AMERICAN SOCIOLOGICAL ASSOCIATION

Sociological Methodology
43(1) 272–311
© American Sociological Association 2013
DOI: 10.1177/0081175012470644
http://sm.sagepub.com
SAGE

7

ESTIMATING THE ASSOCIATION BETWEEN LATENT CLASS MEMBERSHIP AND EXTERNAL VARIABLES USING BIAS-ADJUSTED THREE-STEP APPROACHES

*Zsuzsa Bakk**
*Fetene B. Tekle**
*Jeroen K. Vermunt**

Abstract

Latent class analysis is a clustering method that is nowadays widely used in social science research. Researchers applying latent class analysis will typically not only construct a typology based on a set of observed variables but also investigate how the encountered clusters are related to other, external variables. Although it is possible to incorporate such external variables into the latent class model itself, researchers usually prefer using a three-step approach. This is the approach wherein after establishing the latent class model for clustering (step 1), one obtains predictions for the class membership scores (step 2) and subsequently uses these predicted scores to assess the relationship between class membership and other variables (step 3). Bolck, Croon, and Hagenaars (2004) showed that this approach leads to severely downward-biased estimates of the strength of the relationships studied in step 3. These authors and later also Vermunt (2010) developed

*Tilburg University, The Netherlands

Corresponding Author:
Zsuzsa Bakk, Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
Email: z.bakk@tilburguniversity.edu

methods to correct for this bias. In the current study, we extended these correction methods to situations where class membership is not predicted but used as an explanatory variable in the third step, a situation widely encountered in social science applications. A simulation study tested the performance of the proposed correction methods, and their practical use was illustrated with real data examples. The results showed that also when the latent class variable is used as a predictor of external variables, the uncorrected three-step approach leads to severely biased estimates. The proposed correction methods perform well under conditions encountered in practice.

Keywords

latent class analysis, three-step approach, bias adjustment, covariates, distal outcomes, multiple latent variables

1. INTRODUCTION

The use of latent class analysis (LCA; Lazarsfeld and Henry 1968; Goodman 1974; McCutcheon 1987) is becoming more and more widespread in social science research, especially because of increasing modeling options and software availability. In its basic form, LCA is a statistical method for grouping units of analysis into clusters, that is, to identify subgroups that have similar values on a set of observed indicator variables. Examples of applications include the identification of types of political involvement (Hagenaars and Halman 1989), subgroups of substance abuse among youth (Kam 2011), types of psychological contract (De Cuyper et al. 2008), types of gender role attitudes (Yamaguchi 2000), and types of music consumers (Chan and Goldthorpe 2007).

Identifying the unknown subgroups or clusters is usually just the first step in an analysis since researchers are often also interested in the causes and/or consequences of the cluster membership. In other words, they may wish to relate the latent variable to covariates and distal outcomes. There are two possible ways to proceed with this latter extension, namely, using a one-step or a three-step approach. Using the one-step approach, the relation between the external variables of interest (covariates and/or distal outcomes) and the latent class variable is estimated simultaneously with the model for identifying the latent variable (Dayton and Macready 1988; Hagenaars 1990; Yamaguchi 2000; Muthén 2004). Using the other alternative, the three-step approach, first the underlying latent construct is identified based on a set of observed indicator variables, then individuals are assigned to latent classes, and subsequently the class assignments are used in further analyses (Bolck et al. 2004; Vermunt 2010). When all the model assumptions hold, the more complex one-step approach is better from a statistical point of view. However, most applied researchers prefer using the simpler three-step approach. De Cuyper et al. (2008) and Chan and Goldthorpe (2007) use such a three-step approach with covariates, as do Olino et al. (2010) and Morin et al. (2011) with distal outcomes.

One reason for using the three-step approach is that researchers see constructing a latent typology and investigating how the latent typology is related to external variables as two different steps in an analysis. For instance, in an LCA with distal outcomes, the latent classes will typically be risk groups (e.g., groups of youth delinquents based on delinquency histories or groups of persons with different lifestyles), and the distal outcomes are events in a later life stage (e.g., recidivism or health status). It is substantively difficult to argue that the distal outcomes should be included in the same model as the one that is used to identify the risk groups if one wishes to investigate the predictive validity of the latent classification.

Another argument for the three-step approach as opposed to the one-step is that in applications wherein a possibly large set of external variables is considered, the estimation procedure for the latter approach might fail because of the sparseness of the analyzed frequency table and the potentially large number of parameters (Goetghebeur et al. 2000; Huang and Bandeen-Roche 2004; Clark and Muthén 2009). For example, in a study by Mulder et al. (2012), the association of subgroups of recidivism with 70 possible distal outcomes was analyzed, which would be impossible using the one-step approach.

A related problem with the one-step approach is that the inclusion of covariates or distal outcomes can distort the class solution because additional assumptions are made that may be violated (Bauer and Curran 2003; Tofighi and Enders 2008; Huang et al. 2010; Petras and Masyn 2010). For example, the inclusion of a distal outcome requires specification of its within-class distribution, which if misspecified can distort the whole class solution. It may even happen that rather different class solutions are obtained when different distal outcomes are included separately in the model, though theoretically the latent classes should be based on the indicators and predict only the distal outcome.

Although there are many situations in which researchers may prefer the three-step LCA, the main disadvantage of this approach is that it yields severely downward-biased estimates of the association between class membership and external variables (Bolck et al. 2004; Vermunt 2010). Recently, several correction methods were developed to tackle this problem. Clark and Muthén (2009) proposed a correction method based on pseudo class draws from their posterior distribution, which, however, still maintains a relatively large bias in the log odds ratios of the association of the latent class variable with covariates. Petersen et al. (2012) developed a method based on a translation of the idea of Bartlett scores to the LCA context, which in the simulation study performed by the authors turned out to perform well. Bolck et al. (2004) developed a correction method that involves analyzing a reweighted frequency table and that can be used in three-step LCA with categorical covariates. Later Vermunt (2010) suggested a modification of this method, making it possible to obtain correct standard errors (*SE*s) and accommodate continuous covariates, and also introduced a more direct maximum likelihood (ML) correction method.

A limitation of the currently available adjustment methods for three-step LCA is that they were all developed and tested for the situation wherein class membership is treated as depending on the external variables. Moreover, all these methods were studied using models with only a single latent variable. However, applied researchers are often interested in a much broader use of the latent class solutions, and thus there should be correction methods available for a larger variety of modeling options. Given this gap in the literature, in the current article, we show how the three-step correction methods developed by Bolck et al. (2004) and Vermunt (2010) can be adapted to the situation in which the latent variable is a predictor of one or more distal outcomes, which may be categorical or continuous variables. We also pay attention to the situation in which the distal outcome itself is also a categorical latent variable, which implies that one should adjust for classification errors in both the predictor and the outcome variable.

The content of the article is outlined as follows. First we introduce the basic latent class model and discuss class assignment and quantification of the associated classification error. Then, the two classic ways of handling external variables in LCA will be presented (namely, the one-step and three-step approaches). Next, we discuss the correction methods developed by Bolck et al. (2004) and Vermunt (2010) for three-step LCA and show how these can be generalized for modeling the joint distribution of class membership and external variables, from where specific subcases can be derived. Subsequently, we check the performance of the different correction methods using a simulation study and illustrate them with real data applications.

2. LATENT CLASS MODELING AND CLASSIFICATION

2.1. *The Basic Latent Class Model*

Let us denote the categorical latent variable by X, a particular latent class by t, and the number of classes by T; as such we have $t = 1,2, \dots T$. Let Y_k represent one of the K manifest indicator variables, where $k = 1,2, \dots K$. Let $\mathbf{Y}$ be a vector containing a full response pattern and $\mathbf{y}$ its realization. A latent class model for the probability of observing response pattern $\mathbf{y}$ can be defined as follows:

$$P(\mathbf{Y}=\mathbf{y}) = \sum_t P(X=t)P(\mathbf{Y}=\mathbf{y}|X=t), \tag{1}$$

where $P(X=t)$represents the probability of belonging to class t and $P(\mathbf{Y}=\mathbf{y}|X=t)$ the probability of having response pattern $\mathbf{y}$ conditional on belonging to class t. As we can see from equation 1, the marginal probability of obtaining response pattern $\mathbf{y}$ is assumed to be a weighted average of the t class-specific probabilities.

In a classical LCA we assume local independence, which means that the K indicator variables are assumed to be mutually independent within each class t. This implies that the joint probability of a specific response pattern on the vector of indicator variables is the product of the item-specific probabilities:

$$P(\mathbf{Y}=\mathbf{y}|X=t) = \prod_k P(Y_k=y_\mathrm{k}|X=t). \tag{2}$$

Combining equations 1 and 2, we obtain the following:

$$P(\mathbf{Y}=\mathbf{y}) = \sum_t P(X=t) \prod_k P(Y_k=y_k|X=t). \tag{3}$$

The model parameters of interest are the class proportions $P(X=t)$ and the class-specific response probabilities $P(Y_k=y_k|X=t)$. These parameters are usually estimated by ML.

2.2. *Obtaining Latent Class Predictions*

While the true class memberships cannot be observed, the parameters of the measurement model described in equations 1 to 3 can be used to derive procedures for estimating these class memberships, that is, for assigning individuals to classes (Goodman 1974, 2007; Clogg 1981; Hagenaars 1990). The prediction is based on the posterior probability of belonging to class t given an observed response pattern $\mathbf{y}$, $P(X=t|\mathbf{Y}=\mathbf{y})$, which can be obtained by applying Bayes's theorem, that is,

$$P(X=t|\mathbf{Y}=\mathbf{y})=\frac{P(X=t)P(\mathbf{Y}=\mathbf{y}|X=t)}{P(\mathbf{Y}=\mathbf{y})}. \tag{4}$$

These posterior class membership probabilities provide information about the distribution over the T classes among individuals with response pattern $\mathbf{y}$, which reflects that persons having the same response pattern can belong to different classes. It is important to note that each individual belongs to only one class but that we do not know to which. Using the posterior class membership probabilities, different types of rules can be used for assigning subjects to classes, the most popular of which are modal and proportional assignment.

When using modal assignment, each individual is assigned to the class for which its posterior membership probability is the largest. Denoting the predicted class by W and subject i's response pattern by $\mathbf{y}_i$, the hard partitioning corresponding to modal assignment can be expressed as the following:

$$P(W=s|\mathbf{Y}=\mathbf{y}_i)=\begin{cases}1 & \text{if } P(X=s|\mathbf{Y}=\mathbf{y}_i)>P(X=t|\mathbf{Y}=\mathbf{y}_i)\ \forall\ s\neq t\\ 0 & \text{otherwise}\end{cases}.$$

An individual is assigned with probability or weight equal to 1 to the class with the largest posterior probability and with weight 0 to the other classes. Below we will also use the shorthand notation w_{is} for $P(W=s|\mathbf{Y}=\mathbf{y}_i)$.

To illustrate the class assignment, let us assume that we have a two-class model and that for a particular response pattern containing 20 respondents we find a probability of 0.8 of belonging to class 1, and of 0.2 of belonging to class 2. This means that 16 persons belong to class 1 and 4 to class 2. Under modal assignment, all 20 individuals will be assigned to class 1, which means that 4 will be misclassified (but we do not know which 4). This can be expressed as follows: $16\times(0)+\ 4\times(1) = 4$. It should be noted that modal assignment is optimal in the sense that the number of classification errors is smaller than with any other assignment rule.

An alternative to modal assignment is proportional assignment, which in the context of model-based clustering is referred to as a soft partitioning method (Dias and Vermunt 2008). An individual with the response pattern $\mathbf{y}_i$ will be assigned to each class s with a weight $P(W=s|\mathbf{Y}=\mathbf{y}_i)=P(X=s|\mathbf{Y}=\mathbf{y}_i)$, that is, with a weight equal to the posterior membership probability. In our example, this would mean that each of the 20 observations receive weights of .8 and .2 for belonging to the first and second class, respectively. In practice, this is achieved by creating an expanded data file with one record per class per respondent and by using the class membership probabilities as weights in subsequent analyses.

While at first glance it may seem that proportional assignment prevents introducing misclassifications, this is clearly not the case. In our example, the 16 persons belonging to class 1 receive a weight of .8 for class 1 instead of a weight of 1, which corresponds to a misclassification of .2, and the 4 persons belonging to class 2 receive a weight of .2 for class 2 instead of a weight of 1, which corresponds to a misclassification of .8. The total number of misclassifications for the data pattern concerned is therefore $16 \times (.2) + 4 \times (.8) = 6.4$.

Although modal and proportional assignment are the most common methods, it is also possible to use other rules. An example is the random assignment of individuals to classes based on the posterior class membership probabilities, which is in fact a stochastic version of the proportional assignment rule. The expected number of misclassification is the same under random and proportional assignment. A rule similar to modal assignment involves assigning individuals to class s if the posterior probability is larger than a specific value. For example, in a two-class model, one assigns an individual to class 1 if the posterior membership probability for this class is larger than .7 and otherwise to class 2. Compared to modal assignment, such a rule reduces the number of misclassifications into class 1 but increases the misclassifications into class 2.

It is clear that irrespective of the assignment method used, class assignments and true class scores will differ for some individuals (Hagenaars 1990; Bolck et al. 2004). As is shown in more detail below, the overall proportion of misclassifications can be obtained by averaging the misclassification probabilities of all data patterns. This overall classification error can be calculated irrespective of the assignment rule applied.

2.3. *Quantifying the Classification Errors*

The overall quality of the classification obtained from an LCA can be quantified by $P(W=s|X=t)$, that is, by the probability of a certain class assignment conditional on the true class. The larger the probabilities for $s = t$, the better the classification. Using the LCA parameters, this quantity can be obtained as follows:[1]

$$\begin{aligned} P(W=s|X=t) &= \sum_{\mathbf{y}} P(\mathbf{Y}=\mathbf{y}|X=t)P(W=s|\mathbf{Y}=\mathbf{y}) \\ &= \frac{\sum_{\mathbf{y}} P(\mathbf{Y}=\mathbf{y})P(X=t|\mathbf{Y}=\mathbf{y})P(W=s|\mathbf{Y}=\mathbf{y})}{P(X=t)}. \end{aligned} \tag{5}$$

In fact, the overall classification errors are obtained by averaging the classification errors for all possible response patterns. As indicated by Vermunt (2010), when the possible number of response patterns is very large, it is more convenient to estimate the classification errors by averaging over the patterns occurring in the sample, which involves replacing $P(\mathbf{Y} = \mathbf{y})$ by its empirical distribution:

$$P(W=s|X=t) = \frac{\frac{1}{N}\sum_{i} P(X=t|\mathbf{Y}=\mathbf{y}_i)w_{is}}{P(X=t)}, \tag{6}$$

where N is the sample size and, as indicated above, $w_{is} = P(W=s|\mathbf{Y}=\mathbf{y}_i)$. Below we will show how $P(W=s|X=t)$ is used in the correction methods for three-step LCA.

The concept of classification error is strongly related to the concept of separation between classes. The latter refers to how well the classes can be distinguished based on the available information on **Y.** More specifically, lower separation between classes corresponds to larger classification errors. Measures for class separation, and thus also for classification error, quantify how much the posterior membership probabilities $P(X=s|\mathbf{Y}=\mathbf{y}_i)$ deviate from uniform. For this purpose, one can use, among others, the principle of entropy: $-\sum_t P(X=t|\mathbf{Y}=\mathbf{y}) \log P(X=t|\mathbf{Y}=\mathbf{y})$. The proportional reduction of entropy when **Y** is available compared to the situation in which **Y** is unknown is a pseudo R-squared measure for class separation (Vermunt and Magidson 2005) and thus also for the quality of the classification of a sample.

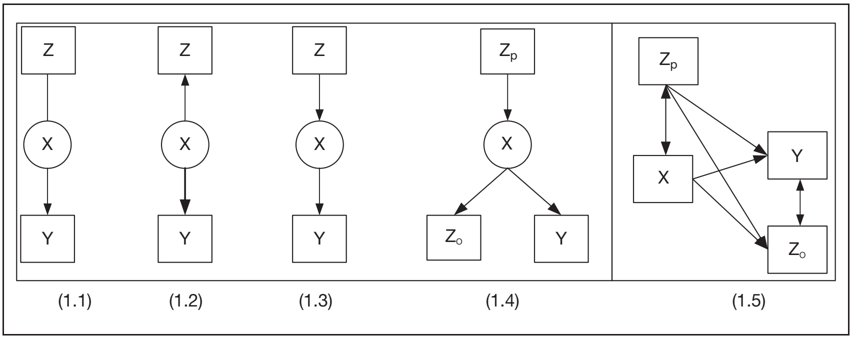

Figure 1. Types of associations between the latent variable (*X*), its indicators (**Y**), and other external variables (*Z*) that can be outcome variables (Z_o) or predictor variables (Z_p) of the latent variable.

Note: While 1.1 through 1.4 can be estimated using any of the methods discussed here, model 1.5 can be estimated only with the one-step approach.

3. LCA WITH EXTERNAL VARIABLES: THE TWO TRADITIONAL APPROACHES

There are a variety of ways in which external variables may play a role in an LCA; the most common ones are depicted in Figure 1 (1.1−1.5). We denote an external variable by *Z*, the latent variable by *X*, and the vector of indicators by **Y.** It should be noted that while the use of multiple latent variables is possible, for simplicity of exposition, in the main part of the current article, we focus on the situation of a single *X* and illustrate the possibility of extension to multiple latent variables in one of the empirical examples.

In its most general form, we can think of the latent class variable (*X*) being measured by its indicators (**Y**) and being associated with external variables (*Z*), without specifying a causal order between *X* and *Z* (see Figure 1.1). More specific cases are when *Z* is a distal outcome (see Figure 1.2), when *Z* is a predictor of *X* (see Figure 1.3), or when *Z* contains both predictors (Z_p) and distal outcomes (Z_o) (see Figure 1.4).

The most general form of an association between *X* and *Z*, without specifying a causal order (see Figure 1.1), involves modeling the joint probability of the three sets of variables as follows:

$$P(Z=z,X=t,\mathbf{Y}=\mathbf{y})=P(Z=z,X=t)P(\mathbf{Y}=\mathbf{y}|X=t). \qquad (7)$$

Note that in this expression we make the assumption that Z and Y are conditionally independent of one another given X. This means that Z is associated with X, but controlling for X it is not associated with the indicators. This is a rather standard assumption in latent variables models with external variables, which is moreover needed for the adjusted three-step approaches.

Based on the substantive theoretical arguments about the causal relationship between X and Z, the joint distribution in equation 7 can be adapted to accommodate specific cases. For instance, if we assume that the latent variable depends on the external variable, the relationship between X and Z can be analyzed using a model of the form (see Figure 1.3):

$$P(Z=z, X=t, \mathbf{Y}=\mathbf{y}) = P(Z=z)P(X=t|Z=z)P(\mathbf{Y}=\mathbf{y}|X=t).$$

Because the marginal distribution of Z is typically not of interest, it can be dropped, and the model can be defined as follows:

$$P(X=t, \mathbf{Y}=\mathbf{y}|Z=z) = P(X=t|Z=z)P(\mathbf{Y}=\mathbf{y}|X=t). \tag{8}$$

Another type of situation that is often of interest is when the latent variable is a predictor of the external variable (see Figure 1.2). In this case, we use a model of the form:[2]

$$P(Z=z, X=t, \mathbf{Y}=\mathbf{y}) = P(X=t)P(Z=z|X=t)P(\mathbf{Y}=\mathbf{y}|X=t) \tag{9}$$

When some of the Z variables are predictors and others outcomes (see Figure 1.4), the model becomes the following:

$$P(Z_o=z_o, X=t, \mathbf{Y}=\mathbf{y}|Z_p=z_p) = P(X=t|Z_p=z_p)$$
$$P(Z_o=z_o|X=t, Z_p=z_p)P(\mathbf{Y}=\mathbf{y}|X=t),$$

where $\mathrm{Z_o}$ is the distal outcome variable and $\mathrm{Z_p}$ a covariate. Note that the latter two models require the specification of the conditional distribution of Z (Z_o) to quantify the effect of X on Z. In the current article, we will use a normal distribution for continuous Z and a multinomial distribution for ordinal and nominal Z. The regression models used are linear, cumulative logistic, and multinomial logistic regression (Agresti 2002).

When the implied conditional independence assumption holds, each of the four variants described above can be investigated using either a one-step or a three-step procedure. However, when this is not the case, one may prefer using a one-step approach, in which it is possible to relax the assumption that Z and $\mathbf{Y}$ are conditionally independent given X (Huang and Bandeen-Roche 2004), contrary to the three-step approaches, where this is yet not possible.[3] Extensions of the standard latent class model using the one-step approach make it possible to include direct effects of covariates on indicators, or residual correlations between indicators and distal outcomes, as shown in Figure 1.5. Readers interested in such extensions are referred to the literature available on these models (Hagenaars 1988; Bandeen-Roche et al. 1997; Huang and Bandeen-Roche 2004). It should be mentioned that when the assumptions of conditional independence of Z and $\mathbf{Y}$ are violated, this can influence model parameters; there is a need to further investigate whether the three- or the one-step approach is more affected by this.

In the following we will restrict ourselves to the situation in which Z and $\mathbf{Y}$ can be assumed to be independent given X. We will show how the relevant models can be estimated using one-step LCA, standard three-step LCA, and bias-adjusted three-step LCA.

3.1. *One-Step Approach*

Using this approach, the external variables are incorporated in the latent class model, and the resulting extended model is estimated simultaneously with the measurement model. The extended model can be seen as being composed of two parts: the measurement model that comprises information on $\mathbf{Y}$ given X and the structural part that deals with the relationship between X and Z.

Both covariates (see Figure 1.3) and distal outcome variables (see Figure 1.2) can be included, possibly in combination with one another (see Figure 1.4), and the inclusion of direct effects of covariates on dependent variables is possible (see Figure 1.5). In situations wherein the class membership is used as a predictor of one or more external distal outcomes Z, the latter have a role similar to those of the indicator variables (Hagenaars 1990:135–42; Muthén 2004; Huang et al. 2010).

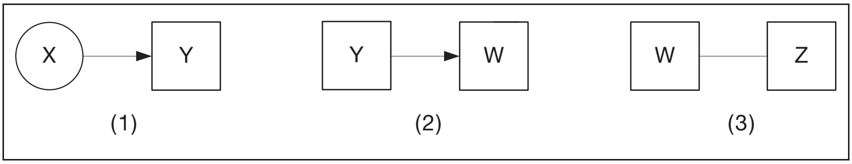

Figure 2. The steps of the standard three-step approach.

3.2. *The Standard Three-step Approach*

The method that is presented graphically in Figure 2 proceeds as follows. In the first step, the measurement model for the relationship between the latent variable and its indicators is built, as described by equation 3 and depicted in Figure 2.1. In the next step, using the information from the first step, subjects are assigned to latent classes based on their scores on the indicator variables, as depicted in Figure 2.2. In this process different assignment rules can be used, the most common ones being modal and proportional assignment. In the third step, the predicted class membership variable (*W*) is used in further analyses, implying analyzing the relationship between *W* and *Z* (see Figure 2.3).

Bolck et al. (2004) proved that the estimates of the log-odds ratios characterizing the relationship between *Z* and *W* will always be smaller than those characterizing the relationship between *Z* and *X* and proposed a correction method that can be used with categorical external predictors (see Figure 1.3). Their correction method was later extended by Vermunt (2010), who showed how to adjust for the downward bias in the *SE*s obtained by the initial method and how to include continuous covariates in the step-three model. Vermunt also proposed an ML-based correction method. In the following, we present these two correction methods and show how these can be generalized to the situation in which the class membership is a predictor instead of an outcome variable.

4. GENERALIZATION OF THE EXISTING CORRECTION METHODS FOR THREE-STEP LCA

While in the standard three-step procedure we estimate the relationship between *W* and *Z*, actually we are interested in the relationship between *X* and *Z* The key to the correction methods lies in the fact that it is possible to show how the *X*-*Z* distribution is related to the *W*-*Z* distribution.

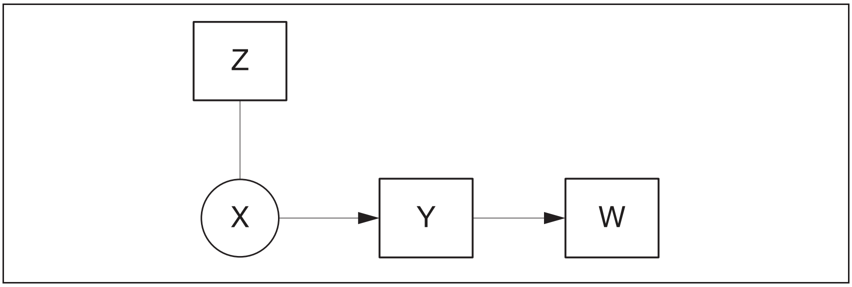

Figure 3. The relationship between variables *W, X,* **Y**, and *Z* in the three-step approach.

Let us first refer to Figure 3, which shows how the four (sets) of variables of interest are connected.

From the joint distribution of *X*, *Z*, *W*, and **Y**, we can derive the marginal distribution of *W* and *Z* by summing over all possible values of *X* and **Y**; that is,

$$\begin{aligned}P(W=s,Z=z) &= \sum_t \sum_{\mathbf{y}} P(X=t,\mathbf{Y}=\mathbf{y},\mathrm{Z}=\mathrm{z},W=s)\\ &= \sum_t P(X=t,Z=z) \sum_y P(\mathbf{Y}=\mathbf{y},W=s|X=t,Z=z)\\ &= \sum_t P(X=t,Z=z)\\ &\quad \sum_{\mathbf{y}} P(\mathbf{Y}=\mathbf{y}|X=t,Z=z)P(W=s|X=t,Z=z,\mathbf{Y}=\mathbf{y}).\end{aligned}$$

Given that *W* depends only on **Y** (as a consequence of the way the class assignment are obtained) and that **Y** is assumed to be independent of *Z* given *X* (the assumption depicted in Figure 1.1), and subsequently replacing $P(\mathbf{Y}=\mathbf{y}|X=t)$ by $P(\mathbf{Y}=\mathbf{y})P(X=t|\mathbf{Y}=\mathbf{y})/P(X=t)$ using Bayes theorem, we obtain:

$$\begin{aligned}P(W=s,Z=z) &= \sum_t P(X=t,Z=z)\frac{\sum_{\mathbf{y}} P(\mathbf{Y}=\mathbf{y})P(X=t|\mathbf{Y}=\mathbf{y})P(W=s|\mathbf{Y}=\mathbf{y})}{P(X=t)}\\ &= \sum_t P(X=t,Z=z)P(W=s|X=t). \qquad (10)\end{aligned}$$

The last substitution follows from the definition presented in equation 5. As can be seen from equation 10, the entries in the W and Z distribution are weighted sums of the entries in the X and Z distribution, where the weights are the misclassification probabilities $P(W=s|X=t)$. This suggests that the relationship between X and Z can be obtained by adjusting the relationship between W and Z for the misclassification probabilities $P(W=s|X=t)$.

The correction methods developed by Bolck et al. (2004) and Vermunt (2010) are based on an equality similar to the one described in equation 10. The difference is that these concern the relationship between the conditional distributions of X given Z and W given Z, so the situation where Z is a covariate and X is the outcome. As we have shown above in equation 10, the correction methods can also be applied to the joint distribution of X and Z. From this joint distribution the conditional distribution of Z given X can be obtained when the latent variable X is considered to be a predictor of external variable Z. The extension of the methods lies on the realization that the classification error depends only on the measurement model. The consequence of this is that irrespective of the role of X and Z in describing their mutual relationship, the adjustments remain the same. The same type of adjustments can also be used with multiple latent variables, as we will discuss shortly in a later section.

4.1. *The Three-Step ML Approach*

The ML-based correction method introduced by Vermunt (2010) involves defining a latent class model with one or more covariates (Z) affecting the latent variable X and with the predicted class membership W as the single indicator of the underlying latent variable X. An important difference compared to a standard LCA is that the conditional response probabilities $P(W=s|X=t)$ are not estimated but fixed to their estimated values from the previous step.

Vermunt's (2010) procedure can easily be adapted for the modeling of the joint distribution of X and Z or the conditional distribution of Z given X. As can be seen from equation 10, even if we have information only on Z and W and if $P(W=s|X=t)$ is known, it is possible to specify a (latent class) model yielding information on the association between X and Z. This requires using W as an indicator of X and defining the form of the X-Z distributions. Equation 10 can also be reexpressed as follows:

$$P(W=s, Z=z) = \sum_t P(X=t)P(Z=z|X=t)P(W=s|X=t), \tag{11}$$

corresponding to the situation in which X is a predictor of Z. Note that this yields a latent class model with two indicators, Z and W, where W comprises all the information on the classification from the first two steps. An assumption underlying this model is that Z and W are conditionally independent given X, which is in agreement with the structure depicted in Figure 3 and is necessary for all of the currently existing three-step approaches. What is also required is that one specifies the distributional form of $P(Z=z|X=t)$.

The parameters of the model in equation 11 can be estimated by maximizing the following log likelihood function:

$$\text{LogL}_{\text{ML}} = \sum_i \log \sum_t P(X=t)P(Z=z|X=t)P(W=s|X=t). \tag{12}$$

This can be achieved with any software for LCA that can accommodate parameters fixed to some specific values. We fix $P(W=s|X=t)$ to the estimates from step 2.

The possibility of using Z variables of different scale types requires that one should be able to specify an appropriate distribution for Z. Logical choices are a normal distribution for continuous Z, a multinomial distribution for nominal or ordinal Z, a Poisson distribution for count Z, and so forth.

4.2. *The Bolck-Croon-Hagenaars (BCH) Approach*

The ML correction method described above uses the classification errors from step 2 directly in a latent class model for W and Z. In contrast, the solution developed by Bolck et al. (2004) for categorical external predictor variables—which we refer to as the BCH approach—involves reexpressing the relationship described in equation 10 as follows:

$$P(X=t, Z=z) = \sum_s P(W=s, Z=z)d^*_{st}, \tag{13}$$

where d^*_{st} represents an element of the inverted T-by-T matrix $\mathbf{D}$ with elements $P(W=s|X=t)$.[4] In other words, if we weight the W-Z distribution by the inverse of the classification errors, we obtain the distribution we are interested in. Bolck et al. (2004) proposed using this relation, which applies at the population level, to reweight the data on W and Z

(the frequency table with observed counts n_{zs}). As shown by Vermunt (2010), their approach involves maximizing the following pseudo (or weighted) log likelihood function:

$$\begin{aligned} LogL_{BCH} &= \sum_z \sum_s n_{zs} \sum_t d^*_{st} \log P(X=t, Z=z) \\ &= \sum_z \sum_t n^*_{zt} \log P(X=t, Z=z), \end{aligned} \tag{14}$$

where the $n^*_{zt} = \sum_s n_{zs} d^*_{st}$ are the reweighted frequencies used to estimate the relationship between X and Z.

4.3. *The Modified BCH Approach*

Vermunt (2010) highlighted three shortcomings of the BCH method: Only categorical predictors can be used, *SE*s are underestimated, and the method needs a tedious data preparation stage that has to be repeated for each external variable. To solve these issues, the author proposed a modification to the BCH method consisting in reexpressing the pseudo log likelihood function in terms of individual observations. That is,

$$\begin{aligned} LogL_{BCH} &= \sum_i \sum_s w_{is} \sum_t d^*_{st} \log P(X=t, Z=z_i) \\ &= \sum_i \sum_t w^*_{it} \log P(X=t, Z=z_i), \end{aligned} \tag{15}$$

where w_{is} is a class assignment weight and $w^*_{it} = \sum_s w_{is} d^*_{st}$. Note that the standard three-step procedure involves using the nonreweighted w_{is} in the third step. To apply this modified BCH method, an expanded data file has to be created containing T records for each subject with X values t = 1,2,3 . . . T and weights w^*_{it}. This weighted data set can be analyzed with standard methods.

While equation 15 shows how to estimate parameters of the joint distribution of X and Z, it can be modified for the estimation of the conditional distribution of Z given X as follows:

$$\begin{aligned} LogL_{BCH} &= \sum_i \sum_t w^*_{it} \log P(X=t) P(Z=z|X=t) \\ &= \sum_i \sum_t w^*_{it} \log P(X=t) \ + \ \sum_i \sum_t w^*_{it} \log P(Z=z|X=t). \end{aligned} \tag{16}$$

Because the first term does not contain parameters of interest, it can be ignored and we can just maximize a pseudo log likelihood function based on the second term. Note that this formulation makes it possible to apply the BCH method to external variables of any scale type and thus also with continuous and ordinal *Z* variables. By applying a robust or sandwich variance estimator, one can prevent *SE*s' being underestimated as is the case with the original BCH approach. The robust variance-covariance matrix of the parameters is the inverse of the matrix obtained by ''sandwiching'' the Hessian by the average outer product of gradients for the independent observations (Skinner, Holt, and Smith 1989).

4.4. *ML Adjustment with Multiple Latent Variables*

For the simplicity of exposition, so far we have focused on the situation in which the step 3 latent class model of interest contains only one latent variable. However, both the ML and BCH methods can easily be extended to be applicable with multiple latent variables. We will illustrate this for the somewhat simpler ML approach.

Suppose one is interested in the association between latent variables X_1 and X_2. A stepwise modeling approach implies that one performs a separate LCA for each of these two latent variables and obtains class assignments W_1 and W_2. Implicitly, this means that an additional assumption is made, namely, that the indicators used in the model for X_1 are independent of X_2 conditionally on X_1 and vice versa. Given these assumptions are met, it is no problem to estimate the measurement models separately. The relationship between the joint distribution of the assigned class memberships and the true class memberships can be expressed in a way similar to equation 10, as follows:

$$\begin{aligned} P(W_1 = s_1, W_2 = s_2) = \sum_{t1} \sum_{t2} & P(X_1 = t_1, X_2 = t_2) \\ & P(W_1 = s_1 | X_1 = t_1) P(W_2 = s_2 | X_2 = t_2). \end{aligned} \tag{17}$$

This is a latent class model that can be estimated using LCA packages that support the use of multiple latent variables—here X_1 and X_2—and fixed value parameters—here $P(W_1 = s_1 | X_1 = t_1)$ and $P(W_2 = s_2 | X_2 = t_2)$.

As shown for the *X-Z* association, rather than modeling the joint distribution of X_1 and X_2, it is also possible to model the conditional distribution—$P(X_2 = t_2 | X_1 = t_1)$—and also when observed predictors are included in the model—$P(X_2 = t_2 | X_1 = t_1, Z = z)$. Moreover, extension to more than two latent variables is straightforward. The ML method with multiple latent variables can be used in any LCA package that supports the use of multiple latent variables and fixed value parameters. We illustrate the use of this method with our second real data example.

The generalized correction methods introduced above will be tested in the following with a simulation study and illustrated with two real data examples. For ease of readability, the simulation study focuses on the situation with one independent latent variable and one dependent external variable. The extension to more complex models is shown using the examples. To show the ease of use and applicability with the real data example, the syntax used in Latent GOLD (Vermunt and Magidson 2005, 2008) will be included as well. Since Vermunt (2010) showed that the *SE*s are underestimated using the original BCH method, here we will use only the modified BCH method with robust *SE*s.

5. SIMULATION STUDY

5.1. *Design*

A simulation study was conducted to check the quality of the proposed adjusted three-step LCA methods in situations in which the latent variable is treated as a predictor of one or more external variables (distal outcomes). In the simulation study, the BCH and ML correction methods were compared with the one-step and the standard three-step approaches. A method can be considered to perform well when the parameter estimates are unbiased and their variation is small, and in general the estimates are accurate. In the simulation study we will manipulate two key factors: the separation between classes (which as explained earlier is strongly related to the size of the classification error)[5] and the sample size, both of which have been found to affect the performance of the correction methods when the three-step LCA involved prediction of class membership using external variables (Vermunt 2010). Separation between classes is manipulated via the strength of the relationship between the classes and the indicators. Other conditions that could have been varied are number of items, number of item categories, and class sizes, but these are all conditions that basically affect the separation between classes. To keep the simulation simple and manageable, we decided to manipulate class separation only via the class-item association.

We tested the performance of the correction methods for three types of distal outcomes, that is, for *Z* nominal, ordinal, or continuous. Two conditions were used for the strength of the *X*-*Z* relationship, corresponding to a weaker and a stronger effect of *X* on *Z*. Data were generated from the full (*X, Y, Z*) model. In the following, the population values for all the parts of the model are provided.

The population model we used is a three-class model for six dichotomous response variables and a single distal outcome variable. The profile of the classes is as follows: Class 1 is likely to give the high response on all indicators, class 2 scores high on the first three indicators and low on the last three, and class 3 is likely to give the low response on all indicators. The separation between classes was manipulated by changing the conditional response probabilities for the indicators. The probability for the likely response was set to .70, .80, and .90, corresponding to a (very) low, middle, and high separation between classes. These settings correspond with entropy based R^2 values of .36, .65, and .90, respectively. In the following we will refer to these conditions as the low, mid, and high separation condition. Sample size is also important because it affects the accuracy of the estimates. The three sample sizes used were 500, 1,000, and 10,000. Note that a class separation of .36 is in fact extremely low and a sample size of 10,000 is rather large.

We used three types of outcome variables, a trichotomous nominal, a trichotomous ordinal, and a continuous outcome, which we modeled using a multinomial logit, a cumulative logit, and a linear model, respectively, with the first class and the first category of the outcome variable as the reference category.

For the nominal outcome, the condition with a strong effect of *X* on *Z* was obtained by setting the intercepts α_2 and α_3 to -2.08 and the effect parameters to 3.87 (β_{22}), 3.17 (β_{23}), 2.08 (β_{32}), and 2.08 (β_{33}), where the first index refers to the distal outcome category and the second to the class. Note that this setup yields some probabilities close to 0, which can cause estimation problems, as we will see in the Results section. For the condition with a weaker effect of *X* on *Z* we set both intercepts equal to -1.098, β_{22} to 2.01, β_{23} to 1.50, and β_{32} and β_{33} to 1.09.

For the ordinal outcome variable, in the high effect condition the thresholds were set to 2.94 (α_2) and 1.55 (α_3) and the effect parameters to −1.55 (β_2) and −4.33 (β_3) for classes 2 and 3, respectively. This setup also yields some probabilities close to 0. In the low effect condition, the thresholds were set to 2.74 (α_2) and 1.82 (α_3) and the effect parameters to −1.23 (β_2) and −3.01(β_3).

For the continuous outcome variable, in the strong effect condition, we set the class-specific means to −1, 0, and 1 (corresponding with an intercept of −1 and slopes of 1 and 2) and the error variance to 1. In the weak effect condition, we set the class-specific means equal to −0.2, 0, and 0.2 and kept the same error variance.

For the simulation study and the real data application, two computer programs were used: Latent GOLD (Vermunt and Magidson 2005, 2008) and R. In Latent GOLD we simulated the data, set up the measurement model, saved the scores on the posterior class assignment, and ran all the correction methods with both modal and proportional assignment. We used R to construct the **D** matrix and compute its inverse and to create the expanded data matrix containing the relevant weights (R code is available in appendix A). The **D** matrix was computed using equation 6, that is, using the empirical distribution of the responses. For each of the 54 conditions, which were obtained by crossing the three separation, three sample size, three types of external variable, and two effect size conditions, we used 500 replications.

5.2. *Results*

The results are presented both averaged across conditions and separately for some of the conditions. We pay attention to parameter bias (measured by comparing the average estimated value with the true values), efficiency (measured by the standard deviation [*SD*] across replications), and the bias in the estimated *SE*s (measured by comparing the average estimated *SE* with the *SD* across replications).

Before looking at these figures, we would like to present an important unanticipated result for the BCH method when applied with a nominal or an ordinal outcome variable *Z*. Some of the replications turned out to contain negative cell frequencies in the adjusted *X-Z* frequency table, in which case the corresponding multinomial distribution is not defined. This happened mainly in the least favorable condition coupling a low-class separation (large classification errors) with a small sample size (large sampling fluctuation). The possibility of such a failure of the BCH method is an important new result because it was not reported by

Table 1. Number of Excluded Replications for the Nominal and Ordinal Outcome Variable due to Negative Frequencies or Boundary Solutions

Sample Size	Separation Level	Correction Methods	One-step Maximum Likelihood
Nominal—strong *X-Z* effect			
500	Low	63	200
1,000	Low	59	59
500	Mid	4	1
1,000	Mid	1	0
Nominal—weak *X-Z* effect			
500	Low	9	46
1,000	Low	5	4
Ordinal—strong *X-Z* effect			
500	Low	20	28
1,000	Low	18	0

Bolck et al. (2004) or Vermunt (2010). While an ad hoc solution could be to fix the probabilities corresponding to negative counts to zero, we decided to exclude replications with negative frequencies from the results reported below. In the replication samples wherein the BCH method gave negative frequencies, the three-step ML method gave logit coefficients going to plus or minus infinity, corresponding to boundary solutions. Boundary solutions also occurred with the one-step ML method in the low separation and low sample size conditions. The replications with negative frequencies and boundary solutions were excluded from further analysis. Table 1 provides information on the number of excluded replications per condition.

Table 2 presents the results averaged over all sample sizes and separation levels for one parameter per outcome variable. It reports the average estimate, average *SE*, and *SD*s of estimates for each method. As can be seen, the proportional standard method has the largest bias. When averaged across conditions, we can see that the correction methods still slightly underestimate the parameters. The bias is less than 5 percent for the continuous and ordinal outcome variable and close to 10 percent for the nominal outcome variable. As shown below, bias varies strongly across separation and sample size conditions (is larger with low separation and small sample size and absent with higher separation and large sample size). As expected, when estimating a correctly specified model, the one-step approach yields a good approximation of the

Table 2. Average Estimate of One Selected β Parameter, and Its Average Estimated Standard Error (*SE*) and Standard Deviation (*SD*) across Replications Aggregated over the Nine Separation and Sample Size Conditions for All Three Types of Outcome Variables (for Strong and Weak *X*-*Z* Association)

	Nominal			Ordinal			Continuous		
	$\beta_{23} = 3.17$			$\beta_2 = -1.55$			$\beta_1 = 1.00$		
Method	Estimate	*SE*	*SD*	Estimate	*SE*	*SD*	Estimate	*SE*	*SD*
One-step ML	3.22	0.55	0.50	−1.58	0.27	0.27	1.00	0.07	0.07
Modal standard	2.06	0.22	0.23	−1.18	0.15	0.20	0.80	0.06	0.08
Proportional standard	1.73	0.21	0.18	−1.07	0.15	0.16	0.72	0.06	0.07
Modal BCH	2.97	0.50	0.51	−1.52	0.26	0.33	0.97	0.07	0.09
Proportional BCH	2.98	0.56	0.48	−1.55	0.25	0.32	0.97	0.07	0.10
Modal ML	2.97	0.50	0.51	−1.52	0.25	0.31	0.97	0.07	0.09
Proportional ML	2.98	0.83	0.51	−1.53	0.30	0.30	0.97	0.07	0.08
	$\beta_{23} = 1.50$			$\beta_2 = -1.23$			$\beta_1 = 0.20$		
One-step ML	1.53	0.40	0.35	−1.26	0.25	0.26	0.20	0.07	0.06
Modal standard	1.05	0.21	0.22	−0.97	0.15	0.17	0.16	0.05	0.05
Proportional standard	0.90	0.19	0.15	−0.88	0.14	0.14	0.14	0.04	0.05
Modal BCH	1.42	0.31	0.32	−1.22	0.23	0.26	0.19	0.06	0.06
Proportional BCH	1.42	0.29	0.30	−1.24	0.22	0.26	0.20	0.06	0.06
Modal ML	1.42	0.31	0.32	−1.22	0.22	0.26	0.19	0.06	0.06
Proportional ML	1.42	0.41	0.30	−1.23	0.28	0.26	0.20	0.07	0.06

Note: ML = maximum likelihood; BCH = Bolck-Croon-Hagenaars.

parameter of interest (bias less than 5 percent). It should be mentioned that with the exception of the low sample size and low separation between classes conditions, the correction methods perform well, having bias less than 5 percent for all outcome variables as well.

As can be seen from the *SD*s across replications, the correction methods perform similarly, in terms of efficiency, with each other and the one-step ML method. Comparison of the average estimated *SE* across replications with the *SD* of the parameter estimate across replications shows that the correction methods slightly underestimate the *SE*, with the exception of the proportional ML method, which overestimates the *SE* for the nominal outcome variable. Overall, the difference between the *SE*s and *SD*s is smallest for the proportional ML method, except for the nominal outcome variable.

When we look separately at the parameter estimates in each of the investigated conditions, we can see large differences between conditions. As seen in Tables 3 and 4, the one-step ML method obtains estimates close to the true values, with the exception of the combination of small sample size and low separation between classes, where it tends to overestimate the parameter. For all outcome variables, the correction methods perform poorly in the low separation and small sample size conditions, a result that is similar to the one reported by Vermunt (2010). Note that this applies to each of the three types of response variables and both for a strong and a weak *X*-*Z* association. The reason for this bad performance with low separation and small sample size is that in this situation the differences between classes are overestimated in the first step, yielding an underestimate of (too optimistic) the classification error, and as a consequence a too moderate adjustment by the BCH and ML correction methods. In the middle and high separation conditions, the correction methods perform well. While in the high separation conditions the performance of the correction methods using modal versus proportional assignment did not differ, in the lower separation condition this is not the case. With middle separation and especially with low separation between classes, the estimates obtained with the proportional assignment approximated better the true values than the ones obtained using modal assignment for all three types of outcome variables.

Table 5 reports the average *SE* and *SD* across replications for one selected parameter (from the condition with a nominal *Z* variable weakly related to the classes) for the nine sample size and class separation combinations. As we can see, in the conditions with a low separation and a

Table 3. Average Estimate of Selected β Parameter Separately for Each of the Nine Separation and Sample Size Conditions for All Three Types of Outcome Variables for Strong *X*-*Z* Association

Separation Level	Low			Medium			High		
Sample Size	500	1,000	10,000	500	1,000	10,000	500	1,000	10,000
Method				Nominal Z: $\beta_{23} = 3.17$					
One-step ML	3.28	3.24	3.11	3.36	3.22	3.17	3.20	3.18	3.16
Modal standard	1.14	1.20	1.34	2.07	2.11	2.13	2.85	2.83	2.82
Proportional standard	0.90	0.87	0.85	1.69	1.69	1.68	2.63	2.61	2.60
Modal BCH and ML	2.03	2.40	3.16	3.17	3.24	3.17	3.21	3.18	3.15
Proportional BCH and ML	2.13	2.58	3.11	3.11	3.15	3.16	3.18	3.16	3.15
				Ordinal Z: $\beta_2 = -1.56$					
One-step ML	−1.64	−1.61	−1.56	−1.60	−1.57	−1.56	−1.59	−1.56	−1.56
Modal standard	−0.83	−0.84	−0.84	−1.22	−1.21	−1.22	−1.51	−1.49	−1.48
Proportional standard	−0.67	−0.65	−0.65	−1.11	−1.09	−1.09	−1.46	−1.44	−1.44
Modal BCH	−1.40	−1.41	−1.54	−1.55	−1.52	−1.55	−1.58	−1.56	−1.56
Proportional BCH	−1.49	−1.50	−1.56	−1.57	−1.55	−1.56	−1.58	−1.56	−1.56
Modal ML	−1.39	−1.42	−1.54	−1.56	−1.52	−1.55	−1.58	−1.56	−1.56
Proportional ML	−1.43	−1.42	−1.56	−1.57	−1.56	−1.55	−1.58	−1.56	−1.56
				Continuous Z: $\beta_1 = 1.00$					
One-step ML	1.00	1.00	0.99	0.99	1.00	1.00	1.03	1.00	1.00
Modal standard	0.57	0.59	0.60	0.81	0.83	0.84	0.96	0.96	0.96
Proportional standard	0.47	0.47	0.45	0.74	0.75	0.75	0.94	0.94	0.94
Modal BCH	0.85	0.91	0.98	0.97	0.99	1.00	1.00	1.00	1.00
Proportional BCH	0.88	0.94	0.98	0.97	1.00	1.00	1.00	1.00	1.00
Modal ML	0.85	0.91	0.99	0.97	1.00	1.00	1.00	1.00	1.00
Proportional ML	0.87	0.93	0.99	0.98	1.00	1.00	1.00	1.00	1.00

Note: ML = maximum likelihood; BCH = Bolck-Croon-Hagenaars.

Table 4. Average Estimate of Selected β Parameter Separately for Each of the Nine Separation and Sample Size Conditions for All Three Types of Outcome Variables for Weak *X*-*Z* Association

Separation Level	Low			Medium			High		
Sample Size	500	1,000	10,000	500	1,000	10,000	500	1,000	10,000
Method					Nominal *Z*: $\beta_{23} = 1.50$				
One-step ML	1.53	1.61	1.51	1.57	1.53	1.51	1.51	1.51	1.50
Modal standard	0.61	0.66	0.72	1.10	1.10	1.10	1.39	1.40	1.39
Proportional standard	0.49	0.49	0.47	0.93	0.91	0.90	1.30	1.31	1.30
Modal BCH and ML	1.02	1.22	1.46	1.50	1.52	1.51	1.50	1.51	1.50
Proportional BCH and ML	1.06	1.29	1.44	1.50	1.50	1.51	1.50	1.51	1.50
					Ordinal *Z*: $\beta_2 = -1.24$				
One-step ML	−1.33	−1.29	−1.25	−1.25	−1.25	−1.23	−1.25	−1.27	−1.24
Modal standard	−0.72	−0.72	−0.72	−0.99	−1.01	−0.99	−1.20	−1.22	−1.19
Proportional standard	−0.59	−0.57	−0.55	−0.91	−0.91	−0.90	−1.17	−1.19	−1.16
Modal BCH	−1.14	−1.18	−1.23	−1.21	−1.24	−1.22	−1.25	−1.27	−1.24
Proportional BCH	−1.22	−1.23	−1.25	−1.23	−1.24	−1.23	−1.25	−1.27	−1.24
Modal ML	−1.14	−1.18	−1.24	−1.22	−1.25	−1.23	−1.25	−1.27	−1.24
Proportional ML	−1.21	−1.22	−1.24	−1.23	−1.24	−1.23	−1.22	−1.27	−1.24
					Continuous *Z*: $\beta_1 = 0.20$				
One-step ML	0.21	0.20	0.20	0.21	0.20	0.20	0.20	0.20	0.20
Modal standard	0.12	0.12	0.12	0.17	0.16	0.17	0.19	0.19	0.19
Proportional standard	0.10	0.09	0.09	0.16	0.15	0.15	0.18	0.18	0.19
Modal BCH	0.18	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Proportional BCH	0.19	0.20	0.20	0.20	0.19	0.20	0.20	0.20	0.20
Modal ML	0.18	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Proportional ML	0.19	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20

Note: ML = maximum likelihood; BCH = Bolck-Croon-Hagenaars.

Table 5. Average Estimated Standard Error (*SE*) and Standard Deviation (*SD*) across Replications for All Nine Conditions, Separately for One Parameter for Nominal Outcome Variable *Z* (β_{23} = 1.50) Obtained Using the One-step Maximum Likelihood (ML) and the Step 3 Correction Methods

Sample Size	500		1,000		10,000	
Method	*SD*	*SE*	*SD*	*SE*	*SD*	*SE*
			Low separation			
One-step ML	0.88	1.28	0.64	0.74	0.09	0.09
Modal BCH	0.66	0.65	0.53	0.51	0.16	0.15
Proportional BCH	0.61	0.58	0.48	0.46	0.15	0.14
Modal ML	0.66	0.64	0.53	0.51	0.16	0.15
Proportional ML	0.61	0.90	0.48	0.79	0.15	0.27
			Mid separation			
One-step ML	0.46	0.43	0.33	0.30	0.09	0.09
Modal BCH	0.47	0.45	0.34	0.31	0.10	0.10
Proportional BCH	0.44	0.42	0.32	0.29	0.09	0.09
Modal ML	0.47	0.45	0.34	0.31	0.10	0.10
Proportional ML	0.44	0.55	0.32	0.39	0.09	0.12
			High separation			
One-step ML	0.34	0.33	0.22	0.23	0.07	0.07
Modal BCH	0.34	0.33	0.22	0.23	0.07	0.07
Proportional BCH	0.34	0.33	0.22	0.23	0.07	0.07
Modal ML	0.34	0.33	0.22	0.23	0.07	0.07
Proportional ML	0.34	0.35	0.22	0.25	0.07	0.08

Note: BCH = Bolck-Croon-Hagenaars.

smaller sample size, the proportional ML and one-step ML methods tend to overestimate parameter uncertainty (*SE* is higher than *SD*). The other correction methods slightly underestimate the *SE*s in all nine conditions. With regard to efficiency, the correction methods perform similarly to the one-step ML method, with the exception of the combination of small sample size coupled with low separation, for which the correction methods are more efficient. Comparison of these results to those for other outcome variables and effect sizes showed that the *SE* bias of the correction methods is slightly larger in the strong effect condition for nominal and ordinal outcomes and smaller for continuous outcomes irrespective of the effect size.

6. TWO EMPIRICAL EXAMPLES

6.1. *Example 1: Types of Psychological Contracts*

To illustrate the working of the correction methods, we analyzed data from the Dutch and Belgian sample of the Psychological Contracts across Employment Situation (PSYCONES) project (European Commission 2006). We used the same questionnaire items as did De Cuyper et al. (2008), who performed an LCA to build a typology for psychological contracts between employers and employees. Out of the eight dichotomous indicators, four refer to employees' obligations (whether a promise was made or not) and four to employers' obligations, where each set of four items contained two items for relational and two for transactional obligations. Examples of the wording of items are, ''This organization promised me a reasonably secure job'' and ''This organization promised me a good pay for the work I do.'' The sample consisted of 1,365 respondents. The distal outcome variable *Z* was perceived job insecurity, measured using the scale from the PSYCONES project. This scale consists of four items with five categories and had a Cronbach's alpha value of .88. We used a composite score calculated from summing the four indicators to measure job insecurity.

In the first step, we fitted the measurement model using the eight indicator variables. Based on the Bayesian Information Criteria (BIC) values and the bivariate residuals between the items, it was concluded that a four-class model fitted the data well. Table 6 presents the parameter estimates for this four-class model. Class 1 (9 percent of respondents) is characterized by mutual low obligations. Class 2 (10 percent) represents employee underobligation: These respondents are likely to

Table 6. Class Proportions and Class-specific Probabilities of a Positive Response for the Four-class Model Estimated for the Psychological Contracts across Employment Situation Data

	Class 1 Mutual Low	Class 2 Underobligation	Class 3 Overobligation	Class 4 Mutual High
Class proportion	.09	.10	.29	.52
Employers' obligations				
Reasonably secure job	.21	.87	.36	.90
Opportunities to advance	.17	.84	.30	.90
Good pay for the work	.26	.75	.28	.87
Safe working environment	.28	.73	.55	.97
Employees' obligations				
Show loyalty to the organization	.08	.36	.72	.96
Volunteer to do tasks outside your job description	.17	.37	.82	.96
Turn up for work on time	.18	.38	.96	.98
Meet the performance expectations for your job	.28	.77	.97	.99

perceive employers' obligations as given and have a lower probability of perceiving their own obligations as promised. Class 3 (29 percent) represents employees who themselves made promises to the organization but received less: the overobligation class. Class 4 (52 percent) scores high on all items, representing mutual high obligations.

After identifying the classes, the posterior class membership probabilities were saved, and the **D** matrix with elements $P(W=s|X=t)$ and its inverse were calculated. The one-step and the corrected and uncorrected three-step methods were used to analyze the relationship between class membership and perceived job insecurity, where the latter is treated as a continuous variable with a constant error variance; that is, in the three-step approaches, we used a linear regression to regress job insecurity on class membership, and in the one-step method, we used job insecurity as a continuous indicator variable. We also ran the analysis treating job insecurity as ordinal variable, and relating the psychological contract type and job insecurity using a cumulative logit link, and obtained comparable results. To keep a more parsimonious model, we opted for treating job insecurity as continuous. This is the Latent GOLD syntax used for three-step ML with modal assignment:

```
variables
    latent X nominal 4;
    dependent Jobinsecurity continuous, W nominal;
equations
    X <- 1;
    JobInsecurity <- 1 + X;
    W <- (D~wei) 1 | X;
    D={.867 .043 .089 .000
        .066 .691 .094 .148
        .014 .026 .860 .101
        .000 .006 .055 .938}
```

As can be seen, an LCA is performed with two indicators (dependent variables), where the specified **D** matrix fixes the *X*-*W* association. For the BCH method, an expanded data set is created with weights based on the inverse of the **D** matrix. Using this weighted data set, *Z* is regressed on *X*. For example, in the case of modal assignment, an individual assigned to class 1 ($W = 1$) receives the weights: 1.154, −0.100, 0.016, and −0.001 for classes 1, 2, 3, and 4, respectively. Under modal assignment, the proportion of classification errors is .11, and the corresponding entropy-based R^2 equals .72. This indicates that the classes are well separated and that as a result the corrected three-step methods can be expected to perform well.

The estimated effect sizes of psychological contract type on job insecurity (and their *SE*s) and the value of the Wald test for the overall effect (and its *p* value) are reported in Table 7. As we can see in the table, the job insecurity of the employee underobligation (class 2) and mutual-high-obligations group (class 3) is lower than for those in the mutual-low-obligations group (class 1). The job insecurity of the employee overobligation group is similar to that of the mutual-low-obligations group. Comparing the effect parameters obtained by the different methods, we can see that the standard three-step procedures yield estimates

Table 7. Effect of Class Membership on Job Insecurity, Standard Errors (*SE*s), Multivariate Wald Test for the Effect, and Its Significance Obtained with the Seven Methods, Using Dummy Coding with First Class as Reference Category

Method	Class 2 (*SE*)	Class 3 (*SE*)	Class 4 (*SE*)	Wald (*df*)	*p*
One-step ML	−0.60 (0.16)	0.10 (0.13)	−0.46 (0.11)	68.04 (3)	< .001
Modal standard	−0.41 (0.13)	0.04 (0.10)	−0.36 (0.10)	56.43 (3)	< .001
Proportional standard	−0.34 (0.13)	0.01 (0.10)	−0.35 (0.10)	42.82 (3)	< .001
Modal BCH	−0.53 (0.12)	0.08 (0.10)	−0.42 (0.09)	89.94 (3)	< .001
Proportional BCH	−0.50 (0.12)	0.08 (0.10)	−0.43 (0.09)	87.76 (3)	< .001
Modal ML	−0.53 (0.16)	0.09 (0.13)	−0.43 (0.11)	57.65 (3)	< .001
Proportional ML	−0.54 (0.19)	0.10 (0.14)	−0.44 (0.12)	44.35 (3)	< .001

Note: ML = maximum likelihood; BCH = Bolck-Croon-Hagenaars.

that are far away from ones of the other methods, while all the other methods yield similar estimates. The correction methods have slightly smaller parameter values than the one-step method, where the three-step ML methods are closer to the one-step method than the BCH methods. Similar to the results of the simulation study, the *SE*s obtained using the proportional ML method are slightly larger than the ones obtained using the other correction methods and the one-step ML. The *SE*s obtained by the other correction methods are similar to the ones obtained using the one-step ML method. Looking at the Wald tests of the correction methods, it can be seen that the Wald value for the proportional ML method is the lowest, meaning that this method is the most conservative.

6.2. *Example 2: The Effect of Religiosity and Social Status on Political Ideology: A Multiple Latent Variable Model*

In many research situations, it is of interest to predict a latent outcome variable from other latent variables. We will illustrate how the ML three-step method can be used for this purpose with data from the Dutch sample of the 1981 European Value Survey (GESIS-Variable Reports No. 2011). More specifically, we investigate how religiosity affects political ideology while controlling for social status. Social status is an observed variable with four ordinal categories: professional/managerial, semiskilled, unskilled, or unemployed or pensioner (1); skilled manual

Table 8. Class Proportions and Class-specific Probabilities of Religiosity for the Three-class Model Estimated for the 1981 Wave of the European Value Survey Data

		Class 1 Nonreligious	Class 2 Middle	Class 3 Religious
Class proportion		0.34	0.33	0.33
Religiosity	no	0.95	0.06	0.01
	yes	0.05	0.94	0.99
Personal God	no	0.99	0.80	0.11
	yes	0.01	0.20	0.89
Traditionalism	nontraditional	0.84	0.15	0.01
	intermediate	0.15	0.65	0.06
	traditional	0.01	0.20	0.93
Religious organization membership	no	0.96	0.66	0.25
	yes	0.04	0.34	0.75
Denomination	yes	0.08	0.80	0.99
	no	0.92	0.20	0.01
Prayer	yes	0.32	0.66	0.96
	no	0.68	0.33	0.04

workers (2); sales, clerical, and other nonmanual (3); and above average lifestyle (4). We used social status as a numeric covariate in the analysis. Similarly to Hagenaars and Halman's (1989) work on the same data set, we modeled political ideology and religiosity as categorical latent variables measured using multiple indicators.

Religiosity was measured with six indicators, among which are praying, belonging to a church, and belonging to a denomination (see Table 8). We selected the three-class model based on the BIC and the goodness of fit (L^2 = 83.68, *df* = 74, *p* = .21). Class separation is good (entropy R^2 = .85). Table 8 shows the class solution. Group 1 (34 percent) is the ''nonreligious,'' scoring low on all items, while group 2, the ''middle'' (33 percent), has mixed scores and the last group, the ''religious,'' (33 percent) score high on all items.

Political ideology was measured with six indicators, among which are party closeness and Left-Right orientation (see Table 9). We fitted latent class models with different numbers of classes and selected the three-class model based on the lowest BIC and a nonsignificant goodness-of-fit statistic (L^2 = 79.38, *df* = 74, *p* = .31). Class separation is moderate (entropy R^2 = .68). Group 1 can be characterized as ''left wing'' (27

Table 9. Class Proportions and Class-specific Probabilities of Political Mentality for the Three-class Model Estimated for the 1981 Wave of the European Value Survey Data

Class		Class 1 Left	Class 2 Middle/Indifferent	Class 3 Right
Class proportion		0.27	0.37	0.35
Left/Right	Left	0.89	0.25	0.02
	middle	0.10	0.53	0.27
	Right	0.01	0.22	0.71
Political interest	no	0.28	0.77	0.39
	yes	0.72	0.23	0.61
Trust in Parliament	no	0.60	0.68	0.26
	yes	0.40	0.32	0.74
Societal change	no	0.17	0.20	0.41
	yes	0.83	0.80	0.59
Equality versus freedom	equality	0.51	0.60	0.76
	freedom	0.49	0.40	0.24
Party closeness	yes	0.92	0.10	0.83
	no	0.08	0.90	0.17

percent), group 2 as ‘‘middle/indifferent’’ (37 percent), and the last group as ‘‘right wing’’ (35 percent).

Note that the two measurement models are estimated separately. Each yields a set of modal class assignments and an estimate of the **D** matrix with the conditional probability of being assigned to a class conditional on the true class membership. The class assignments and the two **D** matrices can be used to set up the model in which (latent) political ideology is predicted from social status and (latent) religiosity. The syntax for the specified model using the three-step ML method is provided in appendix B. Table 10 reports the estimates for the effects of religiosity and social class on political ideology obtained with the one-step, the adjusted three-step, and the standard unadjusted three-step approach. As can be seen, the results obtained with all three methods point toward the same tendencies. While the estimates from the one-step and the corrected three-step method are rather similar, the uncorrected three-step approach yields smaller effect sizes.

Based on the estimated multinomial logit coefficients, one can conclude that controlling for social class, the more religious a person, the more likely it is that he or she is politically Right or middle/indifferent rather than Left. Moreover, the higher the social class, the more likely

Table 10. Multinomial Logit Coefficients from the Regression of Religiosity on Social Class on Political Ideology, Standard Errors (*SE*s), Multivariate Wald Tests, Obtained with Three Methods Using Dummy Coding with the First Class as Reference Category for Religiosity and Political Ideology

	Political = Middle (*SE*)	Political = Right (*SE*)	Wald (*df*)
One step ML			
Religiosity = middle	0.67 (0.35)	0.80 (0.41)	52.80 (4)
Religiosity = religious	1.23 (0.57)	3.09 (0.50)	
Social class	−0.27 (0.47)	1.39 (0.39)	18.33 (2)
Modal ML			
Religiosity = middle	0.48 (0.34)	0.72 (0.45)	39.85 (4)
Religiosity = religious	1.10 (0.46)	2.76 (0.48)	
Social class	−0.15 (0.43)	1.33 (0.41)	14.83 (2)
Uncorrected modal			
Religiosity = middle	0.43 (0.27)	0.65 (0.31)	47.94 (4)
Religiosity = religious	0.99 (0.31)	2.10 (0.33)	
Social class	0.04 (0.33)	1.03 (0.32)	16.74 (2)

Note: ML = maximum likelihood.

he or she is to be right wing rather than left wing, while there is no significant effect of social class on the middle-Left contrast.

7. DISCUSSION

We proposed a generalization of existing correction methods for the attenuation problem appearing in three-step LCA with external variables. We showed how two existing correction methods for latent class models with covariates can be generalized to a broader range of situations, that is, to formulate models for the joint probability of class membership and external variables. The correction methods can therefore now be applied in any situation in which we wish to relate scores on class membership with external variables, irrespective of the hypothesized causal order. Although we focused mainly on the situation in which class membership is a predictor of a continuous, ordinal, or nominal outcome variable, the correction methods can be applied in relation with distal outcome variables having almost any of the distributional forms from the exponential family. We also showed how the ML correction method can be extended to models with more than one latent variable.

The performance of the correction methods was tested by a simulation study and illustrated with two real data examples. The results of the simulation study show, similarly to previously reported results, that using the uncorrected three-step approach leads to seriously biased parameter estimates of the association of class membership with external variables. Although the direction of the effects is correct, the effect sizes are very much attenuated. As such, it is recommended that one use one of the correction methods when deciding to use the three-step approach. All correction methods we tested perform well; both their estimates and *SE*s can be trusted, with the exception of the situations wherein the class separation of the measurement model is very low, in which situation they underestimate the parameter estimates and *SE*s. The most efficient correction method is the proportional ML method. In general, the results obtained with proportional assignment are better than those obtained using modal assignment. A nonanticipated result is that for nominal outcomes the BCH method may fail because of the occurrence of negative cell frequencies, a problem that is much more likely to occur with a (very) low separation between classes. Therefore, the use of the one-step or three-step ML methods is recommended in these situations.

One of the limitations of the current study is that it examined the behavior of the correction methods only for the situation in which model assumptions hold; that is, we did not look at situations in which distributional assumptions about the external variables and/or conditional independence assumptions are violated. As such it is recommended that future research look into adapting the three-step methods to be able to accommodate situations wherein the conditional independence assumption does not hold. Another option to be investigated is testing the performance of the different methods under model misspecification. Our expectation is that the adjusted three-step methods may perform better than the one-step method under misspecification, which is one of the issues we will focus on in future research.

Another limitation is that we focused mainly on parameter bias and less on hypothesis testing. This means that we cannot say anything yet about issues such as amount of power decrease of statistical tests such as the Wald test resulting from using the proposed correction methods. This is another topic for future research.

APPENDIX A

R Code for the Computation of the **D** *Matrix and Restructuring of the Data File*

```
#read in the file containing the posterior class
assignment probabilities to R
library(foreign, pos=4)
classa <- read.spss(‘‘class.sav’’,use.value.
labels=TRUE, max.value.labels=Inf, to.data.
frame=TRUE)

#creating the weights needed for the D matrix
calculation

n<-length(classa$id) # the length of the data file

ctm<-matrix(nrow=n,ncol= 9) #modal weights 3*3
class
ctp<-matrix(nrow=n,ncol= 9)#proportional
weights 3*3 class
modal<-matrix(nrow=n,ncol= 3) # modal class
assignment 3 dummies for 3 classes

for (j in 1:3) # creating dummies for modal
posterior class assignment
{
modal[,j]<- ifelse (classa[,12]==j, 1, 0)
}
# obtaining the elements of the D matrix
i<-1
 for (s in 1:3) # looping over all combinations of
posterior classification probabilities
{
  for (k in 1:3)
{
   ctm[,i]<- classa[,k+8]*modal[,s] # creating the modal correction
   weights
  if (k==1) fixed<-classa[,s+8] #where classa
  [9:11]are the proportional posterior
  classification probabilities
  ctp[,i]<- classa[,k+8]*fixed # creating the
  proportional correction weights
i<-i+1
 }
}
```

```
   combined<-matrix(c(modal,ctm,ctp),ncol=21,
   nrow=n) #collecting all weights to a matrix

   COLSUMS<- matrix(c(apply(combined,2,sum)),
   1,21) #summing all the weights
   # creating the modal D matrix
   DIM <- solve(matrix(c((matrix(c(COLSUMS
   [,4:12]),3,3,byrow=F))/apply(matrix(c(COLSUMS[,4:12]),
   3,3,byrow=F),1,sum)),3,3))

   #creating the proportional D matrix
   DIP<-solve(matrix(c((matrix(c(COLSUMS[,13:21]),
3,3, byrow=F))/apply(matrix(c(COLSUMS[,13:21]),
3,3, byrow=F),1,sum)),3,3))

   #final modal bch weights applied to each case i
   wm1<- ((combined[,1]*DIM[1,1]) + ( combined[,2]
   *DIM[2,1]) + (combined[,3]*DIM[3,1]))
   wm2<- ((combined[,1]*DIM[1,2])+ (combined[,2]
   *DIM[2,2])+ (combined[,3]*DIM[3,2]))
   wm3<- ((combined[,1]*DIM[1,3]) + ( combined[,2]
   *DIM[2,3])+ (combined[,3]*DIM[ [3,3]))

   # final proportional bch weights applied to each
   case i
   wp1<- ((classa [,9]*DIP[1,1])+( classa [,10]
   *DIP[2,1])+(classa [,11]*DIP[3,1]))
   wp2<- ((classa [,9]*DIP[1,2])+( classa [,10]
   *DIP[2,2])+( classa [,11]*DIP[3,2]))
   wp3<- ((classa [,9]*DIP[1,3])+( classa [,10]
   *DIP[2,3])+( classa [,11]*DIP[3,3]))
   #create and save long file

   class_longa<-data.frame (classa[,1:8],
wmodal1=combined[,1], wmodal2=combined[,2],
wmodal3=combined[,3], wprop1=classa[,9],
wprop2= classa[,10],wprop3=classa[,11],
wbchmodal1=wm1, wbchmodal2=wm2,
wbchmodal3=wm3,wbchprop1=wp1,
wbchprop2=wp2, wbchprop3=wp3)

   library(Hmisc)
      class_long<- reShape(class_longa, base=
      c(‘wmodal’,’wprop’,’wbchmodal’,’wbchprop’),
      reps=3)
```

```
    write.table(class_long, ''class_long.txt'',
    sep='' '', col.names=TRUE, row.names=FALSE,
    quote=TRUE, na=''NA'')
```

APPENDIX B

Latent GOLD Syntax Used for Example 2

```
options
  output parameters=first standarderrors
  estimatedvalues;
variables
  dependent PoliticalModal, ReligiosityModal;
  independent SocialClass;
  latent Political nominal 3, Religiosity nominal 3;
equations
  Religiosity <- 1 + SocialClass;
  Political <- 1 + Religiosity + SocialClass;
  PoliticalModal <- (D~wei) 1 | Political;
  ReligiosityModal <- (F~wei) 1 | Religiosity;
  D = {0.854843 0.078100 0.067056
         0.036183 0.890474 0.073343
         0.022912 0.113239 0.863849};
  F = {0.970735 0.029264 0.000000
         0.037784 0.883258 0.078959
         0.000000 0.050674 0.949326};
```

Acknowledgment

We thank the members of the Psychological Contracts across Employment Situation project for allowing us to use data from this project for our work.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by The Netherlands Organization for Scientific Research (grant number 453-10-002).

Notes

1. Note that in equation 5, we implicitly use the equality $P(W|\boldsymbol{Y},X)=P(W|\boldsymbol{Y})$. This follows from the fact that class assignment depends only on Y (and the latent class analysis model parameters) but not directly on X.
2. While in equation 8 it is clear that the extension to more covariates $\mathbf{Z}$ is straightforward, this is also possible using equation 9, assuming conditional independence of outcomes given X.

3. Although in this article we emphasize the need of the conditional independence assumption to hold to be able to use any of the three-step methods, it should be mentioned that an extension of the corrected three-step approaches could be developed that makes it possible to include direct effects of categorical covariates on indicators in the model. This could be done by applying the weighting that we present in the following pages separately at every level of the external covariate.
4. Using matrix algebra, we can write equation 10 as $\mathbf{E} = \mathbf{A}\ \mathbf{D}$, where $\mathbf{E}$ contains the $P(W = \mathrm{s}, Z = z)$, $\mathbf{A}$ the $P(X = \mathrm{t}, \mathrm{Z} = z)$, and $\mathbf{D}$ the $P(W = s|X = t)$. Standard matrix operation yields $\mathbf{A} = \mathbf{E}\ \mathbf{D}^{-1}$, which is what is expressed in equation 13.
5. The separation is measured by the entropy R-squared, which tells how much the prediction of X improved when using the information on $\mathbf{Y}$**.** If $P(X = t|\boldsymbol{Y} = \boldsymbol{y})$ is close to 0 or 1 for most data patterns, the separation between classes is good, and the classification error is low.

References

Agresti, Alan. 2002. *Categorical Data Analysis.* 2nd ed. New York: Wiley.

Bandeen-Roche, Karen, Miglioretti, Diana L., Zeger, Scott L., and Rathouz, Paul J. 1997. ‘‘Latent Variable Regression for Multiple Discrete Outcomes.’’ *Journal of the American Statistical Association* 92:1375–86.

Bauer, Daniel J. and Patrick J. Curran. 2003. ‘‘Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes.’’ *Psychological Methods* 8:338–63.

Bolck, Annabel, Marcel Croon, and Jacques Hagenaars. 2004. ‘‘Estimating Latent Structure Models with Categorical Variables: One-step versus Three-step Estimators.’’ *Political Analysis* 12:3–27. doi:10.1093/pan/mph001.

Chan, Tak Wing and John H. Goldthorpe. 2007. ‘‘European Social Stratification and Cultural Consumption: Music in England.’’ *Sociological Review,* 23:11–19. doi:10.1093/esr/jcl016.

Clark, Shaunna L. and Bengt Muthén. 2009. *Relating Latent Class Analysis Results to Variables Not Included in the Analysis.* Retrieved June 16, 2012 (http://statmodel2.com/download/relatinglca.pdf*).*

Clogg, Clifford C. 1981. ‘‘New Developments in Latent Structure Analysis.’’ Pp. 215–46 in *Factor Analysis and Measurement in Sociological Research,* edited by D. J. Jackson and E. F. Borgotta. Beverly Hills, CA: Sage.

Dayton, C. Mitchell and George B. Macready. 1988. ‘‘Concomitant-variable Latent-class Models.’’ *Journal of the American Statistical Association* 83:173–178. doi:10.2307/2288938.

De Cuyper, Nele, Thomas Rigotti, Hans De Witte, and Gisela Mohr. 2008. ‘‘Balancing Psychological Contracts: Validation of a Typology’’. *International Journal of Human Resource Management* 19:543–61. doi:10.1080/09585190801953590.

Dias, José G. and Jeroen K. Vermunt. 2008. ‘‘A Bootstrap-based Aggregate Classifier for Model-based Clustering.’’ *Computational Statistics, 23,* 643–59. doi:10.1007/s00180-007-0103-7.

European Commission. 2006. *EU Research on Social Sciences and Humanities. Psychological Contracts across Employment Situations: PSYCONES.* Retrieved June 16, 2012 (http://cordis.europa.eu/documents/documentlibrary/100123961EN6.pdf).

GESIS-Variable Reports No. 2011. *EVS 1981—Variable Report Integrated Dataset.* Retrieved June 16, 2012 (http://info1.gesis.org/dbksearch/file.asp?file=ZA4438_r.pdf).

Goetghebeur, Els, Jan Liinev, Marleen Boelaert, and Patrick Van der Stuyft. 2000. ''Diagnostic Test Analyses in Search of Their Gold Standard: Latent Class Analyses with Random Effects.'' *Statistical Methods in Medical Research, 9,* 231. doi:10.1177/096228020000900304.

Goodman, Leo 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* 79:79–259. doi:10.1086/225676.

Goodman, Leo A. 2007. ''On the Assignment of Individuals to Latent Classes.'' Pp. 1–22 in *Sociological Methodology,* Vol. 37, edited by J. Xie. Boston: Blackwell.

Hagenaars, Jacques A. 1988. ''Latent Structure Models with Direct Effects between Indicators: Local Dependence Models.'' *Sociological Methods and Research* 16:379–405.

Hagenaars, Jacques A. 1990. *Categorical Longitudinal Data: Loglinear Analysis of Panel, Trend and Cohort Data.* Newbury Park, CA: Sage.

Hagenaars, Jacques A. and Loek C. Halman. 1989. ''Searching for Idealtypes: The Potentialities of Latent Class Analysis.'' *European Sociological Review* 5(1):81–96.

Huang, David, Mary-Lynn Brecht, Motoaki Hara, and Yih-Ing Hser. 2010. ''Influences of a Covariate on Growth Mixture Modeling.'' *Journal of Drug Issues* 40(1):173–94.

Huang, Guan-Hua and Karen Bandeen-Roche. 2004. ''Building an Identifiable Latent Class Model with Covariate Effects on Underlying and Measured Variables.'' *Psychometrika* 69(1):5–32.

Kam, Jennifer A. 2011. ''Identifying Changes in Youth's Subgroup Membership over Time Based on Their Targeted Communication about Substance Use with Parents and Friends.'' *Human Communication Research* 37(3):324–49.

Lazarsfeld, Paul F. and Neil W. Henry. 1968. *Latent Structure Analysis.* Boston: Houghton Mifflin.

McCutcheon, Allan L. 1987. *Latent Class Analysis.* Newbury Park, CA: Sage.

Morin, Alexandre J. S., Julien Morizot, Jean-Sébastien Boudrias, and Isabelle Madore. 2011. ''A Multifoci Person-centered Perspective on Workplace Affective Commitment: A Latent Profile/Factor Mixture Analysis.'' *Organizational Research Methods* 14:58.

Mulder, Eva, Jeroen Vermunt, Eddy Brand, Ruud Bullens, and Hjalmar van Marle. 2012. ''Recidivism in Subgroups of Serious Juvenile Offenders: Different Profiles, Different Risks?'' *Criminal Behaviour and Mental Health* 22:122–35.

Muthén, Bengt. 2004. ''Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data.'' Pp. 345–68 in *Handbook of Quantitative Methodology for the Social Sciences,* edited by D. Kaplan. Newbury Park, CA: Sage.

Olino, Thomas M., Daniel N. Klein, Peter M. Lewinsohn, Paul Rohde, and John R. Seeley. 2010. ''Latent Trajectory Classes of Depressive and Anxiety Disorders from

Adolescence to Adulthood: Descriptions of Classes and Associations with Risk Factors.'' *Comprehensive Psychiatry* 51:224–35.

Petersen, Janne, Karen Bandeen-Roche, Esben Budtz-Jørgensen, and Klaus G. Larsen. 2012. ''Predicting Latent Class Scores for Subsequent Analysis.'' *Psychometrica* 77(2):244–62. doi:10.1007/s11336-012-9248-6.

Petras, Hanno and Katherine Masyn. 2010. ''General Growth Mixture Analysis with Antecedents and Consequences of Change.'' Pp. 69–100 in *Handbook of Quantitative Criminology,* edited by A. R. Piquero and D. Weisbrud. New York: Springer.

Skinner, C. J., D. Holt, and T. M. F. Smith. 1989. *Analysis of Complex Surveys.* New York: Wiley.

Tofighi, Davood and Craig K. Enders. 2008. ''Identifying the Correct Number of Classes in Growth Mixture Models.'' Pp. 317–41 in *Advances in Latent Variable Mixture Models,* edited by G. R. Hancock and K. M. Samuelsen. Charlotte, NC: Information Age.

Vermunt, Jeroen K. 2010. ''Latent Class Modeling with Covariates: Two Improved Three-step Approaches.'' *Political Analysis* 18:450–69. doi:10.1093/pan/mpq025.

Vermunt, Jeroen K. and Jay Magidson. 2005. *Latent GOLD 4.0 User's Guide.* Belmont, MA: Statistical Innovations.

Vermunt, Jeroen K. and Jay Magidson. 2008. *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module.* Belmont, MA: Statistical Innovations.

Yamaguchi, Kazuo. 2000. ''Multinomial Logit Latent-class Regression Models: An Analysis of the Predictors of Gender-role Attitudes among Japanese Women.'' *American Journal of Sociology* 105:1702–740.

Author Biographies

Zsuzsa Bakk is a PhD student in the Department of Methodology and Statistics at Tilburg University, the Netherlands. She has a BA degree in Sociology, and a MSc in Human Resource Studies from Babes Bolyai University, Romania, and a MSc in Social and Behavioral Sciences from Tilburg University. Her research focuses on step wise latent class analysis.

Fetene B. Tekle is a researcher in the Department of Methodology and Statistics at Tilburg University, the Netherlands. He holds PhD in Statistics from Maastricht University, the Netherlands. His research focuses on optimal designs and lately combined with latent class analysis. He has published a couple of methodological and applied papers in international journals with his colleagues.

Jeroen K. Vermunt is a full professor in the Department of Methodology and Statistics at Tilburg University, the Netherlands. He holds a Ph.D. in Social Sciences from Tilburg University. He has published extensively on categorical data techniques, methods for the analysis of longitudinal and event history data, latent class and finite mixture models, and latent trait models. He is the co-developer (with Jay Magidson) of the Latent GOLD software package. His full CV and publications can be found at http://members.home.nl/jeroenvermunt/