# Relating latent class membership to continuous distal outcomes: improving the LTB approach and a modified three-step implementation

Zsuzsa Bakk

Daniel L. Oberski

Jeroen K. Vermunt

Department of Methodology and Statistics, Tilburg University, The Netherlands

## Author Note

Address:

Room P 1.113,

PO Box 90153, 5000 LE Tilburg

Phone: +31 13 466 2610

Email: z.bakk@tilburguniversity.edu

This work was supported by The Netherlands Organization for Scientific Research (NWO) [VICI grant number 453-10-002].

Abstract

Latent class analysis often aims to relate the classes to continuous external consequences ("distal outcomes"), but estimating such relationships necessitates distributional assumptions. Lanza et al. (2013) suggested to circumvent such assumptions with their "LTB approach": linear logistic regression of latent class membership on each distal outcome is first used, after which this estimated relationship is reversed using Bayes' rule. However, the LTB approach currently has three drawbacks, which we address in this article. First, LTB interchanges the assumption of normality for one of homoskedasticity, or, equivalently, of linearity of the logistic regression, leading to bias; fortunately, we show introducing higher-order terms prevent this bias. Second, we improve coverage rates by replacing approximate standard errors with resampling methods. Finally, we introduce a bias-corrected three-step version of LTB as a practical alternative to standard LTB. The improved LTB methods are validated by a simulation study, while an example application demonstrates their usefulness.

# Relating latent class membership to continuous distal outcomes: improving the LTB approach and a modified three-step implementation

## Introduction

Latent class (LC) analysis is a well-known approach used in the social and behavioral sciences to create subgroups of units with similar scores on a set of observed indicator or response variables. In many applications, the interest lies not only in the clustering of units, but also in investigating whether LCs differ with respect to the mean of one or more continuous distal outcome variable. For example, De Cuyper, Rigotti, Witte, and Mohr (2008) compared the class-specific means of job insecurity for psychological contract type clusters and Mulder, Vermunt, Brand, Bullens, and Van Merle (2012) compared the means of juvenile offenders clusters on outcomes measuring recidivism. Other examples are predicting alcohol dependence from early substance abuse clusters and predicting the contraction of sexually transmitted diseases from sexual risk behavior clusters (Lanza, Tan, & Bray, 2013).

The class-specific means of a continuous distal outcome can be estimated by expanding the LC model with the outcome as an additional indicator. The main problem with this approach, which is referred to as the one-step approach, is that assumptions have to be made about the within-class distribution of the distal outcome. Typically this will be the assumption of normality. However, in case this assumption is violated the whole LC solution can change when including the distal outcome, or even more classes can be extracted than would without this variable included (Bauer & Curran, 2003).

Lanza, Tan, and Bray propose an approach (called "LTB approach" after its developers) that bypasses the difficulties arising from potential violations of distributional assumptions. It involves estimating a LC model in which the distal outcome variable is used as a covariate affecting the LCs instead of a response variable. Subsequently, using the estimates from this model, the class-specific means of the distal outcome variable are calculated (Lanza et al., 2013). The approach is implemented in standard software for LC analysis, such as Mplus 7.1 (Muthén & Muthén, 1998-2012), and Latent GOLD 5.0 (Vermunt & Magidson, 2013).

While promising, the LTB approach has a few shortcomings that we address in

this paper. First of all, when the distal outcome has heteroskedastic errors across classes, the LTB method may yield biased estimates of the class-specific means (Bakk & Vermunt, in press). We show how this bias can be prevented by including a quadratic term in the multinomial regression model for the classes. This is similar to what is done in a quadratic discriminant analysis. Furthermore while in the original article (Lanza et al., 2014) no standard error estimator was proposed, Asparouhov and Muthén (2014) proposed an ad-hoc estimator that is downward biased (Bakk & Vermunt, in press; Asparouhov & Muthén, 2014), thus obtaining too low coverage rates. We propose resolving this problem by using standard resampling methods, i.e. bootstrap-based or jackknifed standard errors.

Furthermore we propose a three-step estimation of the LTB approach. Many applied researchers prefer to first establish a measurement model, and in a later stage relate it to external variables of interest. It is also common that the measurement model is built by one researcher, while the structural model (relating LC membership to external variables) is built by others. It then is useful to have a three-step approach available. This proceeds as follows: 1) a standard LC analysis is performed using only the indicator variables, 2) individuals are assigned to latent classes, and 3) the assigned class scores are regressed on the distal outcome of interest, while correcting for the classification error introduced in the second step (Vermunt, 2010). Based on the parameters obtained in the third step, the class-specific means of the distal outcome can be calculated. This three-step implementation can also be useful when the model estimates using the LTB approach is part of a larger, complex model.

In the remainder of this paper, we first introduce the basic LC model, then present the simultaneous LTB approach (as proposed by Lanza et al. 2014), and subsequently discuss its three-step implementation. Next, we introduce the proposed correction for the situation where the distal outcome has heteroskedastic errors, followed by the introduction of the alternative SE estimators. Subsequently, we present the results of a simulation study investigating the performance of the proposed improvements, and we demonstrate the use of the proposed methods via an example

explaining respondent's income from parents' social status. Lastly, we conclude and suggest directions for future research.

## The basic LC model

Let $Y_{ik}$ denote the response of individual $i$ on one of $K$ categorical indicator variables, where $1 \leq k \leq K$ and $1 \leq i \leq N$. The full response vector is denoted by $\mathbf{Y}_i$. LC analysis assumes that respondents belong to one of the $T$ categories of an underlying categorical latent variable $X$ which affects the responses (McCutcheon, 1987; Goodman, 1974; Hagenaars, 1990). Denoting a particular latent class by $t$, the model can be formulated as follows:

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X = t) P(\mathbf{Y}_i | X = t), \tag{1}$$

where $P(X = t)$ represents the (unconditional) probability of belonging to latent class $t$ and $P(\mathbf{Y}_i|X = t)$ represents the class-specific response probabilities on the indicators. Furthermore, we assume that the $K$ indicator variables are independent within classes, which is known as the local independence assumption. This yields:

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X = t) \prod_{k=1}^{K} P(Y_{ik} | X = t). \tag{2}$$

For categorical responses, $P(Y_{ik}|X = t) = \prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)}$, where $\pi_{ktr}$ is probability of response $r$ on variable $k$ for class $t$, and $I(Y_{ik} = r)$ is an indicator variable taking on the value 1 if $Y_{ik} = r$ and 0 otherwise.

The basic LC model can be extended to include a continuous distal outcome variable, which involves adding this variable to the model as an additional indicator and defining its class-specific distribution. However, this approach is hardly ever used in practice. Alternative approaches are the LTB approach and the three-step approach discussed in the next sections.

### The simultaneous LTB approach

The LTB approach was developed with the goal to make it possible to estimate the association between the LC membership and the distal outcome without making strong distributional assumptions about the latter. This is especially important in case of a continuous outcome variable, in which case the assumption of normal class-specific distribution is often violated. As a consequence of violating this assumption a completely different LC model can be estimated, when the distal outcome is added, a possibility that is often not intended/ desired by researchers. Because of these issues the LTB approach is preferred over the one-step approach in many applications. Using the LTB approach, first a LC model is estimated with the distal outcome, say $Z$, included as covariate to the basic LC model. Subsequently, using Bayes theorem the class-specific means of the distal outcome are calculated (see Figure 1). We will call this approach originally proposed by Lanza, Tan, and Bray (2014) the simultaneous LTB approach.

In step one, $Z$ is included as a covariate to the basic model, by extending Equation 2 to model $P(\mathbf{Y}_i|Z_i)$ instead of $P(\mathbf{Y}_i)$ (Dayton & Macready, 1988; Bandeen-Roche, Miglioretti, & Zegger, 1997):

$$P(\mathbf{Y}_i|Z_i) = \sum_{t=1}^{T} P(X = t|Z_i)) \prod_{k=1}^{K} P(Y_{ik}|X = t). \tag{3}$$

This model assumes that the indicator variables are conditionally independent of the covariate given the latent variable $X$. This is a standard assumption, made by the approaches available in literature to relate LC membership to external variables. If a direct effect is hypothesized, this should be explicitly modeled. The $P(X = t|Z_i)$ is parametrized using a multinomial logistic regression model:

$$P(X = t|Z_i) = \frac{e^{\alpha_t + \beta_t Z_i}}{1 + \sum_{t'=1}^{T-1} e^{\alpha_{t'} + \beta_{t'} Z_i}}, \tag{4}$$

where $\alpha_t$ and $\beta_t$ are the intercept and slope coefficients for class $t$.

Next, in step two, the class-specific means $\mu_t$ are computed:

$$\mu_t = \int_Z Z \ f(Z|X = t) \tag{5}$$

where $f(Z|X = t)$, the class-specific distribution of $Z$, is obtained by applying Bayes' rule (Lanza et al., 2013):

$$f(Z|X = t) = \frac{f(Z)P(X = t|Z)}{P(X = t)}. \tag{6}$$

The quantities $P(X = t|Z)$ and $P(X = t)$ can be obtained from the estimated LC model. The distribution of $Z$, $f(Z)$, can be approximated using the empirical distribution of this variable (Asparouhov & Muthén, 2014). That is, replacing the integral in Equation 5 by a sum over the $N$ sample units and replacing $f(Z)$ in Equation 6 by $\frac{1}{N}$ (Bakk & Vermunt, in press). This yields:

$$\mu_t = \sum_{i=1}^{N} Z_i \frac{P(X = t|Z_i)}{N \ P(X = t)}. \tag{7}$$

Simulation studies show that the estimated class-specific means obtained with this implementation of the LTB approach are unbiased as long as the relation between $X$ and $Z$ is linear-logistic (Asparouhov & Muthén, 2014; Bakk & Vermunt, in press). However, when the linearity does not hold the estimates are biased (Bakk & Vermunt, in press). This occurs for instance when the distal outcome has heteroskedastic errors. It turns out that larger the differences in the variances between classes, the larger the bias. We address this problem in more detail in section 4.

Lanza et al. (2013) did not discuss how to obtain SEs for the class-specific means, which are needed to make statistical inference possible. As a way out, Asparouhov and Muthén (2014) suggested obtaining approximate SEs by taking the square root of the within-class variance divided by the class-specific sample size; that is,

$$\sigma_t^2 = \sum_{i=1}^{N} (Z_i - \mu_t)^2 \frac{P(X = t|Z_i)}{NP(X = t)} \tag{8}$$

Simulation studies show that the approximate SE estimates underestimate the true sampling variability of the class-specific means (Asparouhov & Muthén, 2014; Bakk & Vermunt, in press), a problem that we address in section 5.

## The three-step LTB approach

While originally proposed as a simultaneous estimation procedure, the LTB approach can easily be transformed into a three-step estimation procedure similar to the one proposed by Vermunt (2010). This can be beneficial because it follows the logic of researchers, who often prefer to first establish a measurement model, and later associate it with one or more distal outcomes. Furthermore, in some situations the simultaneous LTB approach cannot be used; for example, when the sample used to estimate the LC measurement model does not (fully) overlap with the sample containing the distal outcomes of interest (Bakk & Vermunt, in press). The approach is also useful in situations where there is a large amount of missingness on the $Z$ variable of interest. In such situations the researcher can choose for a three-step approach, in order to estimate the classification model (step one) using more information.[1]

The three-step LTB approach can be implemented as follows. Steps one and two involve performing a standard LC analysis (without distal outcome) and assigning individuals to classes, whereas in step three the assigned class memberships are related to the external variables of interest while correcting for classification errors, followed by the calculation of the class-specific means (Vermunt, 2010; Bakk, Tekle, & Vermunt, 2013) (see Figure 2). Using this approach, the first two steps need to be performed only once. Step three is repeated for each distal outcome variable, while keeping the measurement model parameters and the resulting classifications fixed.

After estimating the step-one model (that includes only the indicator variables), as described in Equation 2, the units are assigned to the latent classes based on their posterior class membership probabilities: $P(X|\mathbf{Y}_i)$. During the assignment process a

[1]It should be mentioned that in the three-step approach we assume that missingness on the distal outcome is unrelated to the size of the classification errors. Since cases with missing values are removed from the analysis, this missingness is assumed to be MCAR, as is also the case in the simultaneous (original) LTB method.

new variable, $W_i$ is created, which equals the assigned class membership score for person $i$. Different assignment rules can be used, the best-known ones being modal and proportional assignment. Using modal assignment, each unit is assigned to a single class; namely, to the class for which the posterior membership probability is largest (Bolck, Croon, & Hagenaars, 2004), yielding what is called a hard partitioning. Using proportional assignment, each unit is assigned to each of the $T$ classes with a weight equal to $P(W_i = s|\mathbf{Y}_i) = P(X = s|\mathbf{Y}_i)$, leading to what is sometimes referred to as a soft partitioning (Dias & Vermunt, 2008). Irrespective of the assignment rule used, there will be classification errors unless all classifications are perfect. These errors can be quantified as the off-diagonal elements of the $T - by - T$ classification table with entries $P(W_i = s|X = t)$ (Bolck et al., 2004; Vermunt, 2010; Bakk et al., 2013).

In step three, a LC model is estimated with $W$ as a single indicator of class membership and with $Z$ as a covariate affecting the classes:

$$P(W_i = s|Z_i) = \sum_{t=1}^{T} P(X = t|Z_i)P(W_i = s|X = t), \tag{9}$$

where $P(W_i = s|X = t)$ is fixed to the estimated values from step two, and $P(X = t|Z_i)$ contains the logistic parameters to be estimated. Next, just as with the simultaneous LTB approach, with the estimated values for $P(X = t|Z_i)$, the class-specific means of $Z$ are calculated using Equation 7. This three-step LTB analysis is implemented in the Latent GOLD 5.0 program (Vermunt & Magidson, 2013).

### The LTB approach with a quadratic term

While not stated explicitly by Lanza et al. (2013), if $Z$ is normally distributed within classes with means $\mu_t$ and variances $\sigma_t^2$, the logistic model for $P(X = t|Z_i)$ is actually described by the discriminant function (Narsky & Porter, 2013, pp. 221-225). That is:

$$\log P(X = t|Z_i) = \log P(X = t) - \frac{1}{2}\log \sigma_t^2 - \frac{\mu_t^2}{2\sigma_t^2} + \frac{Z_i\mu_t}{\sigma_t^2} - \frac{Z_i^2}{2\sigma_t^2} + C,$$

where $C$ is a constant ($-\frac{1}{2}\log(2\pi)$). This implies that $log(P(X = t|Z_i)$ is a quadratic function of $Z$:

$$\log P(X = t|Z_i) = a_t + b_t Z_i + c_t Z_i^2.$$

Where

$$\begin{aligned} a_t &= \log P(X = t) - \frac{1}{2}\log \sigma_t^2 - \frac{\mu_t^2}{2\sigma_t^2}, \\ b_t &= \frac{\mu_t}{\sigma_t^2}, \\ c_t &= -\frac{1}{2\sigma_t^2}. \end{aligned}$$

Thus, using the logistic formulation, the model for $P(X = t|Z_i)$ equals:

$$P(X = t|Z_i) = \frac{\exp(\alpha_t + \beta_t Z_i + \gamma_t Z_i^2)}{\sum_{t'=1}^{T} \exp(\alpha_{t'} + \beta_{t'} Z_i + \gamma_t Z_i^2)}. \tag{10}$$

In a multinomial logistic regression, one would impose identifying constraints on the $\alpha_t$, $\beta_t$, and $\gamma_t$ terms, for example, set them equal to 0 for class $T$. This means $\alpha_t = a_t - a_T$, $\beta_t = b_t - b_T$, and $\gamma_t = c_t - c_T$.

Since $Z_i$ and $Z_i^2$ are correlated, the estimates of $P(X = t|Z_i)$ based on Equation 4 which does not contain the quadratic term and resulting estimates of the class-specific means will be biased unless the $\gamma_t$ term is equal to 0. It should be noted that $\gamma_t = 0$ when the variances $\sigma_t^2$ are equal across classes, that is, when errors are heteroskedastic. However, when variances are unequal across classes, the quadratic term should be included in the multinomial logistic regression model to obtain the correct estimates for $\mu_t$. By plugging in the estimates for $P(X = t|Z_i)$ obtained using Equation 10 into Equation 7, unbiased estimates of the class-specific means can also be obtained in the case of heteroskedastic errors.

## Alternative SE estimators

Another issue with regard to the LTB approach implementation that needs further attention is the problem of the underestimated standard errors reported by

Asparouhov and Muthén (2014) and Bakk and Vermunt (in press). This occurs because the approximate SEs do not take into account the sampling variability of the logistic parameters defining $P(X = t|Z_i)$ nor the fact that it conditions on the sample distribution of $Z$. A natural way to obtain SEs in such situations is by means of non-parametric resampling methods, that can be done by either a non-parametric bootstrapping or using the jackknife procedure.

### Bootstrap SEs for the LTB approach

For the simultaneous LTB approach, bootstrap SEs (Guan, 2003) are obtained as follows:

1. Draw $B$ random replication samples with replacement from the original data set.

2. Obtain the class-specific means of $Z$ for each of these $B$ bootstrap samples by applying the LTB approach.

3. Calculate the standard deviations of the class-specific means across the $B$ bootstrap replications. This yields the bootstrap SE estimates.

For the three-step LTB method, a choice can be made whether to bootstrap only the third step or also the first step.[2] In the latter case, one would also account for the uncertainty about the classification errors, which are fixed parameters in the step-three analysis. However, such a double bootstrap is much more costly and complex; that is, for each first-step bootstrap replication one should perform a full bootstrap of the third step. Because preliminary analyses showed that the step-three bootstrap is much more important for the SEs, we decided to bootstrap only the step-three parameters and evaluate the performance of this approach in the simulation study.

Bootstrapping the third step is similar to the non-parametric bootstrap described above. The main difference is that we sample from a data set with Z values and posterior class membership probabilities instead of Z values and Y values. That is:

[2]Note that it can seem intuitive to bootstrap only step 1, however this is not a solution to the problem at hand. The step three estimates need to be bootstrapped, because in this step the model for $P(Z|X)$ is obtained, thus the uncertainty around this estimates can be obtained using bootstrap SEs.

1. Draw $B$ random replication samples with replacement from the data set containing the distal outcome(s) of interest and the classification probabilities.

2. Obtain the class-specific means of $Z$ for each of these $B$ samples by applying the step-three LTB approach.

3. Calculate the standard deviation of the class-specific means across the $B$ replications. This gives the bootstrap SE estimates.

### Jackknife standard errors for the LTB approach

When using the jackknife approach, first the ML estimates of the parameters of interests (the class-specific means $\mu_t$) are obtained based on the full sample of size $N$. Following, the estimates are recalculated leaving out one observation $i$ at a time. The jackknife SE estimator is defined as follows:

$$SE(\mu_t) = \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}(\hat{\mu}_t - \hat{\mu}_t(-i))^2}, \tag{11}$$

where $\hat{\mu}_t$ is the original estimate and $\hat{\mu}_t(-i)$ the estimate when leaving out observation $i$.

In the three-step approach, the jackknife estimator can be applied in the step-three analysis, when the parameter estimates pertaining to the $Z - X$ relationship and the corresponding $\hat{\mu}_t$ are obtained.

## Simulation study

We performed a simulation study to evaluate the performance of the proposed adaptations of the LTB approach. These adaptations are the inclusion of a quadratic term, the use of bootstrap and jackknife SEs, and the three-step variant of LTB approach.

In the simulation study we compare the performance of the LTB approach with the different modifications to the BCH approach. This is done because the BCH approach is known to be the most robust stepwise estimator for relating LC membership to continuous distal outcomes (Bakk & Vermunt, in press).

Data sets were generated from a four-class model for eight dichotomous indicators.

The parameter settings were based on the LC application concerning psychological contract types described by Bakk et al. (2013). The class proportions were set to .50, .30, .10, and .10, similarly to those in this application. In class one, the probability of a positive answer was set to .80 for all indicators, and in class four to .20. In class two, it was set to .80 for the first four indicators and to .20 for the last four indicators, while in class three these settings were reversed.

The distal outcome variable $Z$ was specified to have means of -1, -0.5, 0.5, and 1 in the four classes. The variance of $Z$ was fixed to 1 in classes one and four, but varied in classes two and three. More specifically, in these two classes, the variance was set to 1, 4, 9, or 25, which corresponds to four different degrees of heteroskedasticity (none, small, medium, and large), and thus to different degrees of deviation from linearity of the logistic model for the association between $Z$ and the classes.

The second factor that was varied was the sample size, which was specified to be either 500 or 1000. For all combinations of heteroskedasticity and sample size conditions, 500 simulation replications were used. The simulation were done using the computer programs R (Venables, Smith, & the R Core Team, 2013) and Latent GOLD 5.0 (Vermunt & Magidson, 2013).

The LTB approach was applied with and without the quadratic term, in both its simultaneous and its three-step form, where the three-step variant was used with either modal or proportional class assignment. This yielded six different implementations of the LTB method. Each of these was combined with both the approximate SEs, jackknife SEs, and bootstrap-based SEs. For the bootstrap SEs we use $B = 1000$ bootstrap samples to obtain stable estimates. The BCH approach was applied with both modal and proportional assignment, using the sandwich standard error estimator, as proposed by Vermunt (2010).

The six different LTB implementations and two BCH implementations were compared with respect to parameter bias and relative efficiency. The efficiency of the LTB implementations was compared to proportional BCH, by dividing the simulation standard deviations of the estimates by those using BCH. Moreover, coverage rates of

the 95% confidence intervals obtained with the different SE estimators were compared. In the following the results are presented averaged over the classes (a weighted average is used, with the class size as weight).

In Table 1 we show the bias in the estimates of the class-specific means under the different conditions, averaged over 500 replications and over the four classes. As the first three rows of Table 1 show the linear LTB is an unbiased estimator only when there is no heteroskedasticity. In the conditions with heteroskedasticity, it can be seen that as heteroskedasticity increases, bias increases, with both three-step and simultaneous approach (up to values such as .29 with modal, and -.31 with proportional assignment for three-step approaches, and -.22 with simultaneous approach in the worth condition). However using the quadratic term unbiased estimates of the class specific means are obtained also in the high heteroskedasticity conditions. Comparing the estimates obtained with the LTB approach using the quadratic model and BCH we can see that results are comparable. The bias is the lowest using the simultaneous approach (between .000 and .003). However, the differences are negligible (only on the third decimals).

Next Table 2 shows the relative efficiency of the LTB estimators as compared to the BCH method. In almost all conditions the LTB estimators (when used with the correct model) are more efficient then BCH (having relative efficiency scores between .78 and 1.04). The three step LTB with modal assignment is the least efficient among the LTB approaches, with relative efficiency scores between .82 and 1.04. Furthermore the simultaneous LTB is the most efficient estimator in all conditions (with relative efficiency scores ranging between .79 and .92). This results are expected, since in general simultaneous estimators are more efficient then stepwise estimators.

Following, Table 3 shows the coverage rates obtained with the different estimators. When the correctly specified LTB approach is used with the approximate SEs the coverage rate is low (below 90%), even in the larger sample size conditions. The coverage rate obtained using the jackknife and bootstrap approaches is closer to the nominal 95%. The coverage rates obtained with the three-step LTB (with both modal and proportional assignment) is lower (between 90%- 96%) than the coverage using the

simultaneous approach (between 94%- 96%). However when the sample size is large enough even with the three-step implementation the coverage rate with both the jackknife and bootstrap estimators is close to the nominal rate. In all conditions the bootstrap estimator is somewhat better than the jackknife. However, the differences are very small. The coverage rates obtained using the BCH approach while close to the nominal 95% rate (between 90%- 95%), are somewhat smaller than the coverage obtained with the LTB approach using the bootstrap or jackknife SEs.

In summary, when the within-class errors of $Z$ are heteroskedastic, the quadratic term should be used to obtain unbiased estimates of the class-specific means. The three-step LTB approaches perform as well as the original simultaneous approach with regard to bias but they are somewhat less efficient. The bootstrapped and jackknifed standard errors yield coverage rates much closer to the nominal 95% rate than the approximate SEs. Furthermore, the LTB approach (both the simultaneous and the three-step implementation) proved to be more efficient than the BCH approach.

### An application: Predicting income from parents' social status classes

We will now illustrate the different LTB implementations with an application using data from the 1976 and 1977 General Social Survey, a cross-sectional survey of the English-speaking, non institutionalized adult population of the U.S.A., conducted by the National Opinion Research Center ("GENERAL SOCIAL SURVEY 1976-1977", 1977). We built a LC model for parents' social status using mother's education, father's education, and prestige of the father's job as indicators. Education was measured on a five-point scale ranging from 0 to 4, where 0 corresponds to 'lower than high school' and 4 to 'graduate'. Father's job prestige measured on a scale from 12 to 82 which recoded into three categories: low (12-36), medium (37-61), and high (62-82) prestige. As distal outcome variable we chose the real income of the respondent in thousand dollars increments.

In step 1, we fitted various LC models with the three indicators and selected the three-class model as best fitting model ($L^2 = 98.10$, $p = 0.65$, entropy $R^2 = 0.66$). The

bivariate residuals were also small. The parameters of the three-class model are presented in Table 4. Class one, the largest class, comprises of respondents whose parents had a lower social status, while class 2 corresponds to medium, and class 3, the smallest class, to high social status of the parents. Note that in the step-one analysis the full sample of 3029 respondents was used by keeping also cases with missing values on one or more of the indicators in the analysis.

Next we related the respondent's income to the latent classes using the one-step approach in which income is an additional indicator and the LTB approach, both with and without accounting for possible heteroskedastic errors. The LTB approach was used with the original and three-step implementation. The estimated class-specific means obtained with the different approaches are presented in Table 5. The estimates obtained using the four different LTB approaches are very similar. They show that the income is highest among those respondents whose parents have the highest social status, and lowest for those whose parents have the lowest social status. However, the estimates obtained with the one-step approach (with equal or unequal variances) are very different, especially for class 3, which is the result of the fact that its definition changes drastically (for details, see Table A1 and A2). Using unequal variances does also not solve the problem of completely changed class definitions.

These results obtained with this application are in line with previous research. That is, in conditions where the sample size is large and the separation between the classes is good, the LTB approach obtains unbiased estimates, even without the quadratic term (Bakk & Vermunt, in press, Table 5). However, in the one-step approach, the class solution can change to fit the distribution of the distal outcome, which is what happens in this example. It should be mentioned that the changing of the class solution using the one-step approach is only problematic when the model based on the indicators only is seen as the 'true' LC model, a situation that is common in practice. Note that while the simultaneous LTB approach yields similar class-specific means of income as the three-step approach, the class proportions and the class-specific response probabilities on the indicators change somewhat (see Table A3 and A4).

Table 6 presents the SE estimates obtained with the approximate jackknife, and bootstrap estimator for the four LTB approaches. The bootstrap and jackknife SE estimates are larger than the approximate estimates for both the original and three-step approaches with and without quadratic term. In this application, all SE estimators yield the same conclusion with regard to the significance of the income difference across classes.

## Discussion

The central idea of the LTB approach is that LC membership can be related to distal outcome variables by inverting the relationship, that is, by regressing the LC membership on the outcomes of interest. Based on the parameter estimates of this "reversed" model, the class-specific means of the distal outcome are calculated. The main benefit of the LTB approach is that no strong distributional assumptions have to be made about the distal outcomes. Furthermore this approach provides a direct test of the overall association between the LC variable and the distal outcomes.

In this paper, we presented three possible improvements of the LTB approach when applied to continuous distal outcomes. One of these is that we proposed incorporating the LTB approach into a three-step estimation procedure. The advantages of this procedure are that it follows more closely the logic of many researchers who apply latent class analysis by separating the measurement and structural steps. It also allows one to apply the LTB method in situations where the measurement model gives the error rates of a variable in an external dataset, for instance when obtaining medical diagnoses based on several error-prone tests and relating such (modal) diagnoses to external variables in a different, non-overlapping sample. In such situations, the simultaneous approach cannot be used. A disadvantage of the three-step approach is that when the uncertainty about the step one estimates is large (low entropy and/or sample size) the parameters are somewhat biased (Bakk et al., 2013; Vermunt, 2010; Asparouhov & Muthén, 2014). Furthermore, this implementation is asymptotically less efficient than simultaneous LTB.

We showed that omission of the quadratic term in the logistic model for the LC variable yields biased estimates of the class-specific means when the distal outcome has heteroskedastic errors; that is, when means and variances are not independent. In such situations, unbiased estimates of the class-specific means can be obtained by including the quadratic effect of the distal outcome on the classes.

We proposed using a jackknife or bootstrap-based SE estimator as an alternative to the currently used approximate estimator, originally proposed by Asparouhov and Muthén (2014). Contrary to the latter, the bootstrap and jackknife estimators can account for the overall sampling variability in the distal outcome and for the sampling fluctuation in the logistic parameters for the association between the LCs and the distal outcome.

The results of the simulation study showed that unbiased parameter estimates can be obtained if the heteroskedastic error is modeled using the quadratic term in the logistic regression of the LC variable on the outcome. This is the case with both the original and three-step implementation. However, if the quadratic term is not used the estimated class specific means are biased, when the error is heteroskedastic. The bias is somewhat larger with the three-step than with the simultaneous approach, although the difference is minor. At the same time the jackknife and bootstrap SEs obtain coverage rates close to the nominal 95% rate for both the original and three-step implementation of the LTB approach. It should be mentioned that the coverage rate obtained with the bootstrap estimator is slightly better than the one obtained with the jackknife estimator, although both are close to the nominal 95% level.

The real data example illustrated a situation where using a standard LC model with a distal outcome or the LTB approach makes a big difference. That is, in the one-step approach, the full LC solution changed and became hard to interpret, while with the LTB approach this problem did not occur. Furthermore, the different implementations of LTB yielded similar results, which can be explained by the large sample size and the strong measurement model.

We can conclude that the LTB method can be used with confidence for relating

LC membership to continuous distal outcomes. However, attention should be paid to whether a linear or a quadratic model should be used. The results of the simulation study showed that in situations where the Wald test of the quadratic effect is significant, this effect should be added to the model. It is recommended to use the jackknife or the bootstrap standard errors in all conditions. The proposed three-step LTB can be used with a minimum loss of efficiency whenever using the original approach is less practical or not feasible and the uncertainty about the step one model is not too high.

Further research might develop better tools for detecting whether a quadratic term needs to be added. For instance, the EPC-interest measure could be used, which quantifies how much the parameters of interest (here the class-specific means) change when adding the quadratic term (Oberski, 2014). Future research could also analyze the robustness of the LTB approach to other types of violations of implicit assumptions, such as when distal outcome distribution are skewed or show excess kurtosis. In such situations, third- and fourth-order terms might enter the logistic model for the classes. Although we suspect that the impact and importance of such terms will decline with their order, this remains a topic to be investigated. It is also recommended to analyze its performance with continuous distal outcomes coming from other distributions than normal.

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling*.

Bakk, Z., Tekle, F. T., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, 272-311.

Bakk, Z., & Vermunt, J. K. (in press). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*.

Bandeen-Roche, K., Miglioretti, D., & Zegger, P., S.L. Rathouz. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-86.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implication for overextraction of latent trajectory classes. *Psychological methods*, *8*(3), 338-363.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3-27.

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

De Cuyper, N., Rigotti, T., Witte, H. D., & Mohr, G. (2008). Balancing psychological contracts. Validation of a typology. *International Journal of Human Resource Management*, *19*, 543-561.

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*, 643-59.

General social survey 1976-1977. (1977).
doi: doi.org/10.3886/ICPSR07398.v1

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology*, 79-259.

Guan, W. (2003). From the help desk: Bootstrapped standard errors. *The Stata Journal*(1), 71.

Hagenaars, J. A. (1990). *Categorical longitudinal data- loglinear analysis of panel, trend and cohort data.* Newbury Park, CA:Sage.

Lanza, T. S., Tan, X., & Bray, C. B. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling*, *20:1*, 1-26.

McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park, CA:Sage.

Mulder, E., Vermunt, J., Brand, E., Bullens, R., & Van Merle, H. (2012). Recidivism in subgroups of serious juvenile offenders: Different profiles, different risks? *Criminal Behaviour and Mental Health*, *22*, 122-135.

Muthén, L., & Muthén, B. (1998-2012). *Mplus user's guide. Seventh edition.* Los Angeles, CA: Muthén and Muthén.

Narsky, I., & Porter, F. C. (2013). *Statistical analysis techniques in particle physics: Fits, density estimation and supervised learning.* John Willey and Sons, Inc., New JerseyC.

Oberski, D. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45-60.

Venables, W. N., Smith, D. M., & the R Core Team. (2013, April). *An introduction to R. notes on R: A programming environment for data analysis and graphics version 3.0.0.* Retrieved from `http://cran.r-project.org/doc/manuals/R-intro.pdf`

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

Vermunt, J. K., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic and Advanced and Syntax.* Belmont Massachusetts: Statistical Innovations Inc.

Table 1
*Bias in the estimates of class-specific means under eight simulation conditions, averaged over 500 replications and over the four classes.*

| | Condition: heteroskedasticity × sample size condition | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | None | | Small | | Medium | | Large | |
| Estimator | 500 | 1000 | 500 | 1000 | 500 | 1000 | 500 | 1000 |
| *LTB, linear model* | | | | | | | | |
| Modal | 0.005 | 0.004 | 0.020 | 0.012 | 0.063 | 0.074 | 0.103 | 0.293 |
| Proportional | 0.003 | 0.005 | 0.022 | 0.014 | 0.094 | 0.110 | 0.150 | -0.313 |
| Simultaneous | -0.003 | 0.003 | 0.092 | 0.075 | 0.237 | 0.226 | 0.239 | -0.222 |
| *LTB, quadratic model* | | | | | | | | |
| Modal | 0.002 | 0.006 | 0.012 | -0.002 | 0.013 | 0.003 | -0.010 | 0.007 |
| Proportional | 0.006 | 0.007 | 0.010 | -0.002 | 0.011 | 0.003 | -0.013 | -0.005 |
| Simultaneous | 0.001 | 0.000 | 0.003 | -0.003 | 0.002 | 0.000 | -0.003 | 0.000 |
| *BCH* | | | | | | | | |
| Modal | -0.005 | -0.002 | -0.008 | -0.002 | -0.006 | 0.002 | -0.002 | -0.005 |
| Proportional | -0.006 | -0.002 | -0.009 | -0.002 | -0.007 | 0.003 | -0.002 | -0.005 |

Table 2
*Relative efficiency compared with the proportional-assignment BCH estimator. Shown are simulation standard deviations of the estimates divided by those using BCH.*

| | Condition: heteroskedasticity × sample size | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | None | | Small | | Medium | | Large | |
| Estimator | 500 | 1000 | 500 | 1000 | 500 | 1000 | 500 | 1000 |
| *LTB, linear model vs. BCH* | | | | | | | | |
| Modal | 1.024 | 0.976 | 1.251 | 1.308 | 1.819 | 2.262 | 2.258 | 3.523 |
| Proportional | 0.966 | 0.965 | 1.407 | 1.489 | 2.279 | 2.902 | 3.199 | 4.718 |
| Simultaneous | 0.942 | 0.848 | 1.958 | 2.298 | 2.671 | 3.251 | 3.099 | 3.825 |
| *LTB, quadratic model vs. BCH* | | | | | | | | |
| Modal | 1.040 | 0.988 | 1.006 | 0.990 | 0.912 | 0.873 | 0.822 | 0.884 |
| Proportional | 0.983 | 0.976 | 0.958 | 0.981 | 0.898 | 0.866 | 0.780 | 0.852 |
| Simultaneous | 0.942 | 0.848 | 0.946 | 0.904 | 0.848 | 0.834 | 0.792 | 0.834 |

Table 3
*Coverage of 95% confidence intervals under the eight conditions. Performance is shown for three difference standard error estimators for LTB.*

| | Condition: heteroskedasticity × sample size | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | None | | Small | | Medium | | Large | |
| Estimator . | 500 | 1000 | 500 | 1000 | 500 | 1000 | 500 | 1000 |
| *LTB, linear model* | | | | | | | | |
| *Modal* | | | | | | | | |
| Approximate | 0.830 | 0.839 | 0.816 | 0.782 | 0.780 | 0.735 | 0.780 | 0.721 |
| Jackknife | 0.909 | 0.924 | 0.927 | 0.923 | 0.905 | 0.892 | 0.894 | 0.825 |
| Bootstrap | 0.909 | 0.927 | 0.932 | 0.927 | 0.920 | 0.913 | 0.923 | 0.846 |
| *Proportional* | | | | | | | | |
| Approximate | 0.843 | 0.856 | 0.781 | 0.770 | 0.691 | 0.642 | 0.672 | 0.594 |
| Jackknife | 0.903 | 0.926 | 0.913 | 0.923 | 0.845 | 0.822 | 0.791 | 0.720 |
| Bootstrap | 0.903 | 0.926 | 0.919 | 0.933 | 0.874 | 0.847 | 0.812 | 0.748 |
| *Simultaneous* | | | | | | | | |
| Approximate | 0.863 | 0.887 | 0.770 | 0.781 | 0.673 | 0.679 | 0.674 | 0.888 |
| Jackknife | 0.946 | 0.957 | 0.891 | 0.896 | 0.803 | 0.789 | 0.818 | 0.811 |
| Bootstrap | 0.961 | 0.961 | 0.914 | 0.910 | 0.824 | 0.806 | 0.829 | 0.824 |
| *LTB, quadratic model* | | | | | | | | |
| *Modal* | | | | | | | | |
| Approximate | 0.829 | 0.832 | 0.856 | 0.871 | 0.881 | 0.881 | 0.906 | 0.891 |
| Jackknife | 0.900 | 0.921 | 0.935 | 0.931 | 0.932 | 0.946 | 0.959 | 0.941 |
| Bootstrap | 0.906 | 0.921 | 0.936 | 0.936 | 0.940 | 0.947 | 0.962 | 0.942 |
| *Proportional* | | | | | | | | |
| Approximate | 0.836 | 0.856 | 0.872 | 0.885 | 0.888 | 0.907 | 0.922 | 0.904 |
| Jackknife | 0.904 | 0.924 | 0.932 | 0.930 | 0.942 | 0.948 | 0.955 | 0.936 |
| Bootstrap | 0.903 | 0.924 | 0.931 | 0.931 | 0.943 | 0.955 | 0.963 | 0.941 |
| *Simultaneous* | | | | | | | | |
| Approximate | 0.861 | 0.880 | 0.878 | 0.895 | 0.900 | 0.924 | 0.925 | 0.905 |
| Jackknife | 0.947 | 0.956 | 0.944 | 0.954 | 0.951 | 0.953 | 0.963 | 0.936 |
| Bootstrap | 0.954 | 0.965 | 0.952 | 0.956 | 0.959 | 0.959 | 0.967 | 0.944 |
| *BCH* | | | | | | | | |
| Modal | 0.907 | 0.919 | 0.927 | 0.929 | 0.932 | 0.930 | 0.945 | 0.954 |
| Proportional | 0.898 | 0.908 | 0.917 | 0.921 | 0.929 | 0.930 | 0.941 | 0.952 |

Table 4
*The 3-class model of parents social status. Class proportions and conditional response probabilities*

| | Low | Medium | High |
|---|---|---|---|
| Class Size | 0.69 | 0.24 | 0.07 |
| Father's job status | | | |
| low | 0.47 | 0.31 | 0.05 |
| medium | 0.53 | 0.67 | 0.46 |
| high | 0.00 | 0.02 | 0.49 |
| Mother's education | | | |
| < high school | 0.83 | 0.14 | 0.15 |
| high school | 0.16 | 0.78 | 0.44 |
| junior college | 0.00 | 0.03 | 0.01 |
| bachelor | 0.01 | 0.04 | 0.30 |
| graduate | 0.00 | 0.01 | 0.10 |
| Father's education | | | |
| < high school | 0.95 | 0.08 | 0.01 |
| high school | 0.05 | 0.86 | 0.12 |
| junior college | 0.00 | 0.00 | 0.05 |
| bachelor | 0.00 | 0.05 | 0.38 |
| graduate | 0.00 | 0.00 | 0.43 |

Note.The sample consists of 3029 respondents

Table 5
*Estimates of the class-specific means of income for the classes of parental social status obtained with the different methods*

| Method | Model | $\mu$ class 1 | $\mu$ class 2 | $\mu$ class 3 |
|---|---|---|---|---|
| Simultaneous LTB | linear | 25.37 | 36.49 | 44.16 |
| Simultaneous LTB | quadratic | 25.73 | 35.76 | 45.68 |
| 3-step LTB | linear | 26.43 | 37.88 | 44.40 |
| 3-step LTB | quadratic | 26.74 | 36.94 | 44.85 |
| Standard 1-step | equal variances | 25.36 | 36.62 | 162.59 |
| Standard 1-step | unequal variances | 21.33 | 26.98 | 69.73 |

Note. Sample size is 2767 obtained by excluding missing values on income

Table 6
*SE estimates of the class-specific means of income for the different LTB approaches with the approximate and bootstrap SE estimators*

| | SE estimator | Original | | 3-step | |
|---|---|---|---|---|---|
| | | linear | quadratic | linear | quadratic |
| Class1 | approximate | 0.90 | 0.58 | 0.55 | 0.44 |
| | bootstrap | 1.03 | 0.78 | 0.63 | 0.63 |
| | jackknife | 0.98 | 0.83 | 0.65 | 0.59 |
| Class2 | approximate | 1.16 | 1.57 | 1.45 | 1.31 |
| | bootstrap | 1.20 | 1.52 | 1.51 | 1.76 |
| | jackknife | 1.26 | 2.39 | 1.46 | 1.43 |
| Class3 | approximate | 2.62 | 3.42 | 2.81 | 3.04 |
| | bootstrap | 3.00 | 3.30 | 2.96 | 3.15 |
| | jackknife | 2.87 | 4.58 | 2.96 | 3.22 |

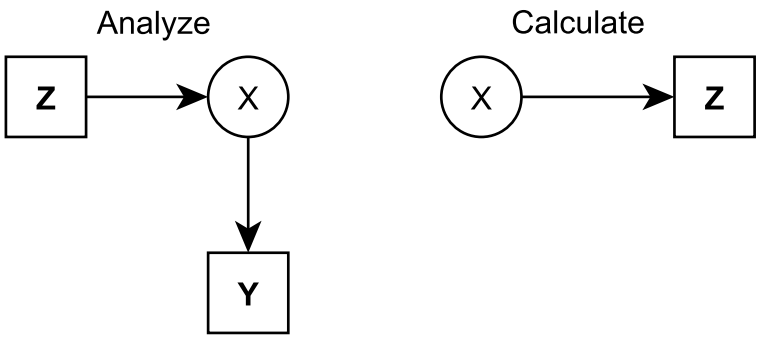


*Figure 1*. Original LTB approach

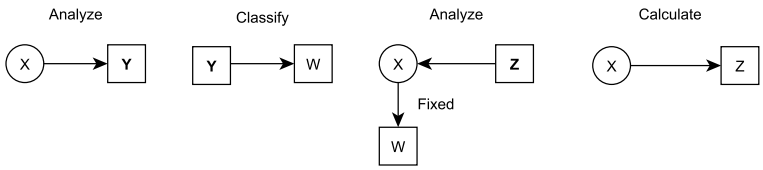


*Figure 2*. Three-step LTB approach

## Appendix 1: Additional tables related to our example application

Table A1
*The modified 3-class model of parents social status obtained using the one-step approach without accounting for heteroskedasticity*

| | | | |
|---|---|---|---|
| Class Size | 0.69 | 0.30 | 0.02 |
| Father's job status | | | |
| low | 0.47 | 0.23 | 0.20 |
| medium | 0.53 | 0.64 | 0.68 |
| high | 0.00 | 0.13 | 0.13 |
| Mother's education | | | |
| lt high school | 0.82 | 0.12 | 0.42 |
| high school | 0.17 | 0.72 | 0.42 |
| junior college | 0.00 | 0.03 | 0.00 |
| bachelor | 0.00 | 0.11 | 0.13 |
| graduate | 0.00 | 0.03 | 0.03 |
| Father's education | | | |
| lt high school | 0.93 | 0.08 | 0.52 |
| high school | 0.07 | 0.66 | 0.26 |
| junior college | 0.02 | 0.00 | 0.00 |
| bachelor | 0.00 | 0.13 | 0.10 |
| graduate | 0.00 | 0.11 | 0.13 |

Table A2
*The modified 3-class model of parents social status obtained using the one-step approach with accounting for heteroskedasticity*

| | | | |
|---|---|---|---|
| Class Size | 0.59 | 0.33 | 0.08 |
| Father's job status | | | |
| low | 0.48 | 0.33 | 0.05 |
| medium | 0.52 | 0.67 | 0.47 |
| high | 0.00 | 0.01 | 0.48 |
| Mother's education | | | |
| lt high school | 0.92 | 0.17 | 0.16 |
| high school | 0.08 | 0.75 | 0.47 |
| junior college | 0.00 | 0.03 | 0.01 |
| bachelor | 0.00 | 0.04 | 0.27 |
| graduate | 0.00 | 0.01 | 0.08 |
| Father's education | | | |
| lt high school | 0.96 | 0.32 | 0.01 |
| high school | 0.04 | 0.65 | 0.17 |
| junior college | 0.00 | 0.01 | 0.05 |
| bachelor | 0.00 | 0.03 | 0.37 |
| graduate | 0.00 | 0.00 | 0.40 |

Table A3
*The modified 3-class model of parents social status obtained using the simultaneous LTB approach without accounting for heteroskedasticity*

| | | | |
|---|---|---|---|
| Class Size | 0.59 | 0.33 | 0.08 |
| Father's job status | | | |
| low | 0.48 | 0.33 | 0.05 |
| medium | 0.52 | 0.67 | 0.47 |
| high | 0.00 | 0.01 | 0.48 |
| Mother's education | | | |
| lt high school | 0.92 | 0.17 | 0.16 |
| high school | 0.08 | 0.75 | 0.47 |
| junior college | 0.00 | 0.03 | 0.01 |
| bachelor | 0.00 | 0.04 | 0.27 |
| graduate | 0.00 | 0.01 | 0.08 |
| Father's education | | | |
| lt high school | 0.96 | 0.32 | 0.01 |
| high school | 0.04 | 0.65 | 0.17 |
| junior college | 0.00 | 0.01 | 0.05 |
| bachelor | 0.00 | 0.03 | 0.37 |
| graduate | 0.00 | 0.00 | 0.40 |

Table A4
*The modified 3-class model of parents social status obtained using the simultaneous LTB approach with accounting for heteroskedasticity*

| | | | |
|---|---|---|---|
| Class Size | 0.61 | 0.31 | 0.08 |
| Father's job status | | | |
| low | 0.47 | 0.33 | 0.06 |
| medium | 0.52 | 0.67 | 0.48 |
| high | 0.00 | 0.00 | 0.47 |
| Mother's education | | | |
| lt high school | 0.92 | 0.15 | 0.16 |
| high school | 0.08 | 0.78 | 0.47 |
| junior college | 0.00 | 0.03 | 0.01 |
| bachelor | 0.00 | 0.04 | 0.27 |
| graduate | 0.00 | 0.01 | 0.09 |
| Father's education | | | |
| lt high school | 0.95 | 0.31 | 0.02 |
| high school | 0.05 | 0.65 | 0.20 |
| junior college | 0.00 | 0.01 | 0.05 |
| bachelor | 0.00 | 0.03 | 0.36 |
| graduate | 0.00 | 0.00 | 0.38 |

## Appendix 2: Latent GOLD syntax for the LTB approach

In this appendix, we present the LG 5.0 syntax used for the real data example.

The following is the syntax for the original one-step LTB approach:

```
 options
   output
        parameters=effect  standarderrors=npbootstrap profile=LTB;
variables
      dependent fatherjob nominal, mothereduc nominal, fathereduc nominal;
       independent income;
      latent
          Cluster nominal 3;
      equations
   Cluster <- 1 + income ;
    fatherjob <- 1 + Cluster;
  mothereduc <-  1 + Cluster;
   fathereduc <- 1 + Cluster;
```

In the options, we indicate what output we would like to have. The command "profile=LTB" yields class-specific means for the independent variables, as well as their standard errors. As can been, in this example, the chosen type of standard error estimator is the nonparametric bootstrap SE (npbootstrap). This can also be jackknife or standard.

The subsection "variables" defines the indicators as dependent variables, the distal outcome as independent variable, and the latent class variable. The section "equations" defines the model of interest. The heteroskedastic model is obtained by changing the first line of the equations into

```
  Cluster <- 1 + income + income * income;
```

When using a step-three approach, one first estimates a LC model and saves the classification information and other variables of interest to an output file. In the third step, this output file is used as the data set to be analyzed. The class assignments are related to the distal outcome as follows:

```
 options
     step3 modal ml simultaneous;
    output
        parameters=effect  standarderrors=npbootstrap profile=LTB;
variables
   independent income;
   latent Cluster nominal posterior = ( Cluster#1 Cluster#2 Cluster#3 ) ;
equations
   Cluster <-\textless- 1 + income;
```

The choice of the type of three-step approach is defined by the command line “step3 modal ml simultaneous”. In the definition of the LC variable one specifies the variables in the data file containing the posterior classification probabilities from the first step: ’(Cluster1 Cluster2 Cluster3)’.