis in evaluation studies can be informative, there are several methodological challenges. First, a study of heterogeneity in treatment effects can be vulnerable to high Type I error rates, particularly in exploratory studies when numerous characteristics are examined as moderators of the treatment effect. This issue of multiple comparisons is described in detail in a technical report by Schochet (2008). Consider as an example pairwise tests for different effects in five race/ethnicity subgroups within each gender, such as the difference in treatment effect between White females and Hispanic females. Based on the required 20 tests (10 pairwise tests within each gender), there is a 64% chance of a spurious finding. Unfortunately, the common Bonferroni correction to adjust the error rate for each test can be overly conservative, and results in a considerable loss of statistical power (Schochet 2008). The second methodological challenge is that treatment effects estimated within each subgroup typically are based on different sample sizes. Thus, the statistical power available to detect a treatment effect will vary across the subgroups. The power to detect differences between subgroups is greater when the subgroups are of equal size (e.g., males versus females) as compared to subgroup variables with less balance (e.g., minorities versus non-minorities). Third, there may be important higher-order interactions in subgroup variables that often cannot be examined due to the aforementioned methodological challenges (i.e., statistical power issues). For example, the detection of differential treatment effects for older White males compared to younger White males would require the four-way interaction between treatment, gender, race/ethnicity, and age group. Along these same lines, it is unlikely that every possible combination of subgroup variable under consideration is an important or even real type of individual in the population. The consideration of every possible combination only makes sense if one can assume no measurement error in the observed variables.

Finite mixture models present an opportunity to approach subgroup analysis from a different perspective. These statistical models are appropriate when one posits that a population is comprised of two or more underlying, latent subgroups defined by the intersection of numerous individual characteristics (Everitt and Hand 1981; McLachlan and Peel 2000; Titterington et al. 1985). In other words, particular combinations of responses to the set of characteristics define the latent subgroups. We suggest that a particular type of finite mixture model, latent class analysis (LCA), is a useful tool for identifying a set of underlying subgroups of individuals based on the intersection of multiple observed characteristics. These latent subgroups can comprise multiple dimensions. In the empirical demonstration presented below, six binary grouping variables are considered. If all possible patterns of responses to the six variables are considered there would be $2^6=64$ possible subgroups, resulting in $C(64,2)=(64*63)/2=2016$ pairwise tests. However, if the number of subgroups could be reduced substantially from 64 to represent only the key patterns of response across all six items, then pairwise comparisons may be feasible. In practice, it is unlikely that every observed response pattern actually reflects a unique and important type of individual; instead, it may be helpful to establish a smaller set of subgroups and assume that there is some level of measurement error in the items used to measure group

membership. By reducing the number of subgroups in this way, a solution can more readily lend itself to drawing prevention and treatment implications.

The overall goal of this study is to introduce LCA as an effective approach to identifying a small set of underlying subgroups characterized by multiple dimensions, which may differ in their response to treatment. In the next section we present a brief introduction to LCA, followed by an empirical demonstration where latent subgroups are identified based on the intersection of six contextual characteristics. In order to demonstrate how latent subgroups can be used to examine differential treatment effects, the empirical data set was augmented with two simulated variables: randomized treatment condition and an outcome (Grade 9 binge drinking). The association between latent subgroup membership and treatment effect was examined using two strategies: 1) a classify-analyze approach involving logistic regression, and 2) a model-based approach to LCA with a distal outcome, based on a reparameterization of the LCA with covariates model. In the Discussion, we comment on implications that can be drawn from the empirical study, and draw more general conclusions about using LCA for subgroup analysis in prevention and treatment research.

Rationale for Modeling Risk Subgroups for Adolescent Substance Use

A risk-focused approach to problem behavior such as adolescent substance use provides an effective framework for targeting at-risk individuals and attempting to mitigate risk through prevention and treatment programs (Catalano and Hawkins 1996; Coie et al. 1993; Hawkins et al. 1992). In addition to providing evidence for targeting individuals with the poorest outcomes, exposure to certain risk factors may be predictive of a stronger or weaker treatment response (Loeber 1990). Previous research has identified a broad range of risk factors for adolescent substance use that can be sorted into five domains, or levels, of risk: neighborhood (e.g., neighborhood disorganization, laws and norms), school (e.g., academic expectations, school bonding), family (e.g., family structure, family management, parent education), peer (e.g., rejection, peer drug use), and individual (e.g., sensation seeking, early initiation of problem behavior) (Hawkins et al. 1992). The various ecological domains, including family, peer, and neighborhood domains, often are considered to be particularly important areas to study in order to prioritize prevention and treatment efforts (e.g., Arthur et al. 2002).

As is illustrated in traditional subgroups analyses, the notion of individualizing programs for the prevention and treatment of problem behaviors is critical because the greatest reduction will likely be realized only if the right program is offered to the right individuals. Additionally, preventive intervention resources can be leveraged more effectively by matching appropriate program components to the different subgroups or targeting resources to those highest-risk subgroups. Several recent studies have demonstrated that a latent subgroup perspective could be successfully applied to the study of multiple demographic and contextual risks (e.g., Coffman et al. 2007; Lanza et al. 2010b; Syvertsen et al. 2010). In the present study, we apply this latent subgroups perspective to identify underlying subgroups of individuals who may respond differently to prevention and treatment programs due to their exposure to various combinations of contextual risk factors.

Latent Class Analysis

LCA (Collins and Lanza 2010; Goodman 1974; Lazarsfeld and Henry 1968) is a mixture model that posits that there is an underlying *unobserved* categorical variable that divides a population into mutually exclusive and exhaustive latent classes. Class membership of individuals is unknown but can be inferred from a set of measured items. LCA has been used in several ways in past prevention research. Most notably, behavioral outcomes such as stage of substance use have been modeled with this approach (e.g., Anderson et al. 2010;

Komro et al. 2010; Laska et al. 2009; Oxford et al. 2003; Scheier et al. 2008; Shin et al. 2010). LCA also has been used to better understand profiles of risk and protection for specific behavioral outcomes (Coffman et al. 2007; Lanza et al. 2010b; Syvertsen et al. 2010). The present study proposes a new application of LCA, where latent subgroups are identified with the goal of examining differential treatment effects within each class.

The mathematical model for LCA can be expressed as follows. Let y_j represent element j of a response pattern $\mathbf{y}$. Let us establish an indicator function $I(y_j=r_j)$ that equals 1 when the response to variable $j=r_j$, and equals 0 otherwise. Then the probability of observing a particular vector of responses is

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

where γ_c is the probability of membership in latent class c and $\rho_{j,r_j|c}^{I(y_j=r_j)}$ is the probability of response r_j to item j, conditional on membership in latent class c. The γ parameters represent a vector of latent class membership probabilities that sum to 1. The ρ parameters represent a matrix of item-response probabilities conditional on latent class membership. The degrees of freedom are calculated as the number of possible response patterns (i.e., number of cells in the contingency table formed by crossing all observed items) minus the number of freely estimated parameters minus one. Parameter estimation is typically performed using an EM algorithm.

Model selection in LCA can involve both absolute fit of a particular model and relative fit of two or more competing models. A common measure of absolute model fit in categorical models is the G^2 likelihood-ratio chi-square statistic, which in our case tests the null hypothesis that the specified LCA model fits the data (Agresti 2002). This statistic has an asymptotic chi-square distribution; thus, when the sample size is sufficient and the degrees of freedom are not too large, this value can be compared to the chi-square distribution with degrees of freedom given by the LCA model. A significant p-value indicates lack of model fit in absolute terms. It is also possible to empirically derive a p-value for this statistic using a parametric bootstrap (see, for example, Collins et al. 1993; Langeheine et al. 1996). The AIC (Akaike 1974), BIC (Schwarz 1978), CAIC (Bozdogan 1987), and aBIC (Sclove 1987) are information criteria that can be used to compare relative fit of models with different numbers of latent classes (for example, four versus five classes). For all of these information criteria, a lower value suggests a more optimal balance between model fit and parsimony. Other approaches for assessing relative fit include a parametric bootstrap of the log-likelihood difference statistic (McLachlan and Peel 2000; Nylund et al. 2007). Model selection is discussed in detail in Collins and Lanza (2010).

Similar to more standard structural equation models, grouping variables such as gender, race/ethnicity, or treatment group can be included in LCA to examine whether measurement invariance of the latent variable holds across groups and, assuming that the same measurement model can be applied to all groups, to examine group differences in the latent class prevalences. To test for measurement invariance, the difference in the G^2 statistic between a model with item-response probabilities freely estimated within each group and a model where these probabilities are constrained to be equal across groups can be compared to the chi-square distribution for the difference in the models' degrees of freedom. A significant p-value indicates evidence for different measurement across groups.

Covariates can also be included in LCA to examine associations between the covariates and subgroup membership. These associations are modeled using a logistic link function,

producing a set of logistic regression coefficients that show how different levels on the covariates predict subgroup membership (Dayton and Macready 1988). When exponentiated, these coefficients are transformed into more easily interpreted odds ratios. If the 95% confidence interval for an odds ratio does not include the value 1.0, the covariate is significantly associated with an increase (or decrease) in odds of membership in a specified latent class relative to a reference latent class corresponding to a different level on the covariate. It is important to note that the latent class measurement model and the prediction model are estimated simultaneously. This is important since each individual's true subgroup membership is not known. Instead, each individual has a (typically nonzero) probability of belonging to each subgroup. See Lanza et al. (2007) and Collins and Lanza (2010) for technical details and empirical examples of multiple-groups LCA and LCA with covariates.

The statistical model used to include covariates in LCA is one of predicting latent class membership from one or more covariates. Therefore, the association is interpreted in a particular direction—in terms of the covariate predicting subgroup membership. Often, however, interest lies in the reverse interpretation. For example, in order to estimate the incidence of a behavioral outcome across the different subgroups, the logical interpretation would involve the latent class variable predicting the covariate. Because each individual's latent class membership is not known with certainty, however, a model-based approach to treating the latent class variable as the independent variable in predicting an outcome is statistically challenging. Fortunately, when the covariate is categorical (such as the behavioral outcomes examined in the present study), the behavior can simply be included as a covariate in the usual sense (i.e., as a predictor). The resultant logistic regression parameters, along with the known marginal frequencies of that behavior, can then be transformed to obtain the desired results. For example, instead of using binge drinking to predict latent class membership, we are able to use the latent classes to predict binge drinking. Further, by incorporating the treatment condition as a grouping variable, we can assess the differential treatment effect across latent classes. This model-based approach to LCA with a distal outcome is demonstrated in the empirical results below. In addition, step-by-step directions on how to reparameterize the LCA with covariates model in order to calculate the probability of an outcome given latent class membership using a model-based approach can be found at http://methodology.psu.edu/. Included on this site is an Excel-based probability calculator that prevention scientists can modify to use in their own work.

A Note on Multiple Modes

For each LCA model under consideration, multiple sets of random starting values should be specified in order to confirm that a solution does not reflect suboptimal estimates caused by a local, as opposed to global, mode (maximum of the likelihood function). If one solution yielding the maximum value of the likelihood function is found for the majority of the sets of starting values, then one can have confidence that the maximum-likelihood solution has been identified. If instead the different random starting values all lead to different modes, the model is unidentified. Model fit should be assessed only for models where the maximum likelihood has been identified.

An Empirical Example: Differential Treatment Effect Across Latent Subgroups of Adolescents

Suppose we have a peer-based intervention program designed to reduce binge drinking that, in reality, is only effective for a subset of adolescents. After an initial examination of the intervention's impact, investigators discover no differences in binge drinking between treatment and control groups, suggesting that the intervention was ineffective. As a follow-up, they decide to conduct traditional moderation analyses to determine if the intervention

had an impact on any specific subgroups. However, because the subgroups for which this intervention is most effective are characterized by multiple, interacting dimensions, the traditional approach to subgroup analysis, which has limited statistical power to uncover higher-order interactions, will be unable to tease out these subgroup effects. This is the hypothetical example upon which the present study is based.

Here, we use LCA to identify underlying subgroups of adolescents at risk for problem behavior using data from The National Longitudinal Study of Adolescent Health (Add Health; Harris et al. 2009). Six unique observed variables were used to measure risk exposure within the household, peer, and neighborhood contexts. Rather than conducting a subgroup analysis for each of the six variables, they are used together to measure underlying, latent subgroups that characterize higher-order interactions across the six different risks. By augmenting the empirical data with a simulated treatment and outcome variable, we then demonstrate how to examine differential treatment effects (i.e., treatment group differences in the rate of Grade 9 binge drinking) using two approaches: 1) a classify-analyze approach involving logistic regression, and 2) a model-based approach to LCA with a distal outcome, based on a reparameterization of the LCA with covariates model.

Participants

Participants were drawn from Add Health. A sample of 80 high schools and 52 middle schools in the US was selected with unequal probability of selection. Incorporating systematic sampling methods and implicit stratification into the Add Health study design ensured this sample is representative of US schools with respect to region of country, urbanicity, school size, school type, and ethnicity. The current study included N=1,900 adolescents who were in eighth grade at Wave 1 (51% female; average age of 13.4 (SD=. 76); 60% White, 21% Black, 12% Hispanic, 7% other).

Measures

Household Risk—Two aspects of household risk were included: *household poverty* and *single-parent household*. For household poverty, youth were considered at-risk if their household income-to-needs ratio was below 1.85, as this is the threshold used to indicate qualification for WIC services, and has also been used to indicate "near poor" (Brooks-Gunn et al. 1997). Income-to-needs ratio was calculated based on the total household income recorded during the parent interview. For single-parent household, youth were considered at-risk if they lived with a parent/caregiver who at the time of interview was widowed, divorced, separated, or never married.

Peer Risk—Two aspects of peer risk were included: *peer cigarette use* and *peer alcohol use*. For these risk factors, youth were considered at-risk if one or more of their three best friends smoke at least one cigarette per day or drink alcohol at least once a month, respectively.

Neighborhood Risk—Two aspects of neighborhood risk were included: *neighborhood unemployment* and *neighborhood poverty*. For neighborhood unemployment, youth were considered at-risk if they lived in a neighborhood where the unemployment rate was 10% or greater within their census block. For neighborhood poverty, youth were considered at-risk if they lived in a neighborhood where the income-to-needs ratio was below 1.85 for at least 50% of the households within their census block.

Analytic Strategy

Identifying Latent Subgroups Using Empirical Data—Using the empirical data from Add Health described above, we first examined models with one latent class, two latent classes, and so on. For each model we assessed the issue of multiple modes by generating ten sets of random starting values. In cases where at least 80% of the sets converged to the same solution, we concluded that the model was identified. When this was not the case, 100 sets of random starting values were used to determine the confidence with which we could identify the maximum-likelihood solution. If a clear mode in the distribution of log-likelihood values emerged at the lowest value, we concluded that the model was identified.

Fit of individual models to the data were examined based on the G^2 statistic (and corresponding degrees of freedom and p-value) when appropriate; selection from the set of models with different numbers of latent classes was conducted based on information criteria (AIC, BIC, CAIC, and aBIC). In selecting our final model, we also took into consideration how well a solution could be interpreted; that is, whether the latent subgroups in a solution showed logical patterns, were distinct from the other subgroups, and could readily be labeled.

Simulating Treatment and Outcome Variables—Next, two simulated variables were created to represent treatment condition and Grade 9 binge drinking. *Treatment condition* was created by randomly assigning individuals, with equal probability, to a hypothetic peer-based intervention program (i.e., treatment) or treatment as usual (i.e., control group). The hypothetical intervention program was designed to reduce binge drinking and began implementation at the beginning of Grade 8 at the same time point when the risk factors were assessed (i.e., baseline).

Grade 9 binge drinking was randomly created as an outcome variable assessed 1 year after adolescents in the treatment group participated in the hypothetical peer-based intervention program. The outcome was simulated as follows. For adolescents in the control group, rates of Grade 9 binge drinking were simulated to be approximately 10% for adolescents in latent subgroups not characterized by peer-level risk and 50% for all others. (These rates were derived from reported rates of substance use from these participants at Wave 2.) For adolescents in the treatment group, rates of Grade 9 binge drinking among those in latent subgroups not characterized by peer-level risk were also simulated to be approximately 10%. Adolescents in a latent subgroup characterized by the intersection of risks across peer, household, and neighborhood levels were simulated to show no beneficial effects of the program; that is, their rate of binge drinking was simulated to be approximately 50% (just as in the control group). Data were simulated such that those in a latent subgroup characterized by peer and household risks, without neighborhood risks, exhibited about a 30% relative risk reduction due to the intervention, with approximately 35% (as opposed to 50% in the control group) engaging in Grade 9 binge drinking. Those in a latent subgroup characterized by only peer-level risk were simulated to exhibit about a 60% risk reduction due to the intervention, with approximately 20% (as opposed to 50% in the control group) engaging in binge drinking.

Examining Differential Treatment Effects—Finally, by augmenting the empirical data with the simulated treatment and outcome variables described above, differential treatment effects were assessed using two different approaches: 1) a classify-analyze approach involving logistic regression, and 2) a model-based approach to LCA with a distal outcome based on a reparameterization of the LCA with covariates model that incorporates the simulated treatment group and outcome directly in the LCA model. For the classify-analyze approach, we obtained each individual's posterior probability of membership in each latent

class based on the model selected above. Individuals were then assigned to the latent class corresponding to their maximum probability (i.e., the "classify" part of this approach). Finally, a logistic model was specified to predict the log-odds of Grade 9 binge drinking from treatment group, assigned risk subgroup, and the interaction between treatment group and assigned risk subgroup (i.e., the "analyze" part of this approach). This is representative of a traditional regression/ANOVA approach to testing for differential treatment effects across subgroups, but in this case the subgroups were derived from a latent class variable.

In the model-based approach, treatment group was incorporated in the model as a grouping variable so that baseline differences between the control and treatment groups in the prevalence of the risk subgroups could be examined. A formal hypothesis test of the equivalence of measurement (i.e., equal item-response probabilities for all item-latent class combinations) across treatment and control groups was conducted before group differences in the prevalences were examined. Next, the outcome variable was included as a covariate so that the associations between latent class membership and Grade 9 binge drinking could be estimated. Treatment group was included as a grouping variable in these models in order to allow the associations between risk subgroups and Grade 9 binge drinking to vary across treatment conditions (put another way, to allow the treatment effect to differ across latent risk subgroups). From this model, we obtained the logistic regression coefficients for adolescents in each treatment condition, and then used those parameter estimates, along with the known marginal distributions of the outcome in each treatment condition, to estimate the probability of Grade 9 binge drinking conditional on treatment condition and latent class membership.

All LCA models were estimated using PROC LCA Version 1.2.5 (Lanza et al. 2010a), a SAS procedure for conducting LCA. PROC LCA is available for download free of charge at http://methodology.psu.edu/.

Descriptive Statistics

Table 1 shows the proportion of adolescents in each (simulated) treatment condition who were exposed to each contextual risk factor. A similar proportion of adolescents in each group were exposed to all six risk factors ($\chi^2 < 4$ for 1 *df, p*>.05 for each risk factor). Rates of exposure ranged from 21.0% to 42.2%. The table also shows the proportion of individuals in the control and treatment groups who reported (simulated) Grade 9 binge drinking. The rates differed significantly across the two groups (27.8% in the control group and 23.4% in the treatment group; $\chi^2 = 4.7$ for 1 df, *p*=.03), suggesting a modest intervention effect overall (i.e., regardless of risk exposure).

Model Selection

Models with one through six latent classes were compared in order to select a model of multiple risks. The solution involving seven latent classes could not be sufficiently identified with these data (i.e., multiple sets of starting values did not suggest a single maximum likelihood solution). Table 2 shows the degrees of freedom and G^2 test statistic for each model, as well as the *p*-value showing the significance level of the test statistic. Models with one through five latent classes had significant test statistics (*p*<.0001 for one through four classes, *p*=.03 for five classes), suggesting that the latent classes in these models may not adequately account for heterogeneity in risk exposure across individuals in this sample. In addition, Table 2 shows values of the information criteria. The AIC suggested that the six-class model was slightly superior (AIC=112.8 for five latent classes v. AIC=110.0 for six latent classes), while the BIC, CAIC, and aBIC more clearly suggested the five-class model (BIC=301.4 for five latent classes v. BIC=337.6 for six latent classes; CAIC=335.4 for five latent classes v. CAIC=378.6 for six latent classes; aBIC=193.4 for five latent classes v.

aBIC=207.3 for six latent classes). A careful examination of both the five- and six-class model solutions led us to select the five-class model because it was more easily identified, had greater parsimony, and its parameter estimates presented a solution with a logical substantive interpretation.

Five Latent Subgroups

Each latent class corresponds to an underlying subgroup of adolescents characterized by a particular pattern of risk exposure; we will refer to these latent classes as risk subgroups. The parameter estimates depicted in Fig. 1 provide the necessary information for interpreting and labeling each risk subgroup. For example, the first latent class is characterized by a low probability of exposure to each of the six risk factors; thus we labeled this subgroup *Low Risk*. The second latent class is comprised of individuals who are likely to report peer cigarette use and peer alcohol use (.662 and .746, respectively), but are not likely to be exposed to any household or neighborhood risk factor as defined in this study. We labeled this subgroup *Peer Risk*. Youth in the third risk class were likely to report household poverty (.744) and resided in a neighborhood marked by high unemployment and high rates of poverty (.684 and .738, respectively; note that around half of adolescents in this latent class lived in a single-parent household). We labeled this subgroup *Economic Risk*. The fourth latent class was labeled *Household & Peer Risk* due to the high probability of adolescents in this subgroup reporting all four household and peer risk factors. This subgroup is characterized by the almost certain absence of neighborhood risk factors. Finally, the fifth latent class is characterized by a high probability of reporting exposure to each of the six risk factors; we labeled this subgroup *Multi-Contextual Risk*.

Measurement invariance held across (simulated) treatment groups, indicating that the same measurement model of risk applied to adolescents in the control and treatment groups (change in G^2=24.2, df=30, p=.76). Figure 2 shows the class membership probabilities for each group. These parameter estimates show the distribution of adolescents in each treatment group across the five risk subgroups. Risk exposure did not vary across treatment conditions (an overall test of group differences in class membership probabilities yielded a change in G^2=2.8, df=4, p=.60). Thirty-one percent of each treatment group were in the Low Risk subgroup, 28% of the control and 30% of the treatment group were in the Peer Risk subgroup, 20% of both the control and treatment group were in the Economic Risk subgroup, 12% of both the control and treatment group were in the Household and Peer Risk subgroup, and 8% of the control group and 6% of the treatment group were in the Multi-Contextual Risk subgroup.

Differential Treatment Effects Across Risk Subgroups

Classify-Analyze Approach—First, individuals were classified in the subgroup that corresponded to their maximum posterior probability. Average posterior probabilities were .84 for Low Risk, .86 for Peer Risk, .87 for Economic Risk, .90 for Household and Peer Risk, and .80 for Multi-Contextual Risk, suggesting low classification error. A logistic regression model was specified to predict the log-odds of (simulated) Grade 9 binge drinking from treatment group, assigned risk subgroup (with Low Risk as the reference group), and the interaction between treatment group and assigned risk subgroup (see Table 3). Results showed that the treatment effect was not significant for the Low Risk group (β=0.52, SE=0.27, p=.06), but that the effect was significantly different from this reference group for adolescents in the Peer Risk subgroup (β=−1.37, SE=0.32, p<.0001). In other words, the intervention was more effective at reducing Grade 9 binge drinking for those adolescents who had friends who used substances during Grade 8, but were not exposed to risk in the household or neighborhood. In order to interpret this effect for adolescents in the Peer Risk subgroup, the logistic regression coefficients can be transformed to odds ratios.

Members of the Peer Risk subgroup assigned to the treatment group were $e^{(0.52-1.37)} = 0.43$ times as likely to report Grade 9 binge drinking compared to their control group counterpart; thus, the program resulted in a risk reduction of more than 50%.

Model-Based Approach—In addition, in order to better understand these associations, Fig. 3 presents the estimated probability of treatment and control participants in each risk subgroup engaging in binge drinking at Grade 9. These probabilities were obtained from the beta coefficient estimates in the LCA model that incorporated the treatment assignment as a grouping variable and Grade 9 binge drinking as a covariate. An overall test of differential treatment effects across risk subgroups provided evidence for important differences (change in twice the log-likelihood value=24.7, *df*=4, *p*=.0001). As described above, treatment effects were greatest for those adolescents in the Peer Risk subgroup (52% of the control group reported Grade 9 binge drinking, compared to just 7% of the treatment group). In addition, a modest reduction was found for the Household & Peer Risk subgroup (54% of the control group, compared to 32% of the treatment group).

Discussion

The prevention and treatment of drug abuse and related societal problems rests largely on the ability of scientists to understand risk factors that can inform intervention efforts. An examination of the complex interplay of multiple risk factors can provide a more thorough and meaningful picture of individuals' environment than that provided by studying risk factors in isolation (Cicchetti and Rogosch 1996). Modeling the intersection of risks by organizing similar individuals into subgroups is conceptually appealing, as this approach lends itself to thinking about individuals more holistically (Bergman and Magnusson 1997; von Eye and Bergman 2003). It is well documented that exposure to additional risk factors corresponds to worse outcomes (e.g., Gerard and Buehler 2004; Luthar 1993; Rutter 1979; Sameroff et al. 1993). However, in much of this previous work risk factors were treated as interchangeable; thus, this work did not provide information about which specific combinations of risk corresponded to poor outcomes. In the current study, because we were able to describe which risk factors characterized the subgroups, we were able to draw conclusions about how multiple risks were organized within individuals.

Finite mixture models provide a statistically sophisticated framework for identifying subgroups of individuals that are not directly observable. In recent years the application of mixture models in intervention research has been highly productive, particularly as applied to substance use research. For example, this framework has been used to identify common patterns of substance use behavior (e.g., Agrawal et al. 2007; Kessler et al. 2005) and common trajectories of use over time (e.g., Chassin et al. 2002; Chung et al. 2004; Colder et al. 2002). This modeling framework also can be used to address many of the statistical challenges posed by more traditional subgroup analysis by modeling the subgroups as a latent categorical variable. These challenges include an inflated Type I error rate, issues of statistical power due to data sparseness, and the fact that traditional subgroup analysis may miss the detection of important subgroups defined by the higher-order intersection of subgroup variables. A latent subgroups perspective can provide important new information about key underlying population subgroups, which may illuminate subgroups that can be the target of tailored prevention and treatment programs and/or that may respond differently to a treatment program.

Several modern statistical programs are available to prevention scientists for conducting latent class analysis and related mixture models. These include the freely-available PROC LCA (Lanza et al. 2010a) for conducting LCA within the SAS environment, and the stand-

alone programs Latent Gold (Vermunt and Magidson 2005), PANMARK (Van de Pol et al. 1996), and Mplus (Muthén and Muthén 1998–2007).

In the empirical study reported above, LCA allowed us to identify a finite set of latent subgroups, and to examine how treatment effects on Grade 9 binge drinking varied according to membership in these latent subgroups. In the latent class model, the intersection of six risk factors (yielding $2^6=64$ possible response patterns) was organized into five straightforward, meaningful subgroups. Approximately one-third of males and females belonged to the Low Risk subgroup and another one-third to the Peer Risk subgroup, with the remaining one-third distributed across the subgroups that conferred the intersection of risks across two or more ecological levels: the Economic Risk subgroup, the Household & Peer Risk subgroup, and the Multi-Contextual Risk subgroup.

Table 3 and Fig. 3 demonstrate, based on simulated outcome data, how this approach could be applied to intervention data in order to obtain tangible evidence of how future intervention resources might best be targeted based on the set of characteristics used in the latent class model. Based on these results, for example, we found that among adolescents in the Peer Risk subgroup, the hypothetical intervention program mitigated risk for Grade 9 binge drinking. Among adolescents in the Household and Peer Risk and the Multi-Contextual Risk subgroups, however, the program had no effect. This indicates that the peer-based program hypothesized in this study should be implemented in lower-risk neighborhoods, but would not be expected to have an impact in higher-risk neighborhoods. These results suggest the need for an additional intervention component, or a program with a different contextual focus, in higher-risk neighborhoods.

As demonstrated in the present empirical example, prevention scientists interested in evaluating differential effects of a randomized prevention or treatment program can use this framework to organize individuals according to an entire set of characteristics (including both individual and contextual factors). Once the latent subgroups have been identified, several techniques can be implemented to explore differential treatment effects. The classify-analyze approach described here is perhaps the most simple one, but results about treatment effects will be biased to the extent that there is classification error. The amount of error is reflected in the degree to which each individual's posterior probabilities diverge from values of zero and one. When the treatment response is coded as a categorical variable, as was the case in this empirical demonstration, the probability of the event (in this case, simulated Grade 9 binge drinking) can be calculated for each latent subgroup using 1) the logistic regression estimates produced by including binge drinking as a covariate and 2) the marginal probabilities of the treatment response. Although the model-based approach demonstrated above (see Fig. 3) is not typically used in practice, it is always preferable to any classify-analyze approach. The model-based approach estimates the actual probability of an outcome conditional on latent class membership, whereas all classify-analyze approaches represent a computationally convenient way to approximate this result. The Excel-based LCA probability calculator, available at http://methodology.psu.edu/, will allow prevention scientists to adopt a model-based approach to LCA with a distal outcome in their own work.

When treatment response is continuous, however, obtaining mean differences in the treatment effect across the latent subgroups is more complicated. LCA is a probabilistic model, where each individual has a (typically nonzero) probability of membership in each latent subgroup. In other words, the true subgroup membership of each individual is unknown. Thus, there is no straightforward model-based approach to testing for mean differences across subgroups on an outcome variable. Although several different classify-analyze approaches have been proposed, including single or multiple pseudoclass draws based on posterior probabilities (Wang et al. 2005) and multiple imputation of the latent

class variable using a Bayesian approach (Lanza et al. 2005), future methodological work in this area is needed.

Limitations of Using LCA for Latent Subgroups Analysis

As with any statistical approach that uses categorical variables, careful consideration must be given to recoding continuous indicators in a way that results in meaningful categories and does not result in a substantial loss of information. It is possible to use continuous variables to measure the latent subgroups by conducting latent profile analysis, or even to use a mix of continuous and categorical variables. In the current study, many of the latent class indicators were categorical by nature (e.g., single-parent home, exposure to peer cigarette use, exposure to peer alcohol use). We did specify cutoffs for the neighborhood risks of greater or equal to 10% unemployment rate and greater or equal to 50% poverty rate in the census block. In some cases a sensitivity analysis may be helpful, where the effect of using different cutoffs is explored in the resulting latent class model.

In any latent class model, the issue of reification is of great importance. More than with traditional analytic approaches such as regression analysis, with LCA it can be easy to conclude that the set of latent classes identified in an analysis represent the actual types of individuals in the population. Instead, the latent classes provide a useful heuristic for representing heterogeneity across the dimensions included in the model. Implications of model misspecification may be important substantively when examining differential treatment effects across the latent classes, particularly if too few latent classes have been identified. In this case, a unique latent subgroup may get overlooked, precluding an examination of the subgroup-specific treatment effect. Cross-validation, where a training data set is used to estimate a model and then the fit of a model using those estimates and a validation data set is assessed, is one method that has been proposed for further confirming the validity of a particular solution (Collins et al. 1994). A related issue is that little is known about the exact effect of sample size on the ability to identify the set of underlying latent classes; this is an important area of future research.

Conclusions

By using LCA, a finite set of latent subgroups characterized by the intersection of numerous characteristics was identified in the present analysis. This LCA perspective can provide important information about how programs may be targeted for or tailored to different population subgroups that are expected to show the strongest response. This study demonstrated how multiple characteristics can be used to identify meaningful latent subgroups, and then subsequently be used to examine differential treatment effects. Overall, this study demonstrated the promise that a latent subgroups perspective holds for advancing subgroup analyses in an effort to better inform prevention and treatment sciences.

Acknowledgments

References

Agrawal A, Lynskey MT, Madden PA, Bucholz KK, Heath AC. A latent class analysis of illicit drug abuse/dependence: Results from the National Epidemiological Survey on Alcohol and Related Conditions. Addiction. 2007; 102:94–104. [PubMed: 17207127]

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

Aiken, LS.; West, SG. Multiple regression: Testing and interpreting interactions. Thousand Oaks, CA: Sage; 1991.

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

Anderson KG, Ramo DE, Cummins KM, Brown SA. Alcohol and drug involvement after adolescent treatment and functioning during emerging adulthood. Drug and Alcohol Dependence. 2010; 107:171–181. [PubMed: 19926231]

Arthur MW, Hawkins JD, Pollard J, Catalano RF, Baglioni AJ. Measuring risk and protective factors for substance use, delinquency, and other adolescent problem behaviors: The Communities that Care Youth Survey. Evaluation Review. 2002; 26:575–601. [PubMed: 12465571]

Baucom BR, Atkins DC, Simpson LE, Christensen A. Prediction of response to treatment in a randomized clinical trial of couple therapy: A 2-year follow-up. Journal of Consulting and Clinical Psychology. 2009; 77:160–173. [PubMed: 19170462]

Bergman LR, Magnusson D. A person-oriented approach in research on developmental psychopathology. Development and Psychopathology. 1997; 9:291–319. [PubMed: 9201446]

Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987; 52:345–370.

Brooks-Gunn, J.; Duncan, GJ.; Maritato, N. Poor families, poor outcomes: The well-being of children and youth. In: Duncan, GJ.; Brooks-Gunn, J., editors. Consequences of growing up poor. New York: Russell Sage; 1997. p. 1-17.

Catalano, RF.; Hawkins, JD. The social development model: A theory of anti-social behavior. In: Hawkins, JD., editor. Delinquency and crime: Current theories. New York: Cambridge University Press; 1996. p. 149-197.

Chassin L, Pitts SC, Prost J. Binge drinking trajectories from adolescence to emerging adulthood in a high-risk sample: Predictors and substance abuse outcomes. Journal of Consulting and Clinical Psychology. 2002; 70:67–78. [PubMed: 11860058]

Chung T, Maisto SA, Cornelius JR, Marti CS. Adolescents' alcohol and drug use trajectories in the year following treatment. Journal of Studies on Alcohol. 2004; 65:105–114. [PubMed: 15000509]

Cicchetti D, Rogosch FA. Equifinality and multifinality in developmental psychopathology. Development and Psychopathology. 1996; 8:597–600.

Coffman DL, Patrick ME, Palen LA, Rhoades BL, Ventura AK. Why do high school seniors drink? Implications for a targeted approach to intervention. Prevention Science. 2007; 8:241–248. [PubMed: 17963040]

Coie JD, Watt NF, West SG, Hawkins JD, Asarnow JR, Markman HJ, et al. The science of prevention: A conceptual framework and some directions for a national research program. American Psychologist. 1993; 48:1013–1022. [PubMed: 8256874]

Colder CR, Campbell RT, Ruel E, Richardson JL, Flay BR. A finite mixture model of growth trajectories of adolescent alcohol use: Predictors and consequences. Journal of Consulting and Clinical Psychology. 2002; 70:976–985. [PubMed: 12182281]

Collins, LM.; Lanza, ST. Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences. New York: Wiley; 2010.

Collins LM, Fidler PL, Wugalter SE, Long JD. Goodness-of-fit testing for latent class models. Multivariate Behavioral Research. 1993; 28:375–389.

Collins LM, Graham JW, Long JD, Hansen WB. Crossvalidation of latent class models of early substance use onset. Multivariate Behavioral Research. 1994; 29:165–183.

Collins LM, Murphy SA, Bierman K. A conceptual framework for adaptive preventive interventions. Prevention Science. 2004; 3:185–196. [PubMed: 15470938]

Conduct Problems Prevention Research Group. A developmental and clinical model for the prevention of conduct disorders: The FAST Track Program. Development and Psychopathology. 1992; 4:509–527.

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

Elkin I, Gibbons RD, Shea MT, Sotsky SM, Watkins JT, Pilkonis PA, et al. Initial severity and differential treatment outcome in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. Journal of Consulting and Clinical Psychology. 1995; 63:841–847. [PubMed: 7593878]

Everitt, BS.; Hand, DJ. Finite mixture distributions. London: Chapman and Hall; 1981.

Fairchild AJ, MacKinnon DP. A general model for testing mediation and moderation effects. Prevention Science. 2009; 10:87–99. [PubMed: 19003535]

Gerard JM, Buehler C. Cumulative environmental risk and youth maladjustment: The role of youth attributes. Child Development. 2004; 75:1832–1849. [PubMed: 15566383]

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

Harris, KM.; Halpern, CT.; Whitsel, E.; Hussey, J.; Tabor, J.; Entzel, P.; Udry, JR. The National Longitudinal Study of Adolescent Health: Research Design [WWW document]. 2009. URL: http://www.cpc.unc.edu/projects/addhealth/design

Hawkins JD, Catalano RF, Miller JY. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. Psychological Bulletin. 1992; 112:64–105. [PubMed: 1529040]

Kessler RC, Chiu WT, Demler O, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry. 2005; 62:617–627. [PubMed: 15939839]

Komro KA, Tobler AL, Maldonado-Molina MM, Perry CL. Effects of alcohol use initiation patterns on high-risk behaviors among urban, low-income, young adolescents. Prevention Science. 2010; 11:14–23. [PubMed: 19639410]

Langeheine R, Pannekoek J, van de Pol F. Bootstrapping goodness-of-fit measures in categorical data analysis. Sociological Methods & Research. 1996; 24:492–516.

Lanza ST, Collins LM, Schafer JL, Flaherty BP. Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. Psychological Methods. 2005; 10:84–100. [PubMed: 15810870]

Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS procedure for latent class analysis. Structural Equation Modeling. 2007; 14:671–694. PMCID: PMC2785099. [PubMed: 19953201]

Lanza, ST.; Lemmon, DR.; Dziak, JJ.; Huang, L.; Schafer, JL.; Collins, LM. PROC LCA & PROC LTA user's guide version 1.2.5 beta. University Park, PA: The Methodology Center, Penn State; 2010.

Lanza ST, Rhoades BL, Nix RL, Greenberg MT, the Conduct Problems Prevention Research Group. Modeling the interplay of multilevel risk factors for future academic and behavior problems: A person-centered approach. Development and Psycholopathology. 2010; 22:313–335.

Laska MN, Pasch KE, Lust K, Story M, Ehlinger E. Latent class analysis of lifestyle characteristics and health risk behaviors among college youth. Prevention Science. 2009; 10:376–386. [PubMed: 19499339]

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

Loeber R. Development and risk factors of juvenile antisocial behavior and delinquency. Clinical Psychology Review. 1990; 10:1–41.

Luthar SS. Annotation: Methodological and conceptual issues in research on childhood resilience. Journal of Child Psychology and Psychiatry & Allied Disciplines. 1993; 34:441–453.

MacKinnon, DP. Introduction to statistical mediation analysis. New York: Lawrence Erlbaum Associates; 2009.

McLachlan, G.; Peel, D. Finite mixture models. New York: Wiley; 2000.

Muthén, LK.; Muthén, BO. Mplus user's guide. 5. Los Angeles, CA: Authors; 1998–2007.

Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling. 2007; 14:535–569.

Ondersma SJ, Winhusen T, Erickson SJ, Stine SM, Wang Y. Motivation enahancement therapy with pregnant substance-abusing women: Does baseline motivation moderate efficacy? Drug and Alcohol Dependence. 2009; 101:74–79. [PubMed: 19101099]

Oxford ML, Gilchrist LD, Morrison DM, Gillmore MR, Lohr MJ, Lewis SM. Alcohol use among adolescent mothers: Heterogeneity in growth curves, predictors, and outcomes of alcohol use over time. Prevention Science. 2003; 4:15–26. [PubMed: 12611416]

Rutter, M. Protective factors in children's responses to stress and disadvantage. In: Kent, MW.; Rolf, JE., editors. Primary prevention of psychopathology: Vol 3 Social competence in children. Hanover, NH: University of New England Press; 1979. p. 49-74.

Sameroff AJ, Seifer R, Baldwin CP, Baldwin A. Stability of intelligence from preschool to adolescence: The influence of social and family risk factors. Child Development. 1993; 64:80–97. [PubMed: 8436039]

Scheier LM, Abdallah AB, Inciardi JA, Copeland J, Cottler LB. Tri-city study of Ecstasy use problems: A latent class analysis. Drug and Alcohol Dependence. 2008; 98:249–263. [PubMed: 18674872]

Schochet, PZ. Technical methods report: Guidelines for multiple testing in impact evaluations (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S Department of Education; 2008.

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

Sclove SL. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika. 1987; 52:333–343.

Shin SH, Hong HG, Hazen AL. Childhood sexual abuse and adolescent substance use: A latent class analysis. Drug and Alcohol Dependence. 2010; 109:226–235. [PubMed: 20197217]

Syvertsen AK, Cleveland MJ, Gayles JG, Tibbits MK, Faulk MT. Profiles of protection from substance use among adolescents. Prevention Science. 2010; 11:185–196. [PubMed: 19844790]

Titterington, D.; Smith, A.; Makov, U. Statistical analysis of finite mixture distributions. Chichester, UK: Wiley; 1985.

Van de Pol, F.; Langeheine, R.; De Jong, W. User's manual: A latent class program. The Netherlands: Voorburg; 1996.

Vermunt, JK.; Magidson, J. Latent GOLD 4.0 user's guide. Belmont, MA: Statistical Innovations Inc; 2005.

Von Eye A, Bergman LR. Research strategies in developmental psychopathology: Dimensional identity an the person-oriented approach. Development and Psychopathology. 2003; 15:553–580. [PubMed: 14582932]

Wang C-P, Brown CH, Bandeen-Roche K. 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. 2005; 100:1054–1076.
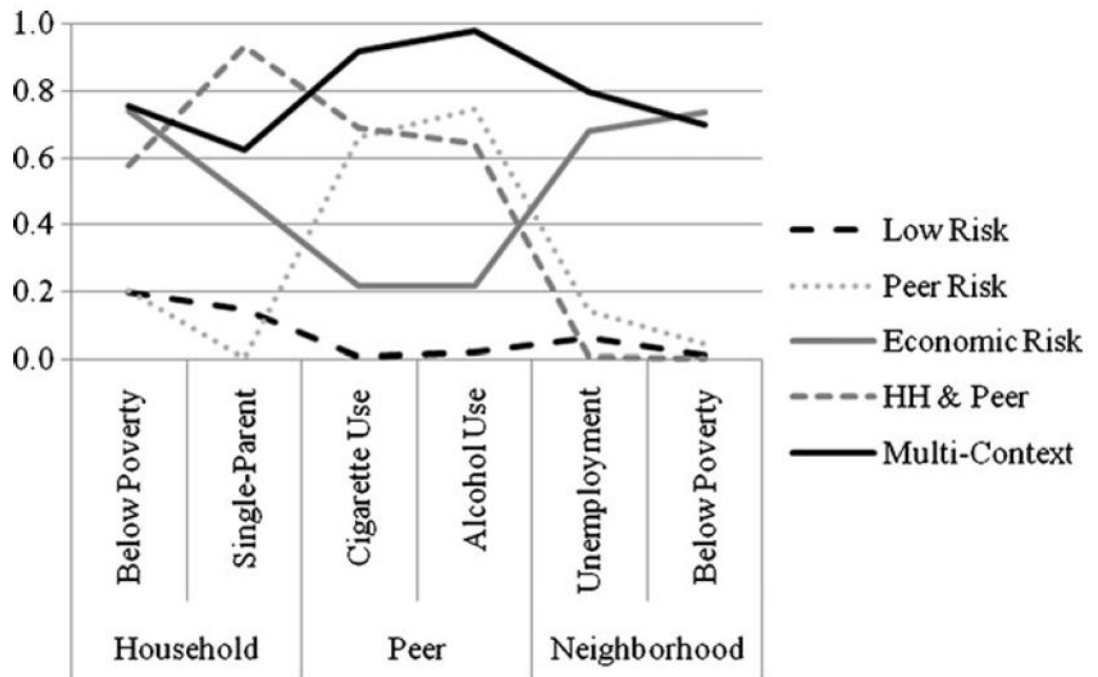
Fig. 1.
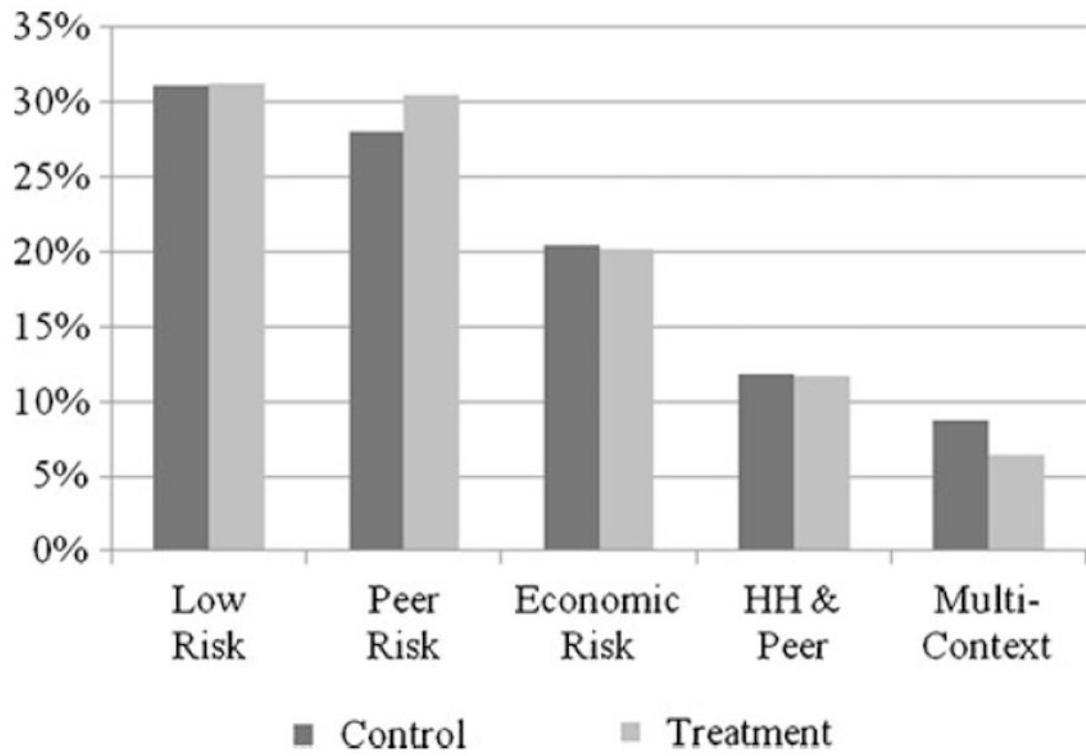Item-response probabilities for five-class model (probability of reporting risk factor given latent class)

Fig. 2.
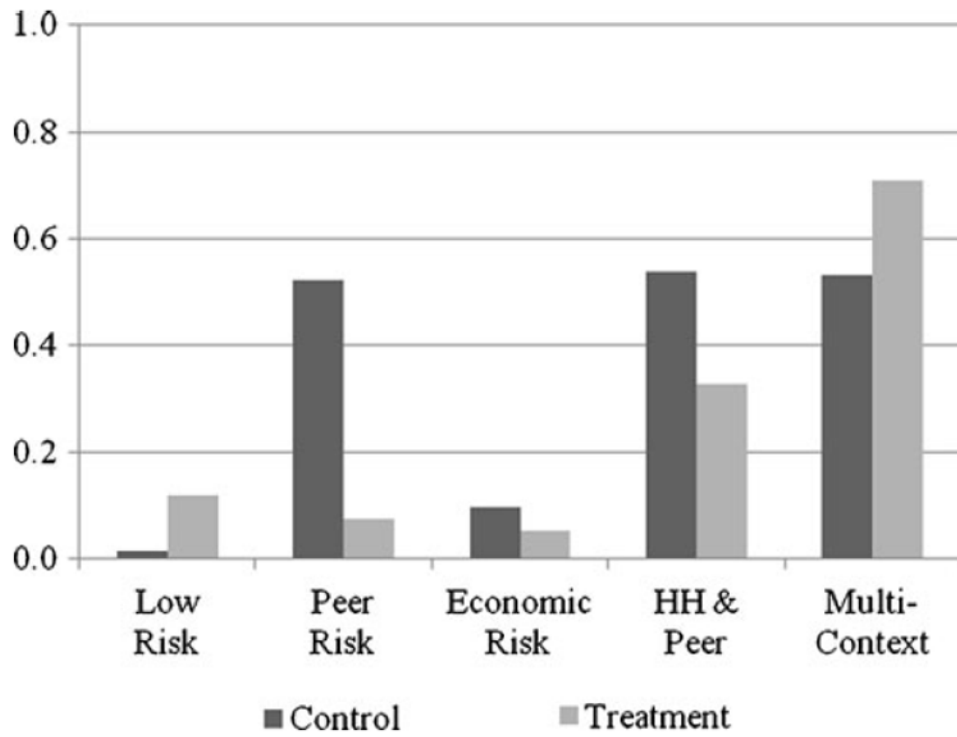Class membership probabilities for each treatment condition

Fig. 3.
Proportion of treatment and control participants in each risk subgroup reporting binge drinking at Grade 9

Table 1

Proportion of adolescents in treatment and control groups exposed to each risk factor and proportion reporting Grade 9 binge drinking

		% At-risk (*N* missing)	
	p-value[a]	Treatment[b] (*N*=927)	Control[b] (*N*=973)
Household Risks			
Below Poverty	.65	39.7 (196)	38.6 (190)
Single-Parent	.52	29.2 (85)	30.6 (88)
Peer Risks			
Cigarette Use	.85	38.7 (28)	39.1 (25)
Alcohol Use	.63	41.1 (34)	42.2 (30)
Neighborhood Risks			
Unemployment	.17	24.1 (28)	26.9 (33)
Below Poverty	.41	21.0 (8)	22.6 (11)
Outcome			
Grade 9 Binge Drinking[b]	.03	23.4 (0)	27.8 (0)

[a]Chi-square test of treatment group differences in proportion reporting risk factor

[b]Variable was simulated

Table 2

Indicators of fit for models with one through six latent classes

Number of classes	df	G^2	p-value	AIC	BIC	CAIC	aBIC
1	57	1267.2	<.0001	1279.2	1312.5	1318.5	1293.5
2	50	576.3	<.0001	602.3	674.5	687.5	633.2
3	43	201.3	<.0001	241.3	352.3	372.3	288.8
4	36	126.2	<.0001	180.2	330.0	357.0	244.2
5	29	44.8	.03	112.8	301.4	335.4	193.4
6	22	28.0	.17	110.0	337.6	378.6	207.3

Table 3

Differential treatment effects across latent subgroups based on classify-analyze approach

Predictor	B	SE B
Intercept	−2.54 ***	.21
Treatment	.52	.27
Risk Classes		
Peer Risk	2.67 ***	.24
Economic Risk	.32	.32
Household & Peer Risk	2.50 ***	.29
Multi-Contextual Risk	2.38 ***	.31
Interactions		
Treatment * Peer Risk	−1.37 ***	.32
Treatment * Economic Risk	−.91	.47
Treatment * Household & Peer Risk	−.54	.40
Treatment * Multi-Contextual Risk	−.06	.44

*
$p<.05$,

**
$p<.01$,

***
$p<.001$.

Assigned class sizes are Low Risk, N=669; Peer Risk, N=551; Economic Risk, N=362; Household & Peer Risk, N=181; Multi-Contextual Risk, N=137