Published in final edited form as:
J R Stat Soc Ser A Stat Soc. 2008 October ; 171(4): 877–897. doi:10.1111/j.1467-985X.2008.00544.x.

Locally dependent latent class models with covariates: an application to under-age drinking in the USA

Beth A. Reboussin, **Edward H. Ip**, and **Mark Wolfson**
Wake Forest University School of Medicine, Winston-Salem, USA

Summary

Under-age drinking is a long-standing public health problem in the USA and the identification of underage drinkers suffering alcohol-related problems has been difficult by using diagnostic criteria that were developed in adult populations. For this reason, it is important to characterize patterns of drinking in adolescents that are associated with alcohol-related problems. Latent class analysis is a statistical technique for explaining heterogeneity in individual response patterns in terms of a smaller number of classes. However, the latent class analysis assumption of local independence may not be appropriate when examining behavioural profiles and could have implications for statistical inference. In addition, if covariates are included in the model, non-differential measurement is also assumed. We propose a flexible set of models for local dependence and differential measurement that use easily interpretable odds ratio parameterizations while simultaneously fitting a marginal regression model for the latent class prevalences. Estimation is based on solving a set of second-order estimating equations. This approach requires only specification of the first two moments and allows for the choice of simple ‘working’ covariance structures. The method is illustrated by using data from a large-scale survey of under-age drinking. This new approach indicates the effectiveness of introducing local dependence and differential measurement into latent class models for selecting substantively interpretable models over more complex models that are deemed empirically superior.

Keywords

Differential measurement; Latent class; Local dependence; Marginal regression; Odds ratio; Second-order estimating equations

1. Introduction

Under-age drinking is a long-standing major public health problem in the USA with widespread individual and social consequences. Although under-age drinking decreased following changes in the minimum purchase age in the mid-1980s, prevalence rates have remained relatively stable and high over the last decade (Faden and Fay, 2004). On the basis of the 2006 National Survey on Drug Use and Health, there are an estimated 10.8 million under-age drinkers in the USA (Substance Abuse and Mental Health Service Administration, 2007). Of these, nearly 7.2 million are considered binge drinkers (five or more drinks on one occasion), and more than 2 million are classified as heavy drinkers. As a result, the US Surgeon General's Office issued its first ‘Call to action’ against under-age drinking in March 2007 (US Department of Health and Social Services, 2007).

Address for correspondence: Beth A. Reboussin, Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA. E-mail: brebouss@wfubmc.edu.

Given the continued pervasiveness of under-age drinking in the USA, recent efforts have focused on understanding the nature of under-age problem drinking rather than abstinence. The term ‘problem drinking’ is most often used by researchers to describe individuals who are not alcohol dependent but who consume enough alcohol to be at risk for a variety of alcohol-related problems (Institute of Medicine, 1990). Evidence suggests, however, that one-dimensional consumption measures that are commonly used to study problem drinking in adult populations fail to identify under-age drinkers suffering alcohol-related problems (Ellickson *et al.*, 1996; Stewart and Power, 2002; Townshend and Duka, 2002). Studies have documented that, in contrast with adults, adolescents tend to be infrequent but heavy drinkers (Deas *et al.*, 2000; Wechsler *et al.*, 2000). Further, many youths who experience harmful consequences as a result of drinking do not meet the clinical definition of dependence (Chung *et al.*, 2000, 2002). For these reasons, it is important to identify patterns of drinking in adolescents that are associated with alcohol-related problems. In this paper, we use latent class analysis (LCA) to identify subtypes of under-age drinkers on the basis of their observed patterns of drinking. LCA is an empirically based statistical approach for explaining the heterogeneity in response profiles in terms of underlying latent classes. This is in contrast with latent trait models in which heterogeneity is described in terms of an underlying continuous trait (Lord and Novick, 1968). In the LCA framework, associations between drinking behaviours are assumed to be due to their relationship to underlying drinking subtypes (classes). In a statistical sense, this means that behaviours are *locally independent* within a drinking class.

Recently, the LCA perspective has found a growing number of important applications not only in the social sciences (Becker and Yang, 1998; Roeder *et al.*, 1999; von Davier and Yamamoto, 2004; Vermunt, 2004) but also in behavioural (Hudziak *et al.*, 1998; Storr *et al.*, 2004; Carlson *et al.*, 2005; Crum *et al.*, 2005; Ferdinand *et al.*, 2005; Reboussin *et al.*, 2006b; Reboussin and Anthony, 2006) and biomedical research (Garrett *et al.*, 2002; Lin *et al.*, 2002; Patterson *et al.*, 2002; Von Korff and Miglioretti, 2005; Strauss *et al.*, 2006). This broadening interest has led to the development of latent class regression models that incorporate covariates as predictors of class membership (Dayton and Macready, 1988; Van der Heijden *et al.*, 1996; Bandeen-Roche *et al.*, 1997; Reboussin and Anthony, 2001; Huang and Bandeen-Roche, 2004). The LCA regression framework introduces a second type of local independence assumption, which is called *non-differential measurement*, that says that within a latent class observed responses and covariates are independent. Although primary interest is most often focused on the relationship between covariates and latent classes, there are situations in which covariates might influence the measurement process directly. For example, males endorse certain alcohol-related problems (e.g. social and legal) at higher rates than females (Perkins, 1992; Wechsler *et al.*, 1995; Humara and Sherman, 2004) despite the fact that both might be considered problem drinkers.

An important issue in LCA is interpretation of the resultant latent classes. Sometimes a statistical analysis yields models that are not immediately interpretable to experts in the field, even though the models that are selected by LCA may be optimal according to criteria that are based on goodness-of-fit statistics. Although the local independence assumption is instrumental in simplifying the representation of potentially large combinations of responses to a smaller number of consistent patterns, it can only be regarded as an idealization. At its best, local independence provides a useful approximation of reality; at its worst, it gives an overly simplified and biased picture of the true state of the subject at hand. For example, when searching for distinct syndromes, for each pair of behaviours it seems unlikely that the occurrence of any one behaviour will be unrelated to the other for any diagnostic category. For this reason, violation of the local independence assumption is expected in applications involving the search for distinct diagnostic categories (e.g. problem drinker) given the presence or absence of several symptoms or behaviours. For example, when searching for problem drinking subtypes symptom questions like 'passing out as a result of drinking' and 'forgetting what happened as a result of drinking' might always be associated.

The local dependence that may be present in survey instruments has implications for statistical inference. First, it leads to biased estimates, even when the number of latent classes is correctly estimated, or known *a priori* (Vacek, 1985; Torrrance-Rynard and Walter, 1997). Second, because we can always achieve local independence by increasing the number of classes (Suppes and Zanotti, 1981), it can yield spurious classes that are not immediately interpretable to experts in the field or truly present at the taxonic level even though the models that are selected by LCA may be optimal according to criteria that are based on goodness-of-fit statistics. Consequently, the researcher is confronted with a choice between a substantively more meaningful model with the potential for model misspecification and an empirically superior model that may be more difficult to interpret.

The idea of relaxing the local independence assumption is not new and dates back to the work of Harper (1972), Hagenaars (1988) and Espeland and Handelman (1989) for evaluating the accuracy of diagnostic tests. Using a log-linear parameterization of the latent class model (Haberman, 1979), they included direct effects between pairs of tests to allow for local dependence. Banfield and Raftery (1993) applied a similar idea in mixture modelling, of which LCA can be considered a special case. Becker and Yang (1998) proposed an LCA for low dimensional cross-classified tables in which the local dependence is specified by a marginal log-linear model. Qu *et al.* (1996) proposed a random-effects model that incorporates local dependence between tests through a single normally distributed random effect. This approach induces an equal and positive correlation between tests. Although this model might be plausible for diagnostic testing when tests, e.g. slides or X-rays, are evaluated by multiple raters or when all tests are based on a similar biological process, it might not accurately reflect the dependence structure of behavioural profiles. For example, when searching for subtypes of behavioural or psychiatric disorders that are represented by distinct patterns of symptoms, some symptoms may be more closely related than others and their correlation may depend on the severity of the disorder. In addition, some correlations might be expected to be negative, such as item pairs for positive and negative moods. Both the random-effects and the log-linear approaches provide conditional interpretations of the relationships between the responses and latent class (e.g. conditional log-odds given random effects in the former and conditional log-odds given all other responses in the latter). If differential measurement is incorporated in the model as well, interpretation of the direct effects of covariates on responses is also conditional on the random effects. More recently, Uebersax (1999) proposed a probit LCA that is similar in spirit to the probit regression model (Hedeker and Gibbons, 1994). The dependence structure within a class is captured via the covariance matrix of a multivariate distribution, which is used as an underlying variable for generating dichotomized or ordinal outcomes. However, the method cannot easily extend to tests that contain a large number of items.

In this paper, we propose a locally dependent latent class model with covariates that uses an easily interpretable pairwise odds ratio to quantify local dependence between binary-coded (yes–no) items. We extend the marginal latent class regression model of Reboussin and Anthony (2001) by jointly estimating three separate models:

- **a.** a nominal logistic regression model for the latent class prevalences conditional on covariates,
- **b.** a model for the log-odds-ratio between pairs of observed responses within a class and
- **c.** a logistic regression model for the probability of a positive response conditional on class and covariates (called response probabilities).

Models (b) and (c) allow for relaxation of the local independence and non-differential measurement assumptions respectively. Existing approaches account for local dependence and differential measurement in the same model, which induces a conditional interpretation of the regression coefficients for the response probabilities. By fitting two separate models, the approach proposed permits flexible modelling of local dependence while providing direct, population-averaged regression parameters for the response probabilities that are not immediately obtainable from previously developed locally dependent latent class models. Parameter estimation involves solving a set of estimating equations. It requires only specification of the first two moments and allows for the choice of simple working covariance structures for the observed responses to reduce the computational burden. This type of approach is widely used for the analysis of longitudinal data in the generalized linear model framework (Zeger and Liang, 1986). Most often it is applied when the correlation between repeated measures is a nuisance but Liang *et al.* (1992) developed a set of second-order estimating equations when modelling the dependence structure is of interest. It was modified for LCA models and shown to perform well and to be consistent with prior studies of this estimation approach in the generalized linear model framework (Reboussin *et al.*, 1999, 2006a; Reboussin and Anthony, 2001).

2. Locally dependent latent class marginal regression model

2.1. A traditional latent class model

Let y_{ij} (j = 1,..., p) be a binary response for the ith subject where $y_{ij} = 1$ if item j is a 'yes' or 'present' response and $y_{ij} = 0$ otherwise. We refer to $\mathbf{y}_i = (y_{i1},...,y_{ip})'$ as the response profile. The hallmark of LCA is the assumption that responses are associated because of an underlying class structure. In a statistical sense, this means that within a class responses are independent. This is the axiom of *local independence* that forms the basis for the traditional LCA. Since latent class membership is not observed without error, this assumption is not verifiable; however, its adequacy as compared under various class assumptions will be discussed later. For estimation purposes, we assume that there are C latent classes indexed by $m = 1,..., C$. The latent (unobserved) class for individual i is denoted by η_i where $\pi_m = P(\eta_i = m)$ is the proportion of subjects in class m or the latent class prevalence.

The joint distribution of responses for an individual under the locally independent LCA is

$$P\left(Y_{i1}=y_{i1},Y_{i2}=y_{i2},\ldots,Y_{ip}=y_{ip}\right)=\sum_{m=1}^{C}\prod_{j=1}^{p}P\left(Y_{ij}|\eta_i=m\right)\ P\left(\eta_i=m\right).$$

This distribution is a mixture of product multinomial processes with mixing weights $\pi_m = P(\eta_i = m)$ subject to the constraint that $\Sigma_{m=1}^{C}\pi_m=1$. The first-order moment for a response j is given by

$$E\left[Y_{ij}\right]=\sum_{m=1}^{C}E\left[Y_{ij}|\eta_i=m\right]\pi_m=\sum_{m=1}^{C}P_{jm}\pi_m \tag{1}$$

where $p_{jm} = P(Y_{ij} = 1|\eta_i = m)$ is the probability that a subject in class m has a positive response on item j. The response probabilities p_{jm} aid in the interpretation of latent classes by characterizing the responses of individuals within a particular latent class. The second-order moment between two responses j and h is expressed as

$$E\left[Y_{ij}Y_{ih}\right]=\sum_{m=1}^{C}E\left[Y_{ij}Y_{ih}|\eta_i=m\right]\pi_m. \tag{2}$$

Under the local independence assumption, the second-order moments reduce to

$$E\left[Y_{ij}Y_{ih}\right]=\sum_{m=1}^{C}E\left[Y_{ij}|\eta_i=m\right]\,E\left[Y_{ih}|\eta_i=m\right]\pi_m=\sum_{m=1}^{C}p_{jm}\,p_{hm}\pi_m \tag{3}$$

whereas the first-order moments in equation (1) are not affected.

2.2. Regression model for local dependence

Let j and h ($j \neq h$) index two items from the response profile $\mathbf{y}_i$. To relax the local independence assumption between this item pair within a latent class, we propose to measure their residual association by using an odds ratio parameterization. Specifically, the pairwise within-class odds ratio between these items is defined by

$$\gamma_{ihm}=\frac{P\left(Y_{ij}=1,Y_{ih}=1|\eta_i=m\right)\,P\left(Y_{ij}=0,Y_{ih}=0|\eta_i=m\right)}{P\left(Y_{ij}=1,Y_{ih}=0|\eta_i=m\right)\,P\left(Y_{ij}=0,Y_{ih}=1|\eta_i=m\right)}$$

where γ_{jhm} = 1 indicates no residual association or local independence. We propose to model the pairwise odds ratio between items j and h within class m by

$$\log\left(\gamma_{jhm}\right)=g\left(\phi\right)$$

where the $\boldsymbol{\phi}$-parameters describe the pattern of association between items within a class to be estimated. For example, when evaluating the accuracy of a diagnostic test that is performed by multiple raters, we might assume a constant residual association for all possible pairs of raters S_1 and across all classes, i.e. an exchangeable correlation structure. This correlation structure is most similar to the random-effects model of Qu *et al.* (1996). It is given by

$$\log\left(\gamma_{jhm}\right)=\phi \qquad \forall\,(j,h)\in S_1, \qquad m=1,\ldots,C. \tag{4}$$

When searching for behavioural or psychiatric subtypes, symptoms are typically not exchangeable. A less restrictive model that specifies local dependences for a subset of symptom pairs $S_2 \subset S_1$ that are both symptom and class specific, i.e. an unstructured correlation structure, may be more appropriate in this situation, specifically

$$\log\left(\gamma_{jhm}\right)=\phi_{jhm} \qquad (j,h)\in S_2, \qquad m=1,\ldots,C. \tag{5}$$

and $\log(\gamma_{jhm}) = 0$ when $(j, h) \notin S_2$ (i.e. j and h locally independent). Models that are specific to the pairing of symptoms but constant across classes might also be considered, i.e. a class exchangeable model such as

$$\log\left(\gamma_{jhm}\right)=\phi_{jh} \qquad (j,h)\in S_2, \qquad \forall m=1,\ldots,C. \tag{6}$$

and $\log(\gamma_{jhm}) = 0$ when $(j, h) \notin S_2$.

Using the result of Dale (1986), the within-class second-order moments for item pairs (j, h) for which local dependence is specified can be written as

$$\begin{aligned}E\left[Y_{ij}Y_{ih}|\eta_i=m\right]=&\Big\{1-\left(p_{jm}+p_{hm}\right)\left\{1-\gamma_{jhm}(\phi)\right\}-\Big(\left[1-\left(p_{jm}+p_{hm}\right)\left\{1-\gamma_{jhm}(\phi)\right\}\right]^2\\&-4\left\{\gamma_{jhm}(\phi)-1\right\}\gamma_{jhm}(\phi)\,p_{jm}\,p_{hm}\Big)^{1/2}\Big\}\left[2\left\{\gamma_{jhm}(\phi)-1\right\}\right]^{-1}\end{aligned} \tag{7}$$

and substituted into equation (2). The second-order moments in equation (2) are now a function of the response probabilities $\mathbf{p}$, the latent class prevalences $\boldsymbol{\pi}$ and the local dependence parameters $\boldsymbol{\phi}$. For item pairs in which local independence is assumed, the second-order moments in equation (2) reduce to equation (3).

2.3. Regression model for latent class prevalences

In addition to observing a set of responses regarding drinking behaviours, we might also observe a set of q risk factors that are thought to be possible determining factors of class membership. Scientific interest then focuses on how the latent class prevalences π_m depend on suspected risk factors. We consider the baseline category logistic marginal regression model that was proposed by Reboussin and Anthony (2001) for the latent class prevalences. It is given by

$$\log\left(\frac{\pi_m}{\pi_1}\right)=\beta_m+\mathbf{z}_i'\zeta_m \qquad (8)$$

where $\pi_m = \mathrm{P}(\eta_i = m|\mathbf{z}_i)$, $m = 2,..., C$, and $\mathbf{z}_i = (z_{i1},..., z_{iq})$ is a $q \times 1$ covariate vector. In the absence of covariates,

$$\pi_m=\exp(\beta_m) \,/\, \left\{1+\sum_{l=2}^{C}\exp(\beta_l)\right\}$$

is the latent class prevalence for class m. The odds ratio relating the latent class prevalence to the covariate z_{ir} is given by $\exp(\zeta_{mr})$. It is the proportional change in the relative probability of membership in latent class m *versus* 1 for a 1-unit increase in the covariate z_{ir} holding all other covariates fixed. Because $\exp(\zeta_{mr})$ is a ratio of population frequencies, it is referred to as a population-averaged parameter.

2.4. Regression model for the response probabilities

When modelling the latent class prevalences as a function of covariates as in equation (8), the traditional LCA assumes that, within a class, responses and covariates are independent, i.e. the covariates do not affect the measurement process. Although the primary interest is often in the effect of covariates on the latent classes, we propose to relax the non-differential measurement assumption and allow covariates to affect the measurement process directly. By relaxing this condition, we can allow a female to be classified as a problem drinker, for example, despite her lower level of symptom endorsement for certain symptoms (e.g. social problems) compared with males within the same class.

To allow for differential measurement, we propose to model the response probabilities by using a logistic regression model. Assuming that a subset of the covariates $\mathbf{z}_i$ that are believed to influence latent class membership in equation (8) are also believed to influence the response probabilities directly, we consider the logistic regression model

$$\log\left(\frac{p_{jm}}{1-p_{jm}}\right)=\alpha_{jm}+\mathbf{x}_i'\lambda_{jm} \qquad (9)$$

where $p_{jm} = P(Y_{ij} = 1|\eta_i = m, \mathbf{x}_i)$ and $\mathbf{x}_i$ is a subset of covariates from equation (8).

3. Parameter estimation

3.1. Second-order estimating equations

We propose to solve a set of second-order estimating equations $U(\boldsymbol{\theta})$ for the parameters of interest $\boldsymbol{\theta}(\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$ that incorporate information from both the first- and the second-order moments of the observed response profile $\mathbf{y}_i$. This is unlike estimating equations for generalized linear models which only use information in the first moments (Zeger and Liang, 1986). Information in the second-order moments is necessary for identification of the latent class model parameters in which the covariance between responses is of scientific interest.

The estimating equations are formed by equating the observed responses $\mathbf{y}_i$ and $\mathbf{w}_i = \{(y_{ij} - \mu_{ij})(y_{ih} - \mu_{ih}); j < h = 1,..., p\}$ to their expected values $\boldsymbol{\mu}_i = \boldsymbol{E}[\mathbf{Y}_i|\mathbf{z}_i]$ and $\boldsymbol{\sigma}_i = \boldsymbol{E}[\mathbf{W}_i|\mathbf{z}_i]$. The first-order moments are given by

$$\mu_{ij}=E\left[Y_{ij}|\mathrm{z}_i\right]=\sum_{m=1}^{C}p_{jm}(\alpha,\lambda)\,\pi_m(\beta,\zeta)$$

for both the traditional and the locally dependent LCA as seen in equation (1). The second-order moments are given by equation (3) for the traditional locally independent LCA and by substituting equation (7) into equation (2) for item pairs in which local dependence is specified.

The estimating equations are weighted by the matrix

$$\mathbf{D}_i=\left(\frac{\partial\mu_i(\theta)}{\partial\theta},\frac{\partial\sigma_i(\theta)}{\partial\theta}\right)$$

of first-order derivatives of the first two moments with respect to the set of parameters $\boldsymbol{\theta}$ and a working $p(p + 1)/2 \times p(p + 1)/2$ covariance matrix $\mathbf{R}_i$ of $\mathbf{Y}_i$ and $\mathbf{W}_i$. The covariance matrix is referred to as working because, as demonstrated by Liang *et al.* (1992), parameter estimates and standard errors remain consistent even if the covariance is misspecified. Various simplifying assumptions can be made to ease computations. In particular, we assume working independence for $\mathbf{R}_i$.

The second-order estimating equations proposed are then

$$U(\theta)=\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{R}_i(\theta)^{-1}\begin{pmatrix}\mathbf{y}_i-\mu_i(\theta)\\ \mathbf{w}_i(\theta)-\sigma_i(\theta)\end{pmatrix}=0. \tag{10}$$

These estimating equations are solved by using a Newton–Raphson iterative procedure. Starting with an initial estimate $\widehat{\theta}^0$, the estimate at the $(r + 1)$th iteration is given by

$$\widehat{\theta}^{r+1}=\widehat{\theta}^{r}+\left(\sum_{i=1}^{N}\widehat{\mathbf{D}}_i'\widehat{\mathbf{R}}_i^{-1}\widehat{\mathbf{D}}_i\right)^{-1}\sum_{i=1}^{N}\widehat{\mathbf{D}}_i'\widehat{\mathbf{R}}_i^{-1}\begin{pmatrix}\mathbf{y}_i-\widehat{\mu}_i\\ \mathbf{w}_i-\widehat{\sigma}_i\end{pmatrix},\qquad r=0,1,\ldots, \tag{11}$$

where $\widehat{\mathbf{D}}_i=\mathbf{D}_i\left(\widehat{\theta}^r\right),\widehat{\mathbf{R}}_i=\mathbf{R}_i\left(\widehat{\theta}^r\right),\widehat{\mu}_i=\mu_i\left(\widehat{\theta}^r\right)$ and $\widehat{\sigma}_i=\sigma_i\left(\widehat{\theta}^r\right)$.

The asymptotically robust covariance estimator for $\widehat{\theta}$ estimated by solution of equation (11) is given by

$$\left(\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{R}_i^{-1}\mathbf{D}_i\right)^{-1}\left\{\sum_{i=1}^{N}\mathrm{D}_i'\mathbf{R}_i^{-1}\mathrm{cov}\begin{pmatrix}\mathbf{y}_i\\ \mathbf{w}_i\end{pmatrix}\mathbf{R}_i^{-1}\mathrm{D}_i\right\}\left(\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{R}_i^{-1}\mathbf{D}_i\right)^{-1}.$$

A consistent estimator of the asymptotic covariance of $\widehat{\theta}$ that is robust to misspecification of the covariance of $\mathbf{Y}_i$ and $\mathbf{W}_i$ is obtained by replacing $\boldsymbol{\theta}$ by $\widehat{\theta}$ and the covariance of $\mathbf{Y}_i$ and $\mathbf{W}_i$ by its empirical estimate. Statistical programs to implement this approach have been written in the Fortran programming language and are available by request from the authors.

3.2. Goodness of fit

Traditionally, LCA proceeds by starting with the most parsimonious one-class model and fitting successive models with an increasing number of classes to determine the most parsimonious model that provides an adequate fit to the data. Because models with different numbers of classes are not nested, precluding the use of a difference likelihood ratio test, LCA must rely on measures of fit such as Akaike's information criterion AIC, which is a global fit

NIH-PA Author Manuscript
NIH-PA Author Manuscript

index which combines goodness of fit and parsimony. In comparing different sets of models with the same set of data, models with lower values are preferred. On the basis of a recommendation of Lin and Dayton (1997), AIC is felt to be more appropriate than the Bayesian information criterion BIC when there are complex models of the type that is encountered in this research. AIC, however, requires a likelihood for model comparison and estimating equations approaches are non-likelihood based. Pan (2001) proposed a version of AIC for estimating equations that is based on the quasi-likelihood under the working independence model with $\hat{\theta}$ estimated by using any general working correlation structure. We consider a modified version of Pan's AIC that was proposed by Reboussin *et al.* (2006a) for the second-order estimating equations (10). They demonstrated that this modified criterion performed well in a small simulation study for LCAs that are estimated by using second-order estimating equations:

$$\mathrm{AIC}=\sum_{i=1}^{N}\{\mathbf{y}_i-\mu_i(\theta)\}\,\mathbf{V}^{-1}\,\{\mathbf{y}_i-\mu_i(\theta)\}+\sum_{i=1}^{N}\{\mathbf{w}_i(\theta)-\sigma_i(\theta)\}\,\mathbf{W}^{-1}\,\{\mathbf{w}_i-\sigma_i(\theta)\}+2r \qquad (12)$$

where $\mathbf{V}$ and $\mathbf{W}$ are the working covariance matrices of $\mathbf{y}_i$ and $\mathbf{w}_i$ calculated under independence and r is the total number of estimated parameters.

Although AIC is preferred over other global goodness-of-fit statistics like BIC when the true model is complex (i.e. based on a relatively large number of parameters), its tendency to select more complex models can result in classes that are not substantively interpretable. Because increasing the number of latent classes always improves the key assumption of local independence in LCA, the researcher is sometimes confronted with choosing between an empirically superior model based on AIC that may be difficult to interpret and a substantively more meaningful model with potential model misspecification. Therefore, in addition to AIC, it is important in model selection both to evaluate the resultant latent structure and to consider a more direct measure of the local independence assumption. If local dependences are detected in a substantively interpretable but empirically inferior model, relaxing the local independence assumption may provide the flexibility that is needed to select a less complex model. In the absence of *a priori* ideas about what items may or may not exhibit local dependence, various diagnostics for detecting local dependences have been proposed (Espeland and Handleman, 1989; Hagenaars, 1988; Qu *et al.*, 1996; Garrett and Zeger, 2000; Vermunt and Magidson, 2000). We use a modified version of Garrett and Zeger's (2000) log-odds-ratio check that was suggested by Uebersax (2000). This method involves calculating the log-odds-ratio ψ in both the observed and the expected√ two-way tables for pairs of items. The standard error of ψ for the expected data is given by $\sqrt{(1/a + 1/b + 1/c + 1/d)}$ where a, b, c and d denote the four frequencies in the expected two-way tables. The observed data ψ are then expressed as a z-score relative to the expected data ψ

$$z=\frac{\psi\,(\text{observed})-\psi\,(\text{expected})}{\sigma\,\{\psi\,(\text{expected})\}}. \qquad (13)$$

This z-value is then used as a guide to detect items that are locally dependent. Uebersax (2000) suggested that p-values should not be interpreted literally but that we should rather focus on the relative magnitude of the z-values.

4. Under-age drinking example

Our application concerns a study of under-age drinkers that were assessed as part of the 'Enforcing underage drinking laws (EUDL) randomized community trial' (CT). The EUDL CT is a programme that is funded by the US Office of Juvenile Justice and Delinquency Prevention to increase enforcement of under-age drinking laws and to reduce under-age drinking (Wolfson *et al.*, 2005). We considered a sample of 7103 youths aged 14–20 years from 70 communities (35 intervention and 35 matched control communities) in five states who were surveyed by telephone at baseline in 2004 about their involvement with alcohol. Of the 7103 youths who were surveyed, 2187 (30.8%) were current drinkers and 1676 of the 2187 (76.6%) had complete data for the analyses that are reported.

We characterized under-age drinking by considering responses to eight questions concerning

- **a.** regular drinking, i.e drinking at least six times during the past month,
- **b.** binge drinking,
- **c.** becoming drunk at least 2 or 3 days a month,
- **d.** driving after drinking,
- **e.** passing out as a result of drinking,
- **f.** being unable to remember what happened as a result of drinking,
- **g.** having a headache or hangover as a result of drinking and
- **h.** social consequences as a result of drinking.

Social consequences included being cited or arrested for drinking, possessing or trying to buy alcohol, missing any school due to drinking, being warned by a friend about drinking, breaking or damaging something as a result of drinking and being punished by parents or guardians as a result of drinking. The five alcohol-related social consequences were combined into a single indicator of social consequences because of their very low prevalence and lack of discriminatory power. More details on these measures can be found in Reboussin *et al.* (2006b). Because community populations represent intact social groups, youths within a community are likely to be more like one another than they are to be like youths in other communities and therefore we expect there to be a correlation of drinking behaviours within a community (Murray, 1998). Failure to account for this correlation could result in inflated type I error rates. For this reason, we used the robust sandwich covariance estimator of Reboussin *et al.* (2006a) for cluster trial designs that incorporates the corrected cross-products for different participants within clusters in the middle term of the variance.

Locally independent LCAs were first applied to examine the structure underlying the set of eight co-occurring drinking behaviours and alcohol-related problems among under-age drinkers. Following the work of Bandeen-Roche *et al.* (1997), we inferred the number of classes ignoring covariates. We started with the most parsimonious one-class model (‘all drinkers the same’) with progression to a less parsimonious model with five classes of under-age drinkers. As seen in Table 1, AIC suggested a best-fitting model that is based on four classes of under-age drinkers. However, rather than rely solely on global indices of fit like AIC, we performed a modified log-odds-ratio check to examine the validity of the local independence assumption more directly. Presented in Fig. 1 are the z-values for the locally independent two-, three- and four-class models. A threshold of ±1.5 was conservatively chosen as suggestive of local dependence. Under the two-class model, there was evidence for a large number of local dependences between pairs of drinking behaviours. The addition of a third class reduced the local dependences considerably. However, it was not until introduction of a fourth class that the local dependences no longer remained. Hence, both AIC and residual diagnostics supported a four-class LCA.

As mentioned previously, a statistical analysis may yield models that are not substantively meaningful despite their optimality based on goodness-of-fit statistics like AIC. Therefore, we examined the resultant latent class structures for the locally independent two-, three- and four-class models to evaluate their interpretability. As displayed in Fig. 2, the two-class model divided under-age drinkers into non-problem drinkers and drinkers experiencing alcohol-related problems. The problem drinkers were characterized by both regular and risky (binge drinking and becoming drunk) drinking behaviours. The three-class model identified two classes of under-age problem drinkers in addition to non-problem drinkers; a class of ‘risky’ problem drinkers whose drinking patterns were characterized by binge drinking and becoming drunk and a class of ‘regular’ problem drinkers who not only exhibited risky drinking behaviours but who were more regular drinkers. This finding is consistent with the developmental sequencing of alcohol use where drinking among adolescents often occurs in a less frequent party setting but one in which consumption is high. It is often not until young or later adulthood that drinking patterns become more regular in a subgroup of at-risk drinkers. The four-class model, which removes evidence of any local dependences, further subdivided the risky drinkers. In addition to the non-problem drinkers and regular problem drinkers, there were two classes of youth with the same pattern of risky drinking behaviours which are herein called ‘moderately risky’ and ‘risky’ drinkers, the difference being that one group had higher probabilities of exhibiting each drinking behaviour. Unexpectedly, the class with the higher probability of risky drinking behaviour had substantially lower probabilities of experiencing problems as a result of drinking compared with those with lower probabilities of risky drinking behaviour. Hence, the substantive meaning of the four-class model was not immediately clear.

Since the two-class model exhibited a significant amount of local dependences based on our residual diagnostic statistics that are shown in Fig. 1 and the three-class model provided a substantively meaningful classification of under-age drinkers, we explored whether we could improve the fit of the three-class model by relaxing local independence assumptions and thereby avoid introduction of a possibly spurious fourth class. We began by fitting a series of local dependence models for pairs of items identified through residual diagnostics in Fig. 1 as exhibiting some evidence of local dependence under the three-class model. First, we fit the least restrictive model, the unstructured model (5), that allowed the local dependence to be different for each pair of items and to vary by latent class. Because some of the local dependences that were identified through residual diagnostics were not statistically significant under this model, we used a backward elimination procedure to simplify the model. The item pair with the least significant local dependence was removed first and the model was refitted. Each subsequent step removed the least significant local dependence in the model until all the remaining local dependences had *p*-values that were smaller than 0.05. The final model resulted in statistically significant local dependences between

- **a.** binge drinking and becoming drunk (γ_{23m}),
- **b.** becoming drunk and experiencing headaches (γ_{37m}), and
- **c.** being unable to remember what happened while drinking and social problems (γ_{68m}).

As seen in Table 2, binge drinking and becoming drunk had a large local dependence within the class of risky drinkers (class 2) (odds ratio OR = γ_{232} = 5.9; 95% confidence interval (CI) = (3.5, 9.9)) and a moderate residual association among non-problem drinkers (class 1: OR = 1.7; 95% CI = (1.2, 2.3)) and regular problem drinkers (class 3: OR = 1.8; 95% CI = (1.7, 1.8)). The large local dependence between binge drinking and becoming drunk within the risky drinker class might explain why the four-class model creates two classes of risky drinkers in an attempt to explain this residual association. By directly modelling this pattern of local dependence, there is a clearer separation between the non-problem drinker and risky problem drinker classes with binge drinking and becoming drunk now stronger indicators of the risky problem drinker class. This is seen in Table 2 by the larger response probabilities for these measures compared with the local independence model.

Because the residual associations were fairly constant across classes for the local dependences between becoming drunk and experiencing headaches and being unable to remember what happened while drinking and social problems, we fitted the class exchangeable model (6) and it provided a better fit, as indicated by the AICs in Table 1, than both the unstructured and the four-class locally independent models. In addition, the class structure for the unstructured and class exchangeable models were equivalent, providing motivation for choosing the more parsimonious class exchangeable model. We then fitted the exchangeable model (4) for the subset of items that were statistically significant under the unstructured model and it did not fit as well as the class exchangeable model (the data are not shown; AIC = 62119). This may be due to the larger local dependence between binge drinking and becoming drunk compared with the other two local dependences. Finally, to compare with the random-effects model of Qu *et al.* (1996), we fitted the exchangeable model (4) for all 28 pairs of items and it did not fit the data well (AIC = 62447). The presence of a large number of items that did not exhibit any local dependence as well as larger residual associations between some items might explain the poor model fit. The class structure for the random-effects local dependence model was also quite different from the structure under the other local dependence models. For example, almost twice as many drinkers were classified as risky problem drinkers under this model (class 2 prevalence 57% compared with 34%) and the likelihood of experiencing alcohol-related problems was lower for both types of problem drinkers compared with the other local dependence and locally independent models.

To learn more about the LCA-derived types of under-age problem drinkers, covariates were included in the baseline category latent class logistic regression model (8) for the three-class locally dependent class exchangeable model that was just described. Specifically, we were interested in simultaneously modelling the odds of being

a. a risky problem drinker relative to a non-problem drinker and

b. a regular problem drinker relative to a non-problem drinker as a function of gender (male *versus* female), age (18–20 and 16–17 years *versus* 14–15 years), past 30-day use of marijuana and beliefs that most friends become drunk.

The addition of covariates to the latent class model for the prevalences introduces a second type of local independence assumption as described previously. We must now assume that responses and covariates are independent conditional on class or, in other words, that covariates do not affect the measurement process directly. However, prior studies have found that males endorse certain alcohol-related problems (e.g. social and legal) at higher rates than females (Perkins, 1992; Wechsler *et al.*, 1995; Humara and Sherman, 2004) despite the fact that both may be considered problem drinkers. Therefore, when fitting model (8) we explored whether the non-differential measurement condition was valid for these data. We did this by allowing the response probabilities to depend on gender by using model (9). Using a backward elimination procedure, the final model resulted in significant associations between gender and three indicators:

a. passing out,

b. having a headache and

c. social problems that were not explained by latent class.

Males were significantly more likely to report social problems than females in the non-problem drinker and regular problem drinking classes supporting prior studies (OR = 2.4, 95% CI (1.4,4.1), and OR = 1.6, 95% CI (1.0,2.8) respectively). Interestingly, males were also less likely to report physical problems such as headaches in the regular problem drinking class (OR = 0.35, 95% CI (0.13,0.95)), and passing out in the non-problem and risky drinking classes (OR = 0.5, 95% CI (0.4,0.5), and OR = 0.7, 95% CI (0.4,1.0) respectively). On the basis of these models with differential measurement for a subset of items, the response probabilities for males and females are presented in Table 3. Interestingly, the probability to report alcohol-related problems is now consistently greater in the regular drinking class compared with the risky drinking class for both males and females in contrast with the model which assumed non-differential measurement.

Presented in Table 4 are the estimated odds ratios and 95% CIs from the latent class prevalence regression models for the locally independent three-class model, the locally dependent model with non-differential measurement and the locally dependent model with differential measurement for gender. In general, the locally dependent LCA regression parameter estimates were larger than the locally independent LCA estimates. Although the magnitude of the effects was greater for the locally dependent LCA model with differential measurement, the inferences were generally consistent. However, gender was marginally significant when comparing the risky and non-problem drinker classes when it had previously not been statistically significant. By allowing males to have lower probabilities of endorsing physical problems from drinking in the risky drinking class, males were approximately 1.4 times more likely to be risky problem drinkers relative to non-problem drinkers than females (OR = 1.35; 95% CI (0.98,1.86)). On the basis of the results from the locally dependent LCA with differential measurement, males were also two and a half times more likely to be regular problem drinkers relative to non-problem drinkers than females (OR = 2.43; 95% CI (1.52,3.87)). Late adolescents aged 16–17 years were five times more likely to be regular problem drinkers relative to non-problem drinkers than adolescents aged 14–15 years (OR = 4.80; 95% CI (1.87,12.3)) and young adults aged 18–20 years were almost 13 times more likely to be in the regular problem drinker class compared with non-problem drinkers (OR = 12.6; 95% CI (4.89,32.3)). Age was not significantly associated with risky problem drinking relative to non-problem drinking. Youths smoking marijuana in the past 30 days were seven times more likely to be risky problem drinkers (OR 6.85; 95% CI (4.05,11.6)) and 18 times more likely to be regular problem drinkers (OR = 17.6; 95% CI (11.1,27.9)) relative to non-problem drinkers. Youths who believe that most of their friends become drunk were three and a half times more likely to be risky problem drinkers (OR = 3.59; 95% CI (2.21,5.83)) and 21 times more likely to be regular problem drinkers (OR = 21.3; 95% CI (14.6,31.3)) compared with non-problem drinkers.

5. Discussion

From a practical standpoint, the most useful feature for introducing local dependence into LCA may be the flexibility that it provides for selecting substantively interpretable models over models that are deemed empirically superior. Our experience suggests that conventional goodness-offit statistics tend to gravitate towards choosing more complex LCAs. The under-age drinking example with a large sample size and no *a priori* hypothesis about the underlying class structure was at particular risk of overfitting. Although four classes provided the best overall fit to the data, the further subdivision of the population was difficult to interpret on the basis of current theory. Although there were no *a priori* ideas about local dependence, the subject matter was such that many of the items were measuring very closely related behaviours. In fact, the strongest local dependence was found between the behaviours of binge drinking and becoming drunk, two behaviours that may frequently be associated. Similarly, becoming drunk and having headaches, which also exhibited local dependence, seem to go hand in hand. By inspection of the observed and expected log-odds-ratios, we could detect local dependences and then relax the local independence assumption. This avoided the introduction of substantively meaningless classes and perhaps biased and less powerful inferences for the regression component of our model for the latent class prevalences. In the absence of methods for relaxing local independence, if we had chosen the three-class LCA because of its substantive interpretation and parsimony, we would have obtained smaller effect sizes (i.e. odds ratios) for covariates in our regression model than if we had allowed for local dependences. For this example, the local dependences were not so great as to have a significant effect on our inferences (i.e. *p*-values). However, by allowing the response probabilities to depend on gender for a subset of items, the three-class model for males and females not only gained in substantive interpretation but also resulted in the detection of a marginally significant relationship between gender and risky problem drinking that was not detected under the locally independent and locally dependent model with non-differential measurement.

As described in Section 1, there are a multitude of methods for incorporating local dependence into LCA models. Early work, such as Harper (1972), used additive models (Darroch and Speed, 1983) for introducing association between item pairs. More recent work involves the use of random effects (Qu *et al.*, 1996), a probit model (Uebersax 1999), conditional log-linear (multiplicative) models (Hagenaars, 1988) and marginal log-linear models (Becker and Yang, 1998). Not surprisingly, parallel methods for introducing dependence structure exist in the longitudinal data analysis literature, and it has been well documented (e.g. Diggle *et al.* (2002) and Fitzmaurice *et al.* (2004)) that in such situations each method can be appropriately applied. This paper deals with the situation where the primary scientific interest is modelling the marginal latent class prevalences as a function of covariates while allowing for the relaxation of the local independence and non-differential measurement assumptions. In terms of how the within-class residual association is modelled, our approach is most similar in spirit to the marginal log-linear LCA of Becker and Yang (1998). The likelihood approach of the marginal log-linear LCA, however, requires the computation of the joint distribution of all item pairs within clusters, which may not be feasible when large clusters exist within the instrument. However, the second-order estimating equations approach requires specification of only the first two moments and is robust to misspecification of the $p(p + 1)/2 \times p(p + 1)/2$ covariance matrix. This allows us to choose a computationally simple alternative—working independence—and still to obtain unbiased parameter estimates. Asymptotic results from Liang *et al.* (1992) have shown that generalized estimating equation type estimates can be reasonably efficient for the regression parameters.

The greatest strength of our approach is the use of a flexible set of models that can be fitted to account for local dependence and non-differential measurement while simultaneously modelling the latent class prevalences as a function of covariates. The use of the odds ratio to measure the local dependence between two binary items has many desirable properties (Bishop *et al.*, 1975) and is easier to interpret than both correlation coefficients and conditional odds ratios. Our models allow for both positive and negative local dependence, local dependence between a subset of items and various assumptions regarding the structure of the local dependence that reflect the subject matter (e.g. full exchangeability, class exchangeability, item exchangeability or unstructured). In addition if, after introducing covariates into the LCA model for the prevalences, the non-differential measurement assumption is not tenable, a model which is separate from the local dependence model can be fitted that models the response probabilities as a function of covariates. The regression parameters from these models then have a marginal interpretation in contrast with random effects and log-linear models which have conditional interpretations.

Limitations in our approach, however, should be noted. In particular, although simultaneous modelling of the mean and association models using second-order estimating equations gives more efficient estimates of the association parameters than first-order estimating equations, to ease the computational burden we have sacrificed some of this efficiency by choosing a working independence covariance matrix. Future research is needed to determine the extent of this loss in efficiency relative to the gain in computational feasibility over choosing both a more complex working covariance matrix or over a full likelihood approach most notably in the presence of large numbers of item responses that are common in these types of models. In addition, although studies of the finite sample properties of the estimating equations approach for LCA have been favourable (see Reboussin and Anthony (2001) and Reboussin *et al.* (2006b)), their performance in the presence of local dependence and differential measurement is an area of future research. Although relaxation of local independence assumptions may lead us to choose more substantively interpretable and less complex models, the potential biases that are associated with misspecifying the local dependence structure also warrants further study. For example, in our EUDL CT data the latent class structure was quite different when assuming the random-effects structure for local dependence compared with the other local dependence structures. Inferences under the three-class local independence model might have been less biased than those under an incorrectly specified local dependence model despite the larger AIC and violation of local independence assumptions. Finally, our models allow the local dependence structure to depend on the specific item pairing and/or class. In other contexts in which latent variables are used, local dependence between items may contain substantive information (Ip *et al.*, 2004). Such information could be captured by including covariates in the model for the pairwise within-class odds ratios similarly to Carey *et al.* (1993) within the generalized linear model framework.

Similarly to an increasing number of researchers, we have attempted to shed light on the nature of a problem, in this case under-age drinking, by modelling patterns of response (or behaviour) rather than relying on unidimensional measures or clinical criteria. As such, it provides a better understanding of the heterogeneous nature of under-age problem drinking that would not have emerged if we had relied on definitions of problem drinking that have developed in adult populations. By relaxing the local independence assumption, we could choose a substantively interpretable model that revealed two subtypes of under-age drinkers experiencing alcohol-related problems, namely risky drinkers and regular drinkers. It is notable that, for a group of under-age drinkers with a moderate prevalence of heavy drinking behaviour (i.e the ‘risky’ drinkers), alcohol-related problems are a serious concern. This finding may suggest that early intervention is necessary before the emergence of both more heavy and regular drinking in this population. It may also set the stage for longitudinal studies which can probe further the nature of the emergence of problem drinking. Even perhaps more importantly, by allowing for differential measurement we found evidence in our sample that, although underage male and female problem drinkers have similar rates of drinking, females tend to endorse physical symptoms at higher rates whereas males endorse more social problems. These findings of gender differences in endorsement of problems that are related to drinking may provide insight into tailoring prevention programmes as well as screening criteria for identifying under-age problem drinkers to be gender specific.

References

Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. J. Am. Statist. Ass 1997;92:1375–1386.

Banfield JD, Raftery AE. Model-based gaussian and non-gaussian clustering. Biometrics 1993;49:803–821.

Becker, MP.; Yang, IS. Latent class marginal models for cross-classifications of counts.. In: Raftery, A., editor. Sociological Methodology. Blackwell; Boston: 1998.

Bishop, YMM.; Fienberg, SE.; Holland, PW. Discrete Multivariate Analysis: Theory and Practice. Massachusetts Institute of Technology Press; Cambridge: 1975.

Carey V, Zeger SL, Diggle P. Modeling multivariate binary data with alternating logistic regressions. Biometrika 1993;80:517–526.

Carlson RG, Wang J, Falck RS, Siegal HA. Drug use practices among MDMA/Ecstasy users in Ohio: a latent class analysis. Drug Alc. Depend 2005;79:167–179.

Chung T, Colby SM, Barnett NP, Rohsenow DJ, Spirito A, Monti PM. Screening adolescents for problem drinking: performance of brief screens against DSM-IV alcohol diagnoses. J. Stud. Alc 2000;61:579–587.

Chung T, Martin CS, Armstrong TD, Labouvie EW. Prevalence of DSM-IV alcohol diagnoses and symptoms in adolescent community and clinical samples. J. Am. Acad. Chld Adolesc. Psychiatr 2002;41:546–554.

Crum RM, Storr CL, Chan YF. Depression syndromes with risk of alcohol dependence in adulthood: a latent class analysis. Drug Alc. Depend 2005;79:71–81.

Dale JR. Global cross-ratio models for bivariate, discrete, ordered responses. Biometrics 1986;42:909–9l7. [PubMed: 3814731]

Darroch JN, Speed TP. Additive and multiplicative models and interactions. Ann. Statist 1983;11:724–738.

von Davier M, Yamamoto K. Partially observed mixtures of IRT models: an extension of the generalized partial-credit model. Appl. Psychol. Measmnt 2004;28:389–406.

Dayton CM, Macready GB. Concomitant-variable latent class models. J. Am. Statist. Ass 1988;83:969–974.

Deas D, Riggs P, Langenbucher J, Goldman M, Brown S. Adolescents are not adults: developmental considerations in alcohol users. Alchlsm Clin. Exptl Res 2000;24:232–237.

Diggle, P.; Heagerty, P.; Liang, K.; Zeger, SL. The Analysis of Longitudinal Data. 2nd edn.. Oxford University Press; Oxford: 2002.

Ellickson PL, McGuigan KA, Adams V, Bell RM, Hays RD. Teenagers and alcohol misuse in the United States: by any definition, it's a big problem. Addiction 1996;91:1489–1503. [PubMed: 8917917]

Espeland MA, Handelman SL. Using latent class models to characterize and assess relative error in discrete measurements. Biometrics 1989;45:587–599. [PubMed: 2765639]

Faden VP, Fay MP. Trends in drinking among Americans age 18 and younger: 1975−2002. Alchlsm Clin. Exptl Res 2004;28:1388–1395.

Ferdinand RF, de Nijs PF, van Lier P, Verhulst FC. Latent class analysis of anxiety and depressive symptoms in referred adolescents. J. Affect. Disorders 2005;88:299–306. [PubMed: 16182373]

Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. Wiley; New York: 2004.

Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. Statist. Med 2002;21:1289–1307.

Garrett ES, Zeger SL. Latent class model diagnosis. Biometrics 2000;56:1055–l067. [PubMed: 11129461]

Haberman, SJ. Analysis of Qualitative Data. Academic Press; New York: 1979.

Hagenaars JA. Latent structure models with direct effects between indicators—local dependence models. Sociol. Meth. Res 1988;16:379–405.

Harper D. Local dependence latent structure models. Psychometrika 1972;37:53–59.

Hedeker D, Gibbons RD. A random-effects ordinal regression-model for multilevel analysis. Bio-metrics 1994;50:933–944.

Huang G, Bandeen-Roche K. Building an identifiable latent class model with covariate effects on underlying and measured variables. Psychometrika 2004;69:5–32.

Hudziak JJ, Heath AC, Madden PF, Reich W, Bucholz KK, Slutske W, Bierut LJ, Neuman RJ, Todd RD. Latent class and factor analysis of DSM-IV ADHD: a twin study of female adolescents. J. Am. Acad. Chld Adolesc. Psychiatr 1998;37:848–857.

Humara MJ, Sherman MF. Situational determinants of alcohol abuse among Caucasian and African-American college students. Addict. Behav 2004;24:135–158. [PubMed: 10189981]

Institute of Medicine. Broadening the Base of Treatment for Alcohol Problems. National Academy Press; Washington DC: 1990.

Ip EH, Wang YJ, De Boeck P, Meulders M. Locally dependent latent trait model for polytomous responses with application to inventory of hostility. Psychometrika 2004;69:191–216.

Liang K-Y, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data (with discussion). J. R. Statist. Soc. B 1992;54:3–40.

Lin TS, Dayton CM. Model selection information criteria for non-nested latent class models. J. Educ. Behav. Statist 1997;22:249–264.

Lin HQ, Turnbull BW, McCulloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. J. Am. Statist. Ass 2002;97:53–65.

Lord, FM.; Novick, MR. Statistical Theories of Mental Test Scores. Addison-Wesley; 1968. Reading

Murray, DM. Design and Analysis of Group Randomized Trials. Oxford University Press; New York: 1998.

Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001;57:120–125. [PubMed: 11252586]

Patterson BH, Dayton CM, Graubard B. Latent class analysis of complex sample survey data: application to dietary data. J. Am. Statist. Ass 2002;97:721–729.

Perkins HW. Gender patterns in consequences of collegiate alcohol abuse: a 10-year study of trends in an undergraduate population. J. Stud. Alc 1992;53:458–462.

Qu YS, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 1996;52:797–810. [PubMed: 8805757]

Reboussin BA, Anthony JC. Latent class marginal regression models for modelling youthful drug involvement and its suspected influences. Statist. Med 2001;20:623–639.

Reboussin BA, Anthony JC. Is there epidemiological evidence to support the idea that a cocaine dependence syndrome emerges soon after onset of cocaine use? Neuropsychopharmacology 2006;31:2055–2064. [PubMed: 16482089]

Reboussin BA, Liang KY, Reboussin DM. Estimating equations for a latent transition model with multiple discrete indicators. Biometrics 1999;55:839–845. [PubMed: 11315015]

Reboussin BA, Lohman KK, Wolfson M. Modeling adolescent drug use patterns in cluster-unit trials with multiple sources of correlation using robust latent class regressions. Ann. Epidem 2006a; 16:850–859.

Reboussin BA, Song EY, Shrestha A, Lohman KK, Wolfson M. A latent class analysis of underage problem drinking: evidence from a community sample of 16−20 year olds. Drug Alc. Depend 2006b; 83:199–209.

Roeder K, Lynch KG, Nagin DS. Modeling uncertainty in latent class membership: a case study in criminology. J. Am. Statist. Ass 1999;94:766–776.

Substance Abuse and Mental Health Service Administration. Publication SMA 07−4293. Office of Applied Studies; Rockville: 2007. Results from the 2006 NSDUH: National Findings..

Stewart C, Power TG. Identifying patterns of adolescent drinking: a tri-ethnic study. J. Stud. Alc 2002;63:156–168.

Storr CL, Reboussin BA, Anthony JC. Early childhood misbehavior and the estimated risk of becoming tobacco-dependent. Am. J. Epidem 2004;160:126–130.

Strauss SM, Rindskopf DM, Astone-Twerell JM, Des Jarlais DC, Hagan H. Using latent class analysis to identify patterns of Hepatitis C service provision in drug-free treatment programs in the U.S. Drug Alc. Depend 2006;83:15–24.

Suppes P, Zanotti M. When are probabilistic explanations possible? Synthese 1981;48:191–199.

Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. Statist. Med 1997;16:2157–2175.

Townshend JM, Duka T. Patterns of alcohol drinking in a population of young social drinker: a comparison of questionnaire and diary measures. Alc. Alchlsm 2002;37:187–192.

Uebersax JS. Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models. App. Psychol. Measmnt 1999;23:283–297.

Uebersax, JS. A practical guide to local dependence in latent class models.. 2000. (Available from http:// ourworld.compuserve.com/homepages/jsuebersax.)

United States Department of Health and Human Services. The Surgeon General's Call to Action to prevent and reduce underage drinking. Office of the Surgeon General, United States Department of Health and Human Services; Washington DC: 2007.

Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 1985;41:959–968. [PubMed: 3830260]

Van der Heijden PGM, Dessens J, Bockenholt U. Estimating the concomitant-variable latent-class model with the EM algorithm. J. Educ. Behav. Statist 1996;21:215–229.

Vermunt JK. Multilevel latent class models. Sociol. Methodol 2004;33:213–239.

Vermunt, JK.; Magidson, J. Latent GOLD User's Guide. Statistical Innovations Inc.; Belmont: 2000.

Von Korff M, Miglioretti DL. A prognostic approach to defining chronic pain. Pain 2005;117:304–313. [PubMed: 16153772]

Wechsler H, Dowdall GW, Davenport A, Castillo S. Correlates of college student binge drinking. Am. J. Publ. Hlth 1995;85:921–926.

Wechsler H, Kuo M, Lee H, Dowdall GW. Environmental correlates of underage alcohol use and related problems of college students. Am. J. Prev. Med 2000;19:24–29. [PubMed: 10865160]

Wolfson, M.; Song, E.; Martin, BA.; Wagoner, K.; Brown, V.; Suerken, C.; Reboussin, B.; Foley, KL.; Preisser, J.; Garner, G.; Hulme, S. National evaluation of the enforcing underage drinking laws randomized community trial: annual report, year 1. Department of Public Health Sciences, Wake Forest University School of Medicine; Winston-Salem: 2005.

Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–130. [PubMed: 3719049]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

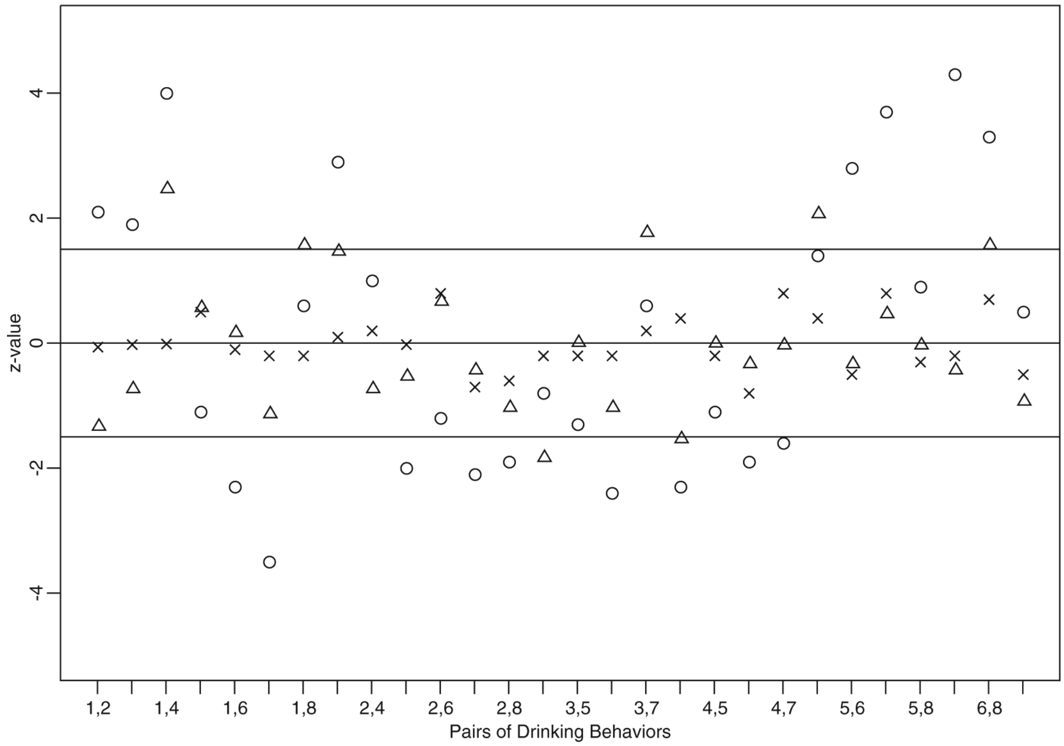

Fig. 1.
z-values for the observed data *ψ* relative to the expected data *ψ* under two- (○), three- (Δ) and four-(×) class locally independent LCAs fitted to the EUDL CT under-age drinking data

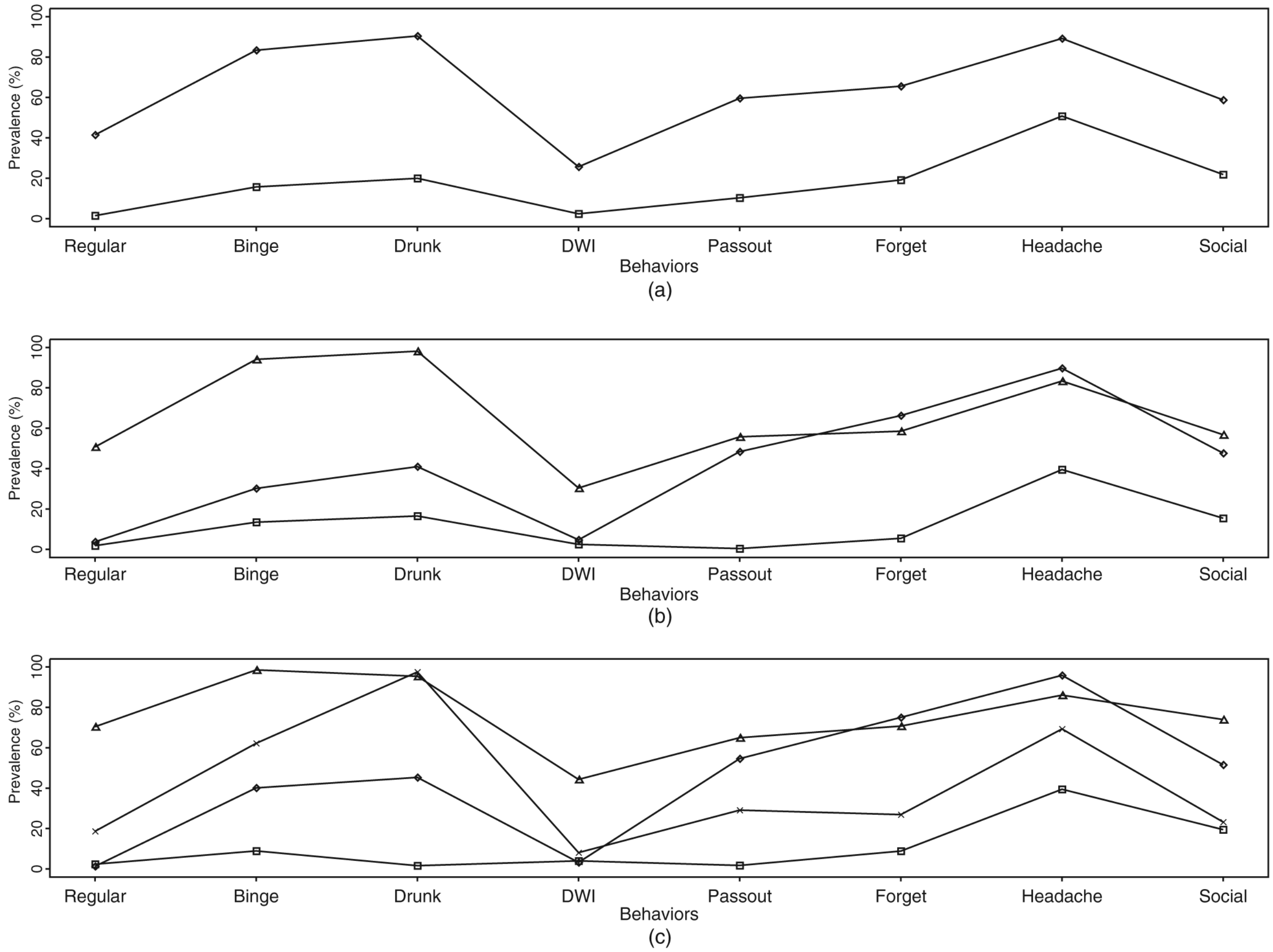

Fig. 2.
(a) Two- (□, non-problem drinker; ◇, problem drinker), (b) three- (□, non-problem drinker; ◇, risky drinker; △, regular drinker) and (c) four- (□, non-problem drinker; ◇, moderately risky drinker; ×, risky drinker; △, regular drinker) class locally independent LCA response profiles estimated from the EUDL CT under-age drinking data (DWI, driving while intoxicated)

Table 1

Goodness-of-fit statistics for locally independent LCAs and a series of three-class locally dependent models fitted to the EUDL CT under-age drinking data

Model	*AIC*	*Number of parameters*
1 class	65393	8
2 class	62212	17
3 class	62136	26
4 class	62124	35
5 class	62138	44
Unstructured locally dependent	62124	35
Class exchangeable locally dependent	62112	29
Exchangeable locally dependent	62119	27
Random-effects locally dependent	62447	27

Table 2

Estimated latent class model parameters from a three-class locally independent LCA and a series of three-class locally dependent LCAs fitted to the EUDL CT under-age drinking data

Class	Response probabilities p_{jm} for the following models:											
	Independence			*Unstructured*			*Class exchangeable*			*Random effects*		
	Non-problem	*Risky*	*Regular*	*Non-problem*	*Risky*	*Regular*	*Non-problem*	*Risky*	*Regular*	*Non-problem*	*Risky*	*Regular*
Regular	0.02	0.03	0.50	0.02	0.05	0.70	0.02	0.05	0.70	0.01	0.15	0.62
Binge	0.13	0.30	0.94	0.15	0.48	0.93	0.15	0.48	0.93	0.06	0.47	1.00
Drunk	0.16	0.40	0.98	0.19	0.55	0.97	0.19	0.55	0.97	0.03	0.58	1.00
Driving while intoxicated	0.02	0.04	0.30	0.02	0.07	0.38	0.02	0.07	0.38	0.05	0.07	0.41
Passout	0.01	0.48	0.56	0.01	0.51	0.58	0.01	0.51	0.58	0.06	0.41	0.35
Forget	0.05	0.66	0.59	0.08	0.64	0.59	0.08	0.64	0.59	0.19	0.47	0.36
Headache	0.39	0.87	0.83	0.40	0.91	0.81	0.40	0.91	0.81	0.34	0.85	0.56
Social problems	0.15	0.47	0.57	0.18	0.44	0.63	0.18	0.44	0.63	0.34	0.35	0.48
Prevalence π_m	43%	26%	31%	45%	34%	21%	45%	34%	21%	29%	57%	14%
Local dependence Odds ratios γ_{jhm}											3.0 (2.5,3.6)	
γ_{23m}				1.7 (1.2,2.3)	5.9 (3.5,9.9)	1.8 (1.7,1.8)		3.5 (2.4, 5.1)				
γ_{37m}				1.7 (1.1,2.7)	1.7 (1.4,2.1)	1.7 (1.6,1.7)		1.7 (1.2, 2.5)				
γ_{68m}				1.7 (1.4,1.9)	1.6 (1.0,2.5)	1.5 (1.1,1.9)		1.6 (1.1, 2.2)				

Table 3

Estimated response probabilities for males and females from a three-class LCA with class exchangeable local dependence structure assuming differential measurement

Class	***Probabilities for females***			***Probabilities for males***		
	Non-problem	***Risky***	***Regular***	***Non-problem***	***Risky***	***Regular***
Regular	0.024	0.030	0.638	0.024	0.030	0.638
Binge	0.100	0.427	0.917	0.100	0.427	0.917
Drunk	0.100	0.548	0.955	0.100	0.548	0.955
Driving while intoxicated	0.030	0.027	0.364	0.030	0.027	0.364
Passout	0.000	0.460	0.549	0.000	0.366	0.616
Forget	0.048	0.549	0.595	0.048	0.549	0.595
Headache	0.376	0.886	0.917	0.314	0.804	0.800
Social problems	0.118	0.435	0.529	0.244	0.362	0.649

Table 4

Estimated odds ratios and 95% CIs for the latent class prevalences, assuming local independence LCA and class exchangeable local dependence with and without non-differential measurement fit to the EUDL CT under-age drinking data

	Results for the locally independent model		*Results for the locally dependent model*			
			Non-differential		*Differential*	
	Risky versus non-problem	*Regular versus non-problem*	*Risky versus non-problem*	*Regular versus non-problem*	*Risky versus non-problem*	*Regular versus non-problem*
Male	1.11 (0.85,1.50)	2.13 (1.39,3.28)	1.09 (0.81,1.47)	2.20 (1.40,3.44)	1.35 (0.98,1.86)	2.43 (1.52,3.87)
16−17 *versus* 14−15 years	1.32 (0.76,2.29)	4.29 (1.79,10.3)	1.33 (0.75,2.36)	4.57 (1.72,12.1)	1.37 (0.76,2.49)	4.80 (1.87,12.3)
18−20 *versus* 14−15 years	1.12 (0.62,2.03)	11.06 (4.56,26.80)	1.18 (0.65,2.17)	12.87 (4.79,34.63)	1.12 (0.61,2.06)	12.56 (4.89,32.27)
Marijuana use	5.80 (3.66,9.18)	14.07 (9.47,20.90)	6.65 (4.08,10.6)	16.15 (10.58,24.67)	6.85 (4.05,11.6)	17.56 (11.07,27.87)
Friends drunk	3.36 (2.11,5.36)	18.04 (11.22,28.98)	3.78 (2.36,6.06)	21.06 (14.35,30.91)	3.59 (2.21,5.83)	21.36 (14.58,31.28)