# Micro-macro multilevel analysis for discrete data: A latent variable approach and an application on personal network data*

Margot Bennink

Marcel A. Croon

Jeroen K. Vermunt

Tilburg University, the Netherlands

* Corresponding author: Margot Bennink, Room P1.113, PO Box 90153, 5000 LE Tilburg, the Netherlands, m.bennink@tilburguniversity.edu. Acknowledgments: We would like to thank the NKPS team for providing the data for the empirical example.

# MICRO-MACRO MULTILEVEL ANALYSIS FOR DISCRETE DATA: A LATENT VARIABLE APPROACH AND AN APPLICATION ON PERSONAL NETWORK DATA

A multilevel regression model is proposed in which discrete individual-level variables are used as predictors of discrete group-level outcomes. It generalizes the model proposed by Croon and van Veldhoven (2007) for analyzing micro-macro relations with continuous variables by making use of a specific type of latent class model. A first simulation study shows that this approach performs better than more traditional aggregation and disaggreagtion procedures. A second simulation study shows that the proposed latent variable approach still works well in a more complex model, but that a larger number of level-2 units is needed to retain sufficient power. The more complex model is illustrated with an empirical example in which data from a personal network are used to analyze the interaction effect of being religious and surrounding yourself with (un)married people on the probability of being married.

**keywords:** generalized linear modeling, multilevel analysis, level-2 outcome, latent class analysis, latent variable, micro-macro analysis, personal network, marriage, religion

# MICRO-MACRO MULTILEVEL ANALYSIS FOR DISCRETE DATA: A LATENT VARIABLE APPROACH AND AN APPLICATION ON PERSONAL NETWORK DATA

In many research situations in the social and behavioral sciences, data are collected within hierarchically ordered systems. For example, data may be collected on individuals nested within groups. Repeated measures carried out on the same individuals can also be treated as nested observations within these individuals. Data collected in a personal or egocentric network are hierarchical as well since data are collected on individuals (egos), and on persons from the network of these individuals (alters) or on ties (ego-alter relations). This data collection procedure is an example of a multilevel design in which the observations on the alters or ties are nested within the egos (Hox and Roberts 2011; Snijders, Spreen, and Zwaagstra 1995). In the current article, data are considered hierarchical when both the level-2 units and the level-1 units are a (random) sample of the population of possible level-2 and level-1 units.

In these two-level settings, two basically different situations can be distinguished. In a first situation, independent variables defined at the higher-level are assumed to affect dependent variables defined at the lower-level.

For example, whether firms have a salary bonus system or not may affect the individual productivity of the employees working in these firms (Snijders and Bosker 1999). Snijders and Bosker (1999) refer to these relationships as macro-micro relations, but they are also referred to as 2-1 relations since a level-2 explanatory variable affects a level-1 outcome variable. In the last few decades many efforts have been made to develop multilevel models for this kind of hierarchical ordering of variables, and although the bulk of this work has emphasized multilevel linear regression models for continuous variables, multilevel regression models for discrete response variables have also been proposed (Goldstein 2003; Snijders and Bosker 1999). Standard multilevel software as implemented in, for instance, SPSS, MLwiN (Rasbash et al. 2005), and Mplus (Muthén and Muthén 2010) is available to estimate these multilevel models.

In a second situation, referred to as a micro-macro situation by Snijders and Bosker (1999), independent variables defined at the lower-level are assumed to affect dependent variables defined at the higher-level. These relations, which can also be referred to as 1-2 relations, have received less attention in the statistical literature than the models for analyzing 2-1 relations. This is rather odd since this type of relation occurs rather frequently in the social and behavioral sciences. As a first example to illustrate the need for appropriate micro-macro methods, consider organizational research that tries to link team performance or team effectiveness to some attributes or characteristics of the individual team members (DeShon et al. 2004; van Veldhoven 2005; Waller et al. 2001). Also in educational psychology these micro-macro relations may be of interest, e.g., when the global school effectiveness is studied in relation to the attributes of the individual students and teachers (Rutter and Maughan 2002).

Two traditional approaches for analyzing micro-macro relationships are commonly in use: either, the individual-level predictors are aggregated to the group level, or the group-level outcome variables are disaggregated to the individual level, and the analysis is concluded with a single-level regression analysis at the appropriate level. More recently, Croon and van Veldhoven (2007) presented an alternative latent variable approach for analyzing micro-macro relations with continuous outcomes. This approach has only been fully worked out yet for the case of linear relationships among continuous explanatory and outcome variables. The present article discusses how to extend this latent variable approach to the analysis of discrete data.

In the remaining of this article, the aggregation, disaggregation and latent variable approaches to deal with a micro-macro hypothesis are described, applied to discrete data and evaluated and compared in a simulation study. Subsequently, a discrete group-level predictor is added to the micro-macro model and this extended model is evaluated in a second simulation study and illustrated with an empirical example on personal network data.

## ANALYZING MICRO-MACRO RELATIONS

### *Aggregation and Disaggregation*

For the analysis of micro-macro relations, two traditional approaches are currently being applied: either the individual-level predictors are aggregated to the group level or the group-level outcome variables are disaggregated to the individual level, and the final analysis is concluded with a single-level regression analysis at the appropriate level.

The first approach to deal with micro-macro relations is to aggregate the individual-level predictors to the group level by assigning a mode, median or mean score to every group based on the scores of the individuals within the group. It is then assumed that the assigned scores perfectly reflect the construct at the group level. This assumption is not realistic in practice since the group-level construct does not represent the heterogeneity within groups, and, moreover, may be affected by measurement error and sampling fluctuation (Lüdtke et al. 2011). Also the number of observations on which the final regression analysis is carried out decreases since the groups are treated as the units of analysis. Consequently, the power of the statistical tests involved may sharply decrease (Krull and MacKinnon 1999). Moreover, aggregation has the additional disadvantage that the information about the individual-level variation within the groups is completely lost.

When disaggregating the outcome variable, on the other hand, each individual in a group is assigned his group-level score, which in the further analysis is treated as if it was an independently observed individual score. Since the scores of all individuals within a particular group are the same, the assumption of independent errors among individuals (Keith 2005), as made in regression analysis, is clearly not valid. This violation leads to inefficient estimates, biased standard errors, and overly liberal inferences for the model parameters (Krull and MacKinnon 1999; MacKinnon 2008). Moreover, by analyzing the data at the individual level in this way, the total sample size is not corrected for the dependency among the individual observations within a group, which causes the power of the analysis to be artificially high.

### *Latent Variable Approach*

Recently, Croon and van Veldhoven (2007) presented an alternative approach for analyzing micro-macro relations with continuous outcomes which overcomes many of the problems associated with aggregation or disaggregation. The general idea of the latent variable approach is illustrated by the model shown graphically in Figure 1. This model covers the situation with a single explanatory variable at the individual level ($Z_{ij}$) affecting a single outcome variable at the group level ($Y_j$). In the notation used here, the subscript $j$ refers to the groups, while the subscript $i$ refers to individuals within a group.

Figure 1 about here

To analyze the relationship between the individual-level independent variable and the group-level outcome, the scores on $Z_{ij}$ are treated as exchangeable indicators for a latent group-level variable $\zeta_j$. The exchangeability assumption implies that the relation between the individual-level observation and the group-level latent variable is assumed to be the same for all individuals within a group. In this way, all individuals are treated as equivalent sources of information about the group-level variable, and none of them is considered as providing more accurate judgments in this respect than his comembers. This assumption is warranted when all group members play similar or identical roles in the group and is probably less vindicated when the group members differ with respect to their functioning in the group. $\zeta_j$ is treated as a predictor or explanatory variable for the group-level outcome variable $Y_j$. In this way, the individual-level observations on $Z_{ij}$ are not assumed to reflect the group-level construct $\zeta_j$ perfectly, but within group heterogeneity, and sampling variability are allowed to exist. This model actually consists of two parts: a measurement part which relates the individual-level scores on $Z_{ij}$ to the latent variable $\zeta_j$ at the group level, and a structural part in which $Y_j$ is regressed on $\zeta_j$.

The latent variable approach can be generalized to situations in which the variables from the measurement or the structural part of the model are not necessarily continuous. With respect to the measurement model, the four different measurement models which are obtained by independently varying the scale type of the observed variable $Z_{ij}$ and the latent variable $\zeta_j$, are shown in Table 1.

Table 1 about here

The basic idea is that groups can be classified or located on either a continuous or discrete latent scale at the group level, and that the group members are acting as 'imperfect' informants or indicators of their group's position on this latent group-level scale. Furthermore, the information the group members provide about the group's position can also be considered as being measured on either a continuous or a discrete scale.

When both the observed variable $Z_{ij}$ and the latent variable $\zeta_j$ are assumed to be continuous, as in Croon and van Veldhoven (2007), a linear factor model links the individual-level scores to the group-level score. Alternatively, one might assume that a discrete latent variable at the group level underlies a continuous observed variable at the individual level. In this situation the measurement part of the model is described by a latent profile model (Bartholomew and Knott 1999). In situations in which the observed explanatory variables at the individual level are discrete, either a latent class model (Hagenaars and McCutcheon 2002) or an item response model (Embretson and Reise 2000) might be considered. A latent class model is appropriate when the underlying latent variable at the group level is discrete as well, whereas an item response model is appropriate when the underlying latent variable is assumed to be continuous.

With respect to the structural part of the model, the regression of $Y_j$ on $\zeta_j$ at the group level can be conceived in different ways depending on the measurement level of the outcome variable $Y_j$. For a continuous outcome variable, Croon and van Veldhoven (2007) defined a linear regression model, but when the group-level outcome variable $Y_j$ is discrete, (multinomial) logit or probit regression models are more appropriate to regress $Y_j$ on $\zeta_j$, irrespective of the scale type of $\zeta_j$. All these models fit within the general framework of generalized latent variable models described by Skrondal and Rabe-Hesketh (2004).

## DISCRETE VARIABLES

The focus of the current paper will be on the application of the latent variable approach to discrete data by combining a latent class model for the measurement model with a (multinomial) logistic regression model at the group level. Readers interested in specifying a continuous latent variable underlying discrete observations are referred to Fox and Glas (2003) and Fox (2005). Our discussion of the model for discrete variables first considers the case in which all variables are dichotomous before sketching the more general case.

Consider again the model shown in Figure 1 but now assume that all variables in the model are dichotomous with values 0 and 1. In this 1-2 model, the relationship between a single dichotomous explanatory variable $Z_{ij}$ at the individual level and a single dichotomous outcome variable $Y_j$ at the group level is at issue. The type of models that are discussed in this article, and of which the model shown in Figure 1 is a first, very basic example, can be seen as a two-level extension of the path models for discrete variables as defined in Goodman's modified path approach (Goodman 1973) and extended to include latent variables by Hagenaars (1990) in the modified Lisrel approach. Moreover, the way in which these models allow for the decomposition of joint probability distributions in terms of products of conditional distributions, indicates their resemblance to the directed graph approach as described by, among others, Pearl (2009) for variables measured at a single level.

We opt for a latent class model with the number of latent classes set equal to the number of response categories of the observed individual-level variable, implying that the scores on $Z_{ij}$ are treated as indicators for a dichotomous latent variable $\zeta_j$ at the group level (score 0 or 1).[1] The number of latent

[1] It should be noted that the latent classes at the group-level underlying $Z_{ij}$ can not only be interpreted as a measurement model for the items, but also as a group-level discrete random effect since the dependence in the responses is summarized in one random score at the group-level. So, in fact, this is how the multilevel structure is taken into account. The predictor $X_j$, and the outcome $Y_j$ are observed at the group-level only, which means that these variables vary only between groups and not within groups.

classes at the group level does not necessarily have to be fixed a priori, but could also be data driven by comparing fit indices for models with varying number of latent classes.[2]

For dichotomous variables, the model can be formulated more formally in terms of two logit regression equations:

$$\text{Logit}[P(Z_{ij} = 1|\zeta_j)] = \log \left[ \frac{P(Z_{ij} = 1|\zeta_j)}{P(Z_{ij} = 0|\zeta_j)} \right] = \beta_1 + \beta_2 \zeta_j \ , \tag{1}$$

and

$$\text{Logit}[P(Y_j = 1|\zeta_j)] = \log \left[ \frac{P(Y_j = 1|\zeta_j)}{P(Y_j = 0|\zeta_j)} \right] = \beta_3 + \beta_4 \zeta_j \ , \tag{2}$$

in which $\beta_1$ and $\beta_3$ are intercepts and $\beta_2$ and $\beta_4$ slopes. The parameters $\beta_2$ and $\beta_4$ are log odds ratios indicating the strength of the association between the latent variable $\zeta_j$ and the observed variables $Z_{ij}$ and $Y_j$, respectively.

For the general case of $K$ nominal response categories for $Z_{ij}$ and $M$ nominal response categories for $Y_j$, multi category logit models can be formulated as described in Agresti (2007).

[2] The number of latent classes could, for example, be determined with the BIC using the number of groups as sample size in the formula (Lukočienė, Varriale and Vermunt 2010)

## ESTIMATION METHODS FOR THE LATENT VARIABLE APPROACH

For continuous outcomes, Croon and van Veldhoven (2007) proposed a stepwise estimation method in which the two parts of the model are estimated separately by what they called an 'adjusted regression analysis'. In this approach the aggregated group means of the variables measured at the individual level are adjusted in such a way that a regression analysis at the group level using these adjusted group means produces consistent estimates of the regression coefficients. Full information maximum likelihood (FIML) estimates can be obtained by either the 'persons as variables approach' (Curran 2003; Metha and Neale 2005), or by fitting the model as a two-level structural equation model (Lüdtke et al. 2008) as made possible in software packages like Mplus (Muthén and Muthén 2010), LISREL (Jöreskog and Sörbom 2006), or EQS (Bentler 1995). These maximum likelihood methods estimate the parameters from the two parts of the model simultaneously.

Applied to the 1-2 model with discrete data, let $\mathbf{Z}_j$ be the vector containing the $I_j$ individual-level responses for group $j$; that is $\mathbf{Z}_j = \{Z_{1j}, Z_{2j}, ..., Z_{I_j j}\}$ . The joint density of $\mathbf{Z}_j$, $Y_j$ and $\zeta_j$ equals

$$\begin{aligned} P(\mathbf{Z}_j, Y_j, \zeta_j) &= P(\zeta_j)P(Y_j|\zeta_j)P(\mathbf{Z}_j|\zeta_j) \\ &= \underbrace{P(\zeta_j)P(Y_j|\zeta_j)}_{between} \underbrace{\left\{\prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j)\right\}}_{within} \end{aligned} \tag{3}$$

It consists of a product of a between and a within component. In the between component only relations among variables defined at the group-level are defined, whereas in the within component the individual-level scores are related to the group-level variable. By taking the product of $P(\mathbf{Z}_j, Y_j, \zeta_j)$ over all groups, the complete data likelihood is obtained which is the likelihood function if $\zeta_j$ would have been observed. The log likelihood for the observed data is then obtained by summing $\log(P(\mathbf{Z}_j, Y_j, \zeta_j))$ over all groups.

Integrating out the latent variable $\zeta_j$ from the complete log likelihood by summing over its possible values yields the log likelihood function for the observed data $Z_{ij}$ and $Y_j$; that is,

$$\log \mathcal{L} = \sum_{j=1}^{J} \log \left[ \sum_{l=1}^{L} \left[ P(\zeta_j = l)P(Y_j|\zeta_j = l) \prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j = l) \right] \right] \tag{4}$$

in which $L$ represents the number of latent classes.

In practice, this incomplete data likelihood function can be constructed in two equivalent ways: with the ‘two-level regression approach’ and with the 'persons as variables approach'(Curran 2003; Metha and Neale 2005). For the first approach, data need to be organized in a 'long file' while for the second approach the data need to be organized in a 'wide file'. More details about these equivalent approaches and the construction of the likelihood accordingly, can be found in Appendix A. The Latent GOLD software (Vermunt and Magidson 2005) can be used to estimate the model in both ways.

## SIMULATION STUDY TO EVALUATE LATENT CLASS APPROACH IN 1-2 MODEL

### *Aim of the simulation*

This section reports the results of a Monte Carlo simulation study which evaluated the (statistical) performance of the latent class approach for analyzing micro-macro relations among dichotomous variables using the 1-2 model. A first aim of the simulation study is to investigate the bias of the ML estimates of the relevant regression parameters describing the micro-macro relationship. Additionally, the power and observed type-I error rate of the test of the regression coefficients are determined.

Two different ways to test for the significance of individual parameters are compared. First, significance is tested by means of the Wald test. This test is easy to implement and only requires the maximum likelihood estimation of the unrestricted model (leaving the estimation of $\beta$ free). However, evidence exists that in small samples the likelihood ratio test may be preferred (Agresti 2007). The latter testing procedure requires estimating both the unrestricted model and restricted model with $\beta = 0$.

Besides looking at the absolute performance of the latent class approach, its relative performance is assessed by comparing it to three more traditional approaches: mean aggregation, mode aggregation, and disaggregation. The present simulation study investigates how, for all four approaches, the bias in the parameter estimates, their type-I error rate and the power of the associated tests are affected by (1) the strength of the micro-macro relation, (2) the degree to which the individual-level scores reflect the (latent) group-level score, and (3) the sample sizes at both the individual and group level.

### *Method*

Data were generated according to the 1-2 model shown in Figure 1 and formally described by Equations 1 and 2. In the population model, four factors were systematically varied. First, the micro-macro relation was assumed to be either absent, ($\beta_4 = 0$), moderate ($\beta_4 = 1$), or strong ($\beta_4 = 2$). Second, the individual-level observed variable $Z_{ij}$ was either a poor ($\beta_2 = 1$), a good ($\beta_2 = 3$), or a perfect indicator ($\beta_2 = 200$) of the construct at the group level. In most applications the later assumption is unrealistic but it was included in order to compare the other two situations with the perfect situation. Third, the number of groups was set to either 40 or 200, and, fourth, the number of individuals within a group was either 10 or 40. Finally, the intercept values $\beta_1$ and $\beta_3$ were not varied independently, but were chosen so to guarantee uniform marginal distributions for $Z_{ij}$ and $Y_j$, implying that these marginal distributions were held constant across simulation conditions. Completely crossing the four factors resulted in $3 \times 3 \times 2 \times 2 = 36$ conditions. For each condition, 100 data sets were generated with Latent GOLD (Vermunt and Magidson 2005).

Each data set was analyzed in four different ways. First, they were analyzed according to the latent class approach and the estimate of the micro-macro regression coefficient is represented by the term $\beta_4$ from Equation 2.

Second, the same data were analyzed at the group level by aggregating the individual-level predictor scores using the group means, $\bar{Z}_{.j}$, or, third, using the group mode, denoted by $\breve{Z}_{.j}$. The logistic regression analyses at the group level are defined by

$$\text{Logit}[P(Y_j = 1|\bar{Z}_{.j}] = \beta_5 + \beta_6 \bar{Z}_{.j} \tag{5}$$

and

$$\text{Logit}[P(Y_j = 1|\breve{Z}_{.j}] = \beta_7 + \beta_8 \breve{Z}_{.j} \ . \tag{6}$$

The estimate of the micro-macro regression coefficient is now represented by $\beta_6$ and $\beta_8$, respectively.

Finally, in the fourth analysis the group-level outcome variable $Y_j$ is disaggregated to the individual level by assigning the group score to every group member as if the score was unique to the individual, so $Y_{ij} = Y_j$ for each individual $i$ in group $j$. The disaggregated variable $Y_{ij}$ is then regressed on $Z_{ij}$ at the individual level and the corresponding logistic regression equation becomes

$$\text{Logit}[P(Y_{ij} = 1|Z_{ij})] = \beta_9 + \beta_{10} Z_{ij} \ . \tag{7}$$

The estimate of the micro-macro regression coefficient is now represented by $\beta_{10}$.

Power was determined with a Wald test by computing the percentage of times that the hypothesis $\beta = 0$ was rejected when in fact there was a non-zero effect present in the population ($\beta = 1$ and $\beta = 2$). The observed type-I error rate was given by the proportion of significant results for the same hypothesis when there was zero effect in the population ($\beta = 0$). The observed type-I error rate and power of the likelihood ratio test were determined in a similar way. In order to assess the main effects of each of the manipulated factors, the results were collapsed over the three other factors.

***Results***

### Bias of the parameter estimates

The mean and standard deviations of the estimates of the micro-macro relation are summarized in Table 2.

Table 2 about here

When the micro-macro relation was estimated with the latent class approach, the micro-macro effect was estimated without severe bias in all conditions. When $Z_{ij}$ was aggregated to the group level using mean scores, the estimated micro-macro effect was overestimated in all conditions where a micro-macro relation was present, except when the individual-level scores perfectly reflected the construct at the group level. When the mode instead of the mean was used to aggregate the individual-level scores, the bias decreased. This method also seems to work when the individual-level scores were good, and not necessarily perfect, indicators of the construct at the group level. When $Y_j$ was disaggregated to the individual level, the estimated micro-macro effect is estimated with a downwards bias, except when the individual-level scores perfectly reflected the construct at the group level. When the true micro-macro relation was absent in the population, all four approaches estimated the effect unbiasedly.

The information in Table 2 indicates that increasing the number of groups from 40 to 200 reduces the bias of the estimates a little, and leads to much smaller standard deviations of the estimates for all four approaches. Increasing the number of group members from 10 to 40, improving the quality of the individual-level scores to reflect the group-level construct, or increasing the effect size of the micro-macro relation did not cause large changes in the bias of the mean estimates, nor in the value of their standard deviations.

## Power and observed type-I error rates

The results with respect to power and type-I error rates were also collapsed for each factor over the three remaining factors and are shown in Table 3.

Table 3 about here

The observed power to detect the micro-macro effect could be determined in the 24 conditions in which an effect was present in the population. For the latent class approach, mean aggregation, and mode aggregation, the observed power was, larger than 0.70 when the true effect was large. A moderate micro-macro effect could only be detected with power larger than .70 in samples with 200 groups. When disaggregating, power is always above 0.70, except when the individual-level scores are poor indicators of the group-level construct.

The observed type-I error rates could be evaluated in the 12 conditions with a zero micro-macro effect in the population. In these conditions the observed type-I error rate was expected to lie between .02 and .09 with a probability of 0.935.[3] When the data were analyzed with the latent class approach, mean aggregation or mode aggregation, all the observed type-I error

[3] This probability is based on a binomial distribution with 100 trials and a success probability equal to .05

rates lay between these boundaries. When $Y_j$ is disaggregated to the individual level, the observed type-I error rates were unacceptable high, ranging from 0.18 to 0.60, indicating that this approach leads to an unacceptably liberal significance test for the micro-macro effect.

Increasing the sample sizes, the quality of the individual-level scores to reflect the construct at the group level, or the effect size all lead to increased power, regardless of the way in which the micro-macro relation is modeled. On the other hand, the observed type-I error rates do not seem to vary as a function of the four manipulated factors. The results reported above are very similar for the Wald and the likelihoodratio test.

### *Conclusion*

Overall the latent class approach obtains unbiased parameters even when the individual-level scores poorly reflect the (latent) group-level score with reasonable power and type-I error rate. Aggregation only works with perfect (mean aggregation) or good (mode aggregation) indicators, which are however rather unrealistic conditions in practice. Using disaggregation, the observed type-I error rates were unacceptable high so this approach should be avoided anyhow. Since the latent class approach estimates the 1-2 model with dichotomous variables better than the other 3 approaches, only this approach is evaluated in a more complex model.

## ADDING A LEVEL-2 PREDICTOR TO THE MODEL

The 1-2 model can be extended to a 2-1-2 model by adding a predictor $X_j$ at the group level as shown in Figure 2. In the present discussion $X_j$ is assumed to be dichotomous, but the extension to the general case of $Q$ response categories or to continuous variables is straightforward.

Figure 2 about here

At the group level two logistic regression equations are defined and a latent class model is used to link the individual and group level, so that for dichotomous data the model can be formulated in terms of three logit regression equations:

$$\mathrm{Logit}(P(\zeta_j = 1|X_j)) = \beta_1 + \beta_2 X_j \ , \tag{8}$$

$$\mathrm{Logit}(P(Y_j = 1|X_j, \zeta_j)) = \beta_3 + \beta_4 X_j + \beta_5 \zeta_j + \beta_6 X_j \cdot \zeta_j \ , \tag{9}$$

and

$$\text{Logit}(P(Z_{ij} = 1|\zeta_j)) = \beta_7 + \beta_8\zeta_j \ , \tag{10}$$

in which $\beta_1$, $\beta_3$ and $\beta_7$ are intercepts and $\beta_2$, $\beta_4$, $\beta_5$, $\beta_6$, and $\beta_8$ slopes. The regression model for $Y_j$ contains the main effects of $\zeta_j$ and $X_j$ and their mutual interaction effect represented by the product variable $X_j \cdot \zeta_j$. Furthermore, $\zeta_j$ itself is regressed on $X_j$.

The joint probability density of $X_j$, $\mathbf{Z}_j$, $Y_j$, and $\zeta_j$ for an arbitrary group $j$ is defined as

$$\begin{aligned} P(X_j, \mathbf{Z}_j, Y_j, \zeta_j) &= P(X_j)P(\zeta_j|X_j)P(Y_j|X_j,\zeta_j)P(\mathbf{Z}_j|\zeta_j) \\ &= \underbrace{P(X_j)P(\zeta_j|X_j)P(Y_j|X_j,\zeta_j)}_{between} \underbrace{\left\{\prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j)\right\}}_{within} , \end{aligned} \tag{11}$$

while the observed or incomplete data log likelihood function is

$$\log \mathcal{L} = \sum_{j=1}^{J} \log \left[ \sum_{l=1}^{L} \left[ P(X_j)P(\zeta_j = l|X_j)P(Y_j|X_j, \zeta_j = l) \prod_{i=1}^{I} P(Z_{ij}|\zeta_j = l) \right] \right] , \tag{12}$$

in which $L$ represents the number of latent classes. The likelihood function can be maximized in the same two ways as described for the 1-2 model in

Appendix A, namely the 'persons-as-variables approach' and the 'two-level regression approach', requiring the data to be appropriately structured. The model can again be estimated with the Latent GOLD software (Vermunt and Magidson 2005).

## SIMULATION STUDY TO EVALUATE LATENT CLASS APPROACH IN 2-1-2 MODEL

### *Aim of the simulation study*

The latent class approach, which seems to work well for a simple micro-macro relation with dichotomous variables, is now evaluated in the slightly more complex 2-1-2 model. The Monte Carlo simulation study reported in this section intends to investigate how the bias in parameter estimates, the type-I error rates, and the power of tests for individual regression coefficients are influenced by (1) the strength of the true relations, (2) the degree to which the individual-level scores reflect the latent group-level score, and (3) the sample sizes at both the individual and group level. As in the previous simulation study, the significance of the parameters is evaluated with both Wald and likelihoodratio tests.

***Method***

Data are generated according to the 2-1-2 Model shown in Figure 2 and formally described by Equations 8, 9, and 10. In the population models, all three main effects at the macro-level were assumed to be either absent ($\beta = 0$), moderate ($\beta = 1$) or strong ($\beta = 2$) and the interaction effect between $X_j$ and $\zeta_j$ was either negative ($\beta_6 = -1$), absent ($\beta_6 = 0$), or positive ($\beta_6 = 1$). The scores on $Z_{ij}$ were either poor indicators ($\beta_8 = 1$), good indicators (($\beta_8 = 3$), or perfect indicators($\beta_8 = 200$) of the latent group score $\zeta_j$. The number of groups was set to either 40 or 200, and the number of individuals within a group to either 10 or 40. The intercept values $\beta_1$ and $\beta_3$, and $\beta_7$ were determined in such a way that the marginal distributions of $Z_{ij}$, $\zeta_j$, and $Y_j$ were uniform. The marginal probability of $X_j$ was made uniform. Crossing the 7 factors resulted in $3 \times 3 \times 3 \times 3 \times 2 \times 2 = 972$ conditions.

Again 100 data sets were generated for each condition using Latent GOLD (Vermunt and Magidson 2005) and the data sets were analyzed with the latent class approach. Power and observed type-I error rates were determined for both the Wald and likelihoodratio tests as described in the method section of the previous simulation study. The power for the main effects of $X_j$ and

$\zeta_j$ on $Y_j$ was only determined in those conditions in which there was no interaction between $X_j$ and $\zeta_j$ in the population. In order to assess the (main) effect of a particular factor in the simulation study, the results obtained in the different conditions were collapsed over the other factors.

***Results***

### Bias in the parameter estimates

A summary of the estimated effects at the group level is given in Table 4.

Table 4 about here

First, Table 4 indicates that there is some bias in the estimates. Moreover, the magnitude of the bias seems to be proportional to the value of true effect since there is no bias when the true effect equals zero. Bias slightly decreases when the number of groups is increased, but remains about the same when the number of individuals within a group is increased, or when the quality of the individual-level scores reflecting the latent group-level score is improved. The standard deviations of the estimates are quite large and, consistent with the first simulation study, only increasing the number of groups reduces the standard deviations. Increasing the number of group members and the quality of the indicators have only small effects on the standard deviations.

**Power and observed type-I error rates**

The results with respect to power and type-I error rates are summarized in Table 5.

Table 5 about here

The power of the macro-level effects could be observed in the 646 conditions in which their was a non-zero effect in the population. The results can be summarized as follows. First, the power of the test $H_0 : \beta_2 = 0$ for the main effect of $X_j$ on $\zeta_j$ is larger than 0.70 when the true value of the effect is strong. When the true effect is moderate, power is above .70 when the number of groups is 200 but for the other factors the power to detect an moderate effect of $X_j$ on $\zeta_j$ lies between .26 and .63. The results are similar for the Wald and likelihoodratio tests. Second, the power to test $H_0 : \beta_4 = 0$ for the main effect of $X_j$ on $Y_j$ and the power of the test $H_0 : \beta_5 = 0$ for the main effect of $\zeta_j$ on $Y_j$ are above .70 when the true effects are strong except for $\beta_5 = 2$ with 40 groups. Moderate main effects can again only be detected with sufficient power when the number of groups is 200. For the other factors the power to detect moderate main effects lies between .22 and .61 and although the obtained power is a little larger with a likelihood ratio test compared to a Wald test, the difference is rather small. Third, the power of the test $H_0 : \beta_6 = 0$ for the interaction effect of $X_j$ and $\zeta_j$ on $Y_j$ is very low but higher for the likelihood ratio test than for the Wald test, especially when there are only 40 groups. For the Wald test the power to detect an interaction effect lies between .02 and .30 while for the likelihood ratio test power lies between .11 and .32.

The observed type-I error rates could be evaluated in the 324 conditions in which the macro-level effect was absent. As before, the type-I error rate was expected to lie between 0.02 and 0.09. This was indeed the case except that the observed type-I error rates were too low for the test of $\beta_6$ when determined with a Wald test in the conditions with 40 groups. The observed type-I error rates seem to be independent of the manipulated factors. Within acceptable boundaries, the Wald test seems to be slightly too conservative while the likelihood ratio test seems to be slightly too liberal.

### *Conclusion*

From this second simulation study it can be concluded that the latent class approach produces almost unbiased parameters in the 2-1-2 model but standard deviations are quite high and can be reduced by using a large number of groups. Especially for the interaction effect, the power is low in most conditions but can be improved by using a likelihoodratio test instead of a Wald test. The type-I error rates seem correct with both the Wald and the likelihood ratio tests.

## EMPIRICAL DATA EXAMPLE

The discrete latent variable approach is illustrated with an empirical application on data from personal networks, in which individuals (egos) are interviewed together with persons from their network (alters). Up till now only research questions could be answered when the dependent variable was defined at the lowest level of the alters or ties (van Duijn, Busschbach, and Snijders 1999; Snijders et al. 1995). The latent variable approach allows to answer research questions with a dependent variable at the higher ego-level, so providing new possibilities for investigating a broad range of research questions in studies of personal networks. More specifically, in the current example the effect of belonging to a particular type of personal network on the behavior of the ego himself is explored.

### *Data and conceptual model*

The data come from the 'Netherlands Kinship Panel Study' (NKPS), which is a large-scale database on Dutch families that yields information for individual respondents (the egos) and some of their family members and friends (the alters). The data are publicly available and can be retrieved from http://www.nkps.nl. For the present example, data were available for 8161 egos with maximally six alters nested within each ego: the parents in law, two siblings, two children, and a friend.

Kalmijn and Vermunt (2007) used the same data to investigate whether selection in networks is based on age and marital status, but in the present paper a different perspective is chosen. Instead of expecting that persons choose the persons in their network based on their marital status, we suppose that egos are members of a network in which either many or few people are married. The latent variable $\zeta_j$ then represents latent class membership of an ego's network: $\zeta_j = 0$ if the ego belongs to a network in which few members are married versus $\zeta_j = 1$ if the ego belongs to a network in which many members are married. The marital status of the alters, $Z_{ij} = 0$ for unmarried alters and $Z_{ij} = 1$ for married alters, are taken as exchangeable indicators of the type of network an ego belongs to. The dependent level-2 variable in this analysis is the dichotomous variable $Y_j$ indicating whether an ego is married or not: $Y_j = 0$ if the ego is not married versus $Y_j = 1$ if the ego is married. The religiosity of the ego, $X_j = 0$ if ego $j$ is not religious, $X_j = 1$ if ego $j$ is religious, is treated as the level-2 explanatory variable that affects the probability of an ego to belong to a particular type of network. Eggebeen and Dew (2009) already pointed out that religion is a very important factor in family formation during young adulthood. In the present analysis it is expected that non-religious persons rather belong to the latent class with few married members than to the class with many married members. For religious people, we expect the opposite. Furthermore, we allow for an interaction effect of type of network and religiosity on the dependent variable, implying that the effect of the network on being married can be different for religious and not-religious) persons. The model as formulated here can be extended in several ways. First, the exchangeability assumption, stating that all alters are equivalent indicators of the type of network, can eventually be relaxed if the parents in law, siblings, children, and friend to the network provide (partly) different network information. Second, if necessary, a model with more than two latent classes at the network level could be considered. These extensions will not be further discussed here.

### *Method*

The model, shown in Figure 2, is defined by Equations 8, 9, and 10 and the model parameters can be estimated with the software package Latent GOLD (Vermunt and Magidson 2005) by applying either the 'two-level regression' or the 'persons as variables approach' as described in Appendix B.

### *Results*

Since the 'persons as variables approach' and the 'two-level regression approach' yield the same results, only the results of the 'two-level regression approach' are presented here. Looking at the regression coefficients in Table 6, it can be seen that the Wald tests for all slopes coefficients are significant at at least the 5%-level, except the main effects of $X_j$ and $\zeta_j$ on the level-2 outcome variable $Y_j$. Their interaction effect, however, is significant.

Table 6 about here

By substituting the estimated parameter values in the logit regressions equations 8, 9, and 10 and transforming them into the probability scale, the probabilities as given in Table 7(a), 7(b), and 7(c) are obtained.

Table 7 about here

As can be seen from Table 7(a), alters in the two network classes have a probability of being married of 0.43 and 0.60, respectively. So, the latent classes can be interpreted in terms of the egos belonging to a network with either a minority or a majority of married persons.

Second, Table 7(b) indicates that when an ego is not religious, the probability of having a network in which the majority of the persons is married is 0.79 while it is 0.49 for an ego that is religious.

Third, Table 7(c) shows that the probability of an ego being married depends on whether he is religious or not, and on the type of network the ego belongs to. The egos that are religious and have a network of mostly married persons, have the highest probability of being married (0.99) whereas for egos that are not religious and have a network of mostly married persons that probability is equal to 0.31. The egos that are not religious and have a network in which a minority is married, and the egos that are religious and have a network in which a minority is married have both a very low probability of being married themselves (0.02 and 0.01, respectively). So, the positive effect of belonging to a married network on the probability of being married is much stronger for religious persons than for non-religious persons.

### *Conclusion*

An interesting thing emerging from this analysis is the strong interaction effect of religiosity and latent class membership at the network level, showing that the positive effect of having a married network on the probability of being married is stronger for religious persons compared to non religious persons. Contrary to our expectations, religiosity has a negative affect ($\hat{\beta}_2 = -1.39$) on the probability of belonging to a network with many married people. Table 8(a) also confirms that non-religious people have a probability of .79 of belonging to a network with many married people, whereas for religious people this probability is only .49. We have no clear-cut explanation for this counter-intuitive result.

## DISCUSSION

Although a wide variety of research questions in the social and behavioral sciences involve micro-macro relations, specific methods to analyze such relationships are not yet fully developed. The current article is contributing to this development by showing how a latent variable approach which was originally proposed for continuous outcomes (Croon and van Veldhoven, 2007) can be modified for the application to discrete outcomes.

We showed that, in a simple 1-2 model, the latent variable approach outperforms more traditionally aggregation and disaggregation strategies with respect to bias with reasonable power and correct observed type-I error rates. In a more complex 2-1-2 model, there is small bias and standard deviations are a little higher. These can be reduced by using a larger number of groups. Power is acceptable for the main effects but relatively low for the interaction effect. The latter might be due to general power problems associated with detecting interaction effects by including product terms in the regression equation (McClelland and Judd 1993; Whisman and McClelland 2005). Using a likelihood ratio test instead of a Wald test increases power. Observed type-I error rates are correct although. Overall, the latent variable approach seems to work well for analyzing micro-macro relations with discrete variables and this enables investigating research questions that could not be addressed appropriately before.

The current research was restricted to models with only one lower-level predictor. Further research should be devoted to models with multiple level-1 variables. In this context it might be more practical to use 3-step estimation procedures as described by Bakk, Tekle and Vermunt (in press), instead of the currently suggested 1-step estimation procedure. Furthermore, in the current article the focus was set at two-level situations in which the predictors and outcome variable were observed variables. It would be interesting to explore the possibilities to extend the model to the situation in which the outcome variable and/or predictors are latent constructs measured with multiple indicators.

# References

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis.* Hoboken, NJ: John Wiley & Sons.

Bakk, Zsuzsa, Fetene Tekle, and Jeroen K. Vermunt. In press. “Estimating the Association between Latent Class Membership and External Variables using Bias Adjusted Three-step Approaches. *Sociological Methodology.*

Bartholomew, David J. and Martin Knott. 1999. *Latent Variable Models and Factor Analysis: Kendall's Library of Statistics 7.* London, United Kingdom: Arnold.

Bentler, Peter M. 1995. *EQS Structural Equations Program Manual.* Encino, CA: Multivariate Software.

Croon, Marcel A. and Marc J. P. M. van Veldhoven. 2007. “Predicting Group-level Outcome Variables from Variables Measured at the Individual Level: A Latent Variable Multilevel Model.” *Psychological Methods* 12(1): 45-57.

Curran, Patrick J. 2003. “Have multilevel models been structural equation models all along?" *Multivariate Behavioral Research* 38(4): 529-569.

DeShon, Richard P., Steve W. J. Kozlowski, Aaron M. Schmidt, Karen R. Milner, and Darin Wiechmann. 2004. " Multiple-goal Multilevel Model of Feedback Effects on the Regulation of Individual and Team Performance." *Journal of Applied Psychology* 89(6): 1035-1056.

Eggebeen, David and Jeffrey Dew. 2009. "The Role of Religion in Adolescence for Family Formation in Young Adulthood." *Journal of Marriage and Family* 71(1): 108-121.

Emberson, Susan E. and Steven P. Reise. 2000. *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fox, Jean-Paul. 2005. "Multilevel IRT using Dichotomous and Polytomous Items." *British Journal of Mathematical and Statistical Psychology* 58(1): 145172.

Fox, Jean-Paul and Cees A. W. Glas. 2003. "Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory." *Psychometrika* 68(2): 169-191.

Goldstein, Harvey. 2003. *Multilevel statistical models.* London, United

Kingdom: Arnold.

Goodman, Leo A. 1973. "The Analysis of Multidimensional Contingency Tables when Some Variables are Posterior to Others: A Modified Path Analysis Approach" *Biometrika* 60(1): 179-192.

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

Hagenaars, Jacques A. P. and McCutcheon, Allan L., ed. 2002. *Applied Latent Class Analysis.* Cambridge, United Kingdom: Cambridge University Press.

Hox, Joop J. and J. Kyle Roberts. 2011. *Handbook of Advanced Multilevel Analysis.* New York: Routledge Taylor & Francis Group.

Jöreskog, Karl G. and Dag Sörbom. 2006. *LISREL 8.8 for Windows.* Lincolnwood, IL: Scientific Software International.

Kalmijn, Matthijs and Jeroen K. Vermunt. 2007. "Homogeneity of Social Networks by Age and Marital Status: A Multilevel Analysis of Ego-centered Networks." *Social Networks* 29(1): 25-43.

Keith, Timothy Z. 2005. *Multiple Regression and Beyond.* Boston, MA: Pearson Education.

Krull, Jennifer L. and David P. MacKinnon. 1999. "Multilevel Mediation Modeling in Group-based Intervention Studies." *Evaluation Review* 23(4): 418-444.

Lukočienė, Olga, Varriale, Roberta, and Vermunt, Jeroen K. 2010. The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. Sociological Methodology, 40(1), 247-283.

Lüdtke, Oliver, Herbert W. Marsh, Alexander Robitzsch, Ulrich Trautwein, Tihomir Asparouhov, and Bengt Muthén. 2008. "The Multilevel Latent Covariate Model: A New, more Reliable Approach to Group-level Effects in Contextual Studies." *Psychological Methods* 13(3): 203-229.

Lüdtke, Oliver, Herbert W. Marsh, Alexander Robitzsch, and Ulrich Trautwein. 2011. "A 2 x 2 Taxonomy of Multilevel Latent Contextual Models: Accuracy-bias Trade-offs in Full and Partial Error Correction Models". 2011. *Psychological Methods* 16(4): 444-467.

MacKinnon, David P. 2008. *Introduction to Statistical Mediation Analysis.*

New York, NY: Lawrence Erlbaum Associates Taylor & Francis Group.

McClelland, Gary H. and Charles M. Judd. 1993. "Statistical Difficulties of Detecting Interactions and Moderator Effects". *Psychological Bullitin* 114(2): 376-390.

Metha, Paras D. and Michael C. Neale. 2005. "People are Variables too: Multilevel Structural Equations Modeling." *Psychological Methods*: 10(3): 259-284.

Muthén, Linda K. and B. O. Muthén. 2010. *Mplus Users Guide sixth edition.* Los Angeles, CA: Muthén & Muthén.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference.* $2^{nd}$ ed. Cambridge, United Kingdom: Cambridge University Press.

Rasbash, Jon, Christopher Charlton, William J. Browne, Michael Healy, and Bruce Cameron. 2005. *MLwiN Version 2.02.* Bristol, United Kingdom: Centre for Multilevel Modelling, University of Bristol.

Rutter, Michael and Barbara Maughan. 2002. "School Effectiveness Findings 1979-2002." *Journal of School Psychology* 40(6): 451-475.

Skrondal, Anders and Sophia Rabe-Hesketh. 2004. *Generalized Latent*

*Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.* Boca Raton, FL: Chapman & Hall/CRC.

Snijders, Tom A. B. and Roel J. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* London, United Kingdom: SAGE.

Snijders, Tom A. B., Marinus Spreen, and Ronald Zwaagstra. 1995. "The Use of Multilevel Modeling for Analysis of Personal Networks: Networks of Cocaine Users in an Urban Area." *Journal of Quantitative Anthropology* 5(2): 85-105.

van Duijn, Marijtje A. J., Jooske T van Busschbach, and Tom A.B Snijders. 1999. "Multilevel Analysis of Personal Networks as Dependent Variables." *Social Networks* 21(2): 187-209.

van Veldhoven, Marc J. P. M. 2005. "Financial Performance and the Long-term Link with HR Practices, Work Climate and Job Stress." *Human Resource Management Journal* 15(4): 30-53.

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

Waller, Mary J., Jeffrey M. Conte, Cristina B. Gibson, and Mason A. Carpenter. 2001. "The Effect of Individual Perceptions of Deadlines on Team Performance." *The Academy of Management Review* 26(4): 586-600.

Whisman, Mark A. and Gary H. McClelland. 2005. "Designing, Testing, and Interpreting Interactions and Moderator Effects in Family Research." *Journal of Family Psychology* 19(1): 111-120.

# A TWO EQUIVALENT ESTIMATION PROCEDURES

This appendix shows how the likelihood function of the 1-2 model as defined in Equation 4 can be constructed in two equivalent ways, that is, with the 'two-level regression approach' and with the 'persons as variables approach' (Curran 2003; Metha and Neale 2005).

## Two-level regression

The 'two-level regression approach' is illustrated in Figure 1. In practice, the group-level variables are treated as individual-level variables but the group-level score of a particular group is assigned to a single individual from that group, while the scores of the other individuals within that group on this variable are defined as missing. Note that this is not the same as disaggregating the group-level variable since that would come down to assigning the group score to each and every group member. Since the individuals within the same group are exchangeable, it does not matter to which individual the group-level score is assigned, but for convenience it will be assumed here that assignment is to the first individual in a group.

The data are stored in a long file in which each row of the data matrix corresponds to an individual, but an additional group identification variable is defined that indicates to which group an individual belongs. The group-level outcome defined as an individual-level variable is denoted by $Y_{ij}^*$, so that $Y_{ij}^* = Y_j$ for $i = 1$, and $Y_{ij}^*$ is missing for $i \neq 1$. The variables originally measured at the individual level are simply reproduced in the data matrix. Table 8 provides an example data matrix with three groups, the first two groups consisting of three individuals, and the third group of two individuals.

Table 8 about here

The joint density of $\mathbf{Z}_j$, $\mathbf{Y}_j^*$ and $\zeta_j$ equals

$$\begin{aligned} P(\mathbf{Z}_j, \mathbf{Y}_j^*, \zeta_j) &= P(\zeta_j)\left\{\prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j)P(Y_{ij}^*|\zeta_j)\right\} \qquad (13)\\ &= P(\zeta_j)\left\{\prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j)\right\}\left\{\prod_{i=1}^{I_j} P(Y_{ij}^*|\zeta_j)\right\}\\ &= P(\zeta_j)\left\{\prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j)\right\}P(Y_{1j}^*|\zeta_j)\left\{\prod_{i=2}^{I_j} P(Y_{ij}^*|\zeta_j)\right\} \end{aligned}$$

Aggregating over the missing values $Y_{2j}^*, Y_{3j}^*, ..., Y_{I_j j}^*$ in applying full information maximum likelihood yields

$$\begin{aligned} P(\mathbf{Z}_j, Y^*_{1j}, \zeta_j) &= P(\zeta_j)\left\{\prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j)\right\} P(Y^*_{1j}|\zeta_j)\left\{\prod_{i=2}^{I_j}\sum_{Y^*_{ij}} P(Y^*_{ij}|\zeta_j)\right\} \\ &= P(\zeta_j)\left\{\prod_{i=1}^{I_j} P(Z_{ij}|\zeta_j)\right\} P(Y^*_{1j}|\zeta_j) \qquad (14) \end{aligned}$$

The latter simplification follows from the fact that $\sum_{Y^*_{ij}} P(Y^*_{ij}|\zeta_j) = 1$. Since $Y^*_{1j} = Y_j$, this is equivalent to the log likelihood described Equation 4.

## Persons as variables

The ‘persons as variables approach’ is illustrated in Figure 3 for the case that each group consists of maximum three members.

Figure 3 about here

Each of the three individuals within a group defines a different variable at the group level and as a consequence, there are as many ‘person variables’ as there are individuals in the groups. A separate equation is needed to describe the relationship between each ‘person variable’ and $\zeta_j$. Since the individuals from the same group are assumed to be exchangeable, the relationships between the different ‘person variables’ and $\zeta_j$ are required to be completely identical. As a consequence of these exchangeability constraints, it does not matter who is assigned to $Z_1$, who to $Z_2$, etc.. This approach can still be applied with unequal group sizes: the number of 'person variables' needed then is equal to the largest group size, and smaller groups have missing scores on the non used 'person variables'. In this way, the log likelihood described Equation 4 is obtained.

The 'persons as variables approach' requires the data matrix to be structured in a wide file format in which each of the rows represent a group, while its columns correspond to the persons variables as defined above. As an example Table 9 shows the data matrix with three groups, the first two groups consisting of three individuals, and the third group of two individuals.

Table 9 about here

# B LATENT GOLD SYNTAX EMPIRICAL EXAMPLE

This appendix explains how the 2-1-2 model from the Latent GOLD software (Vermunt and Magidson 2005) by either the 'persons as variables approach' or the ' two-level regression approach'. Estimation by the 'two-level regression approach' requires that the data are structured in a long file format with

8161 (# level-2 units) × 6 (# level-1 units) = 48966 rows. An indicator variable `egoid` is needed for identifying the different egos within which the alters are nested.

The relevant parts of the syntax for this approach are:

```
options
  bayes categorical=1
  missing includeall;
variables
  caseid      egoid;
  dependent   z nominal, y nominal;
  independent x nominal;
  latent      zeta nominal 2;
equations
  zeta        <- (b1)1 + (b2)x;
  y           <- (b3)1 + (b4)x + (b5)zeta + (b6)x*zeta;
  z           <- (b7)1 + (b8)zeta;
```

In the `options` section of syntax, all default settings can be accepted with two exceptions. First, `bayes categorical=1` is declared to prevent boundary solutions. Second, by default Latent GOLD applies list-wise deletion of cases with missing data. For obtaining maximum likelihood estimates with missing data, `missing includeall` should be declared. In the `variables` section the `egoid` variable should be defined as the `caseid`. In the same section a list of the dependent, independent, and latent variables should be provided. For nominal latent variables, the number of latent classes is specified after the definition of the scale type. The regression equations defining the model are formulated in the `equations` section of the syntax.

When the same model is estimated with the 'persons as variables approach', the results will be the same but the data file is constructed as a wide file with 8161 (level-2 units) rows corresponding to the different egos and with the columns corresponding to the variables defined on the egos and their alters. A separate equation has to be specified for each alter as shown in the part of the syntax that differs from the syntax of the 'persons as variables' approach:

```
variables
  dependent   z1 nominal, z2 nominal, z3 nominal,
              z4 nominal, z5 nominal, z6 nominal,
              y nominal;
  independent x nominal;
  latent      zeta nominal 2;
equations
  zeta <- (b1)1 + (b2)x;
  y    <- (b3)1 +  (b4)x + (b5)zeta + (b6)x*zeta;
  z1   <- (b7)1 + (b8)zeta;
  z2   <- (b7)1 + (b8)zeta;
  z3   <- (b7)1 + (b8)zeta;
  z4   <- (b7)1 + (b8)zeta;
  z5   <- (b7)1 + (b8)zeta;
  z6   <- (b7)1 + (b8)zeta;
```

To establish that the alters are exchangeable indicators of the latent variable at the ego-level, the regression coefficients are restricted to be equal using the arbitrary chosen value labels `(b7)` for the intercepts and `(b8)` for the indicator loadings. It is also possible to substitute the last six equations by `z1-z6 <- (b7)1 + (b8)zeta`.

Table 1: *Measurement model*

| | $\zeta_j = continuous$ | $\zeta_j = discrete$ |
|---|---|---|
| $Z_{ij} = continuous$ | linear factor model | latent profile model |
| $Z_{ij} = discrete$ | item response model | latent class model |

Table 2: Mean and standard deviations of estimates of micro-macro relationship estimated with latent class approach, mean aggregation, mode aggregation, and disaggregation, after collapsing

| | | Latent class | Mean aggregation | Mode aggregation | Disaggregation |
|---|---|---|---|---|---|
| | $\beta_4$ | $\bar{b_4}(\bar{SD})$ | $\bar{b_6}(\bar{SD})$ | $\bar{b_8}(\bar{SD})$ | $\bar{b_{10}}(\bar{SD})$ |
| $L_2 = 40$ | 0 | 0.00(0.79) | 0.05(1.25) | 0.00(0.64) | -0.01(0.43) |
| | 1 | 1.06(0.77) | 1.61(1.25) | 0.92(0.63) | 0.65(0.42) |
| | 2 | 2.18(0.84) | 3.41(1.45) | 1.87(0.70) | 1.29(0.48) |
| $L_2 = 200$ | 0 | 0.01(0.32) | 0.03(0.50) | 0.01(0.28) | 0.00(0.19) |
| | 1 | 1.01(0.33) | 1.58(0.56) | 0.91(0.28) | 0.64(0.19) |
| | 2 | 2.02(0.38) | 3.18(0.62) | 1.79(0.31) | 1.23(0.20) |
| $L_1 = 10$ | 0 | 0.02(0.61) | 0.04(0.76) | 0.03(0.46) | 0.01(0.32) |
| | 1 | 1.08(0.60) | 1.38(0.80) | 0.88(0.44) | 0.66(0.30) |
| | 2 | 2.13(0.65) | 2.79(0.86) | 1.71(0.48) | 1.26(0.33) |
| $L_1 = 40$ | 0 | -0.01(0.50) | 0.03(0.99) | -0.01(0.46) | -0.02(0.30) |
| | 1 | 0.99(0.50) | 1.81(1.00) | 0.96(0.48) | 0.63(0.30) |
| | 2 | 2.07(0.56) | 3.79(1.20) | 1.96(0.53) | 1.27(0.35) |
| $\beta_2 = 1$ | 0 | 0.04(0.74) | 0.14(1.47) | 0.05(0.46) | 0.01(0.16) |
| | 1 | 1.05(0.74) | 2.25(1.55) | 0.71(0.47) | 0.23(0.16) |
| | 2 | 2.21(0.78) | 4.79(1.79) | 1.43(0.49) | 0.47(0.14) |
| $\beta_2 = 3$ | 0 | 0.00(0.47) | 0.00(0.71) | 0.00(0.47) | 0.00(0.31) |
| | 1 | 1.02(0.46) | 1.50(0.70) | 1.00(0.45) | 0.65(0.29) |
| | 2 | 2.07(0.50) | 3.06(0.77) | 2.03(0.49) | 1.24(0.27) |
| $\beta_2 = 200$ | 0 | -0.02(0.45) | -0.02(0.45) | -0.02(0.45) | -0.02(0.46) |
| | 1 | 1.03(0.46) | 1.03(0.46) | 1.03(0.46) | 1.05(0.47) |
| | 2 | 2.03(0.54) | 2.03(0.54) | 2.03(0.54) | 2.07(0.60) |

Table 3: Power and observed type-I error rates of micro-macro relationship estimated with latent class approach, mean aggregation, mode aggregation, and disaggregation, after collapsing

| | | Latent class | | Mean aggregation | | Mode aggregation | | Disaggregation | |
|---|---|---|---|---|---|---|---|---|---|
| | | $P(H_0\,rejected)$ | | $P(H_0\,rejected)$ | | $P(H_0\,rejected)$ | | $P(H_0\,rejected)$ | |
| | $\beta_4$ | Wald | LR | Wald | LR | Wald | LR | Wald | LR |
| $L_2 = 40$ | 0 | .04 | .05 | .04 | .05 | .03 | .04 | .43 | .43 |
| | 1 | .24 | .28 | .25 | .29 | .25 | .27 | .71 | .71 |
| | 2 | .72 | .78 | .74 | .78 | .74 | .76 | .93 | .93 |
| $L_2 = 200$ | 0 | .05 | .06 | .05 | .06 | .05 | .06 | .41 | .41 |
| | 1 | .86 | .87 | .84 | .85 | .85 | .85 | .93 | .93 |
| | 2 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $L_1 = 10$ | 0 | .03 | .05 | .04 | .05 | .04 | .05 | .30 | .30 |
| | 1 | .50 | .54 | .51 | .53 | .51 | .52 | .77 | .77 |
| | 2 | .81 | .87 | .85 | .88 | .83 | .85 | .94 | .94 |
| $L_1 = 40$ | 0 | .06 | .06 | .05 | .06 | .05 | .06 | .54 | .54 |
| | 1 | .60 | .61 | .58 | .60 | .59 | .60 | .87 | .87 |
| | 2 | .90 | .91 | .89 | .90 | .90 | .90 | .99 | .99 |
| $\beta_2 = 1$ | 0 | .04 | .06 | .04 | .05 | .04 | .04 | .18 | .18 |
| | 1 | .40 | .45 | .41 | .43 | .40 | .42 | .59 | .60 |
| | 2 | .73 | .82 | .77 | .81 | .76 | .77 | .90 | .90 |
| $\beta_2 = 3$ | 0 | .05 | .06 | .06 | .06 | .05 | .06 | .48 | .48 |
| | 1 | .61 | .62 | .59 | .62 | .61 | .62 | .92 | .92 |
| | 2 | .94 | .95 | .94 | .95 | .95 | .95 | 1.00 | 1.00 |
| $\beta_2 = 200$ | 0 | .04 | .05 | .04 | .05 | .04 | .05 | .60 | .60 |
| | 1 | .64 | .65 | .64 | .65 | .64 | .65 | .95 | .95 |
| | 2 | .89 | .91 | .89 | .91 | .89 | .91 | 1.00 | 1.00 |

Table 4: Mean and standard deviations of estimates of group-level effects 2-1-2 Model estimated with latent class approach, after collapsing

| | $\beta_2$ | $\bar{b_2}(\bar{SD})$ | $\beta_4$ | $\bar{b_4}(\bar{SD})$ | $\beta_5$ | $\bar{b_5}(\bar{SD})$ | $\beta_6$ | $\bar{b_6}(\bar{SD})$ |
|---|---|---|---|---|---|---|---|---|
| $L_2 = 40$ | 0 | 0.00(0.75) | 0 | -0.02(1.40) | 0 | -0.03(1.47) | -1 | -1.23(2.10) |
| | 1 | 1.05(0.78) | 1 | 1.15(1.48) | 1 | 1.19(1.52) | 0 | -0.03(2.15) |
| | 2 | 2.08(0.86) | 2 | 2.38(1.60) | 2 | 2.35(1.64) | 1 | 1.17(2.30) |
| $L_2 = 200$ | 0 | 0.00(0.32) | 0 | -0.02(0.53) | 0 | -0.01(0.55) | -1 | -1.05(0.80) |
| | 1 | 1.01(0.34) | 1 | 1.02(0.54) | 1 | 1.03(0.57) | 0 | 0.01(0.82) |
| | 2 | 2.03(0.37) | 2 | 2.06(0.60) | 2 | 2.09(0.63) | 1 | 1.07(0.92) |
| $L_1 = 10$ | 0 | 0.00(0.58) | 0 | -0.02(1.01) | 0 | -0.02(1.09) | -1 | -1.14(1.55) |
| | 1 | 1.04(0.61) | 1 | 1.08(1.07) | 1 | 1.12(1.12) | 0 | 0.00(1.58) |
| | 2 | 2.06(0.67) | 2 | 2.20(1.14) | 2 | 2.21(1.21) | 1 | 1.12(1.72) |
| $L_1 = 40$ | 0 | -0.01(0.49) | 0 | -0.03(0.92) | 0 | -0.02(0.94) | -1 | -1.14(1.35) |
| | 1 | 1.02(0.51) | 1 | 1.09(0.95) | 1 | 1.11(0.97) | 0 | -0.03(1.38) |
| | 2 | 2.05(0.57) | 2 | 2.24(1.05) | 2 | 2.23(1.07) | 1 | 1.13(1.50) |
| $\beta_8 = 1$ | 0 | 0.00(0.68) | 0 | -0.02(1.12) | 0 | 0.02(1.24) | -1 | -1.17(1.77) |
| | 1 | 1.06(0.70) | 1 | 1.11(1.17) | 1 | 1.17(1.27) | 0 | -0.04(1.80) |
| | 2 | 2.07(0.75) | 2 | 2.22(1.24) | 2 | 2.26(1.35) | 1 | 1.10(1.92) |
| $\beta_8 = 3$ | 0 | -0.01(0.47) | 0 | -0.03(0.90) | 0 | -0.04(0.91) | -1 | -1.12(1.29) |
| | 1 | 1.02(0.49) | 1 | 1.08(0.93) | 1 | 1.07(0.92) | 0 | 0.01(1.34) |
| | 2 | 2.04(0.55) | 2 | 2.22(1.03) | 2 | 2.20(1.05) | 1 | 1.14(1.47) |
| $\beta_8 = 200$ | 0 | -0.01(0.46) | 0 | -0.02(0.88) | 0 | -0.03(0.89) | -1 | -1.12(1.29) |
| | 1 | 1.02(0.49) | 1 | 1.07(0.93) | 1 | 1.09(0.94) | 0 | -0.01(1.31) |
| | 2 | 2.04(0.55) | 2 | 2.22(1.03) | 2 | 2.20(1.02) | 1 | 1.12(1.45) |

Table 5: Power and observed type-I error rates of group-level effects 2-1-2 Model with Wald test and Likelihood ratio test, after collapsing

| | | $P(H_0\,rejected)$ | | | $P(H_0\,rejected)$ | | | $P(H_0\,rejected)$ | | | $P(H_0\,rejected)$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta_2$ | Wald | LR | $\beta_4$ | Wald | LR | $\beta_5$ | Wald | LR | $\beta_6$ | Wald | LR |
| $L_2 = 40$ | 0 | .04 | .05 | 0 | .03 | .06 | 0 | .03 | .07 | -1 | .04 | .12 |
| | 1 | .26 | .30 | 1 | .26 | .29 | 1 | .22 | .29 | 0 | .01 | .06 |
| | 2 | .73 | .78 | 2 | .72 | .73 | 2 | .61 | .69 | 1 | .02 | .11 |
| $L_2 = 200$ | 0 | .05 | .06 | 0 | .05 | .05 | 0 | .05 | .05 | -1 | .30 | .32 |
| | 1 | .87 | .88 | 1 | .87 | .86 | 1 | .80 | .81 | 0 | .05 | .06 |
| | 2 | 1.00 | 1.00 | 2 | 1.00 | .99 | 2 | .99 | .99 | 1 | .24 | .28 |
| $L_1 = 10$ | 0 | .04 | .05 | 0 | .04 | .06 | 0 | .04 | .07 | -1 | .15 | .20 |
| | 1 | .52 | .55 | 1 | .54 | .54 | 1 | .46 | .52 | 0 | .03 | .06 |
| | 2 | .82 | .86 | 2 | .84 | .83 | 2 | .75 | .81 | 1 | .11 | .17 |
| $L_1 = 40$ | 0 | .05 | .05 | 0 | .04 | .06 | 0 | .04 | .06 | -1 | .19 | .24 |
| | 1 | .61 | .62 | 1 | .59 | .60 | 1 | .55 | .58 | 0 | .03 | .06 |
| | 2 | .91 | .92 | 2 | .88 | .89 | 2 | .85 | .88 | 1 | .15 | .21 |
| $\beta_8 = 1$ | 0 | .04 | .06 | 0 | .04 | .06 | 0 | .03 | .07 | -1 | .11 | .17 |
| | 1 | .46 | .51 | 1 | .52 | .52 | 1 | .39 | .46 | 0 | .02 | .06 |
| | 2 | .75 | .82 | 2 | .82 | .79 | 2 | .66 | .75 | 1 | .08 | .14 |
| $\beta_8 = 3$ | 0 | .05 | .05 | 0 | .05 | .06 | 0 | .04 | .06 | -1 | .20 | .24 |
| | 1 | .62 | .63 | 1 | .58 | .60 | 1 | .57 | .59 | 0 | .03 | .06 |
| | 2 | .92 | .93 | 2 | .88 | .90 | 2 | .87 | .89 | 1 | .16 | .22 |
| $\beta_8 = 200$ | 0 | .05 | .05 | 0 | .04 | .05 | 0 | .04 | .06 | -1 | .20 | .25 |
| | 1 | .62 | .63 | 1 | .59 | .61 | 1 | .57 | .59 | 0 | .03 | .06 |
| | 2 | .92 | .92 | 2 | .88 | .90 | 2 | .87 | .89 | 1 | .16 | .22 |

Table 6: Regression coefficients empirical example

| Independent variable | $\beta$ | SE |
|---|---|---|
| *Dependent variable: network ego* | | |
| Intercept ($\beta_1$) | 1.34** | 0.32 |
| Religion ego ($\beta_2$) | -1.39** | 0.38 |
| *Dependent variable: married ego* | | |
| Intercept ($\beta_3$) | -4.17* | 2.10 |
| Religion ego ($\beta_4$) | -0.73 | 2.00 |
| Network ego ($\beta_5$) | 3.38 | 2.12 |
| Religion ego * Network ego ($\beta_6$) | 5.88* | 2.11 |
| *Dependent variable: married alter* | | |
| Intercept ($\beta_7$) | -0.28** | 0.07 |
| Network ego ($\beta_8$) | 0.69** | 0.07 |

* $p < .05$, ** $p < .01$

Table 7: Estimated probabilities empirical data example

(a)

| network ego | P(married alter = 1 \| network ego) (SE) |
|---|---|
| 0 | .43 (0.017) |
| 1 | .60 (0.004) |

(b)

| religion ego | P(network ego = 1 \| religion ego) (SE) |
|---|---|
| 0 | .79 (0.053) |
| 1 | .49 (0.033) |

(c)

| religion ego | network ego | P(married ego = 1 \| religion ego, network ego) (SE) |
|---|---|---|
| 0 | 0 | .02 (0.032) |
| 0 | 1 | .31 (0.032) |
| 1 | 0 | .01 (0.016) |
| 1 | 1 | .99 (0.008) |

Table 8: Example data matrix 'two-level regression approach'

| Groupid | $Y^*_{ij}$ | $Z_{ij}$ |
|---|---|---|
| 1 | $Y_1$ | $Z_{11}$ |
| 1 | . | $Z_{21}$ |
| 1 | . | $Z_{31}$ |
| 2 | $Y_2$ | $Z_{12}$ |
| 2 | . | $Z_{22}$ |
| 2 | . | $Z_{32}$ |
| 3 | $Y_3$ | $Z_{13}$ |
| 3 | . | $Z_{23}$ |

Table 9: Example data matrix ‘persons as variables approach’

| $Y_j$ | $Z_{1j}$ | $Z_{2j}$ | $Z_{3j}$ |
|---|---|---|---|
| $Y_1$ | $Z_{11}$ | $Z_{21}$ | $Z_{31}$ |
| $Y_2$ | $Z_{12}$ | $Z_{22}$ | $Z_{32}$ |
| $Y_3$ | $Z_{13}$ | $Z_{23}$ | . |

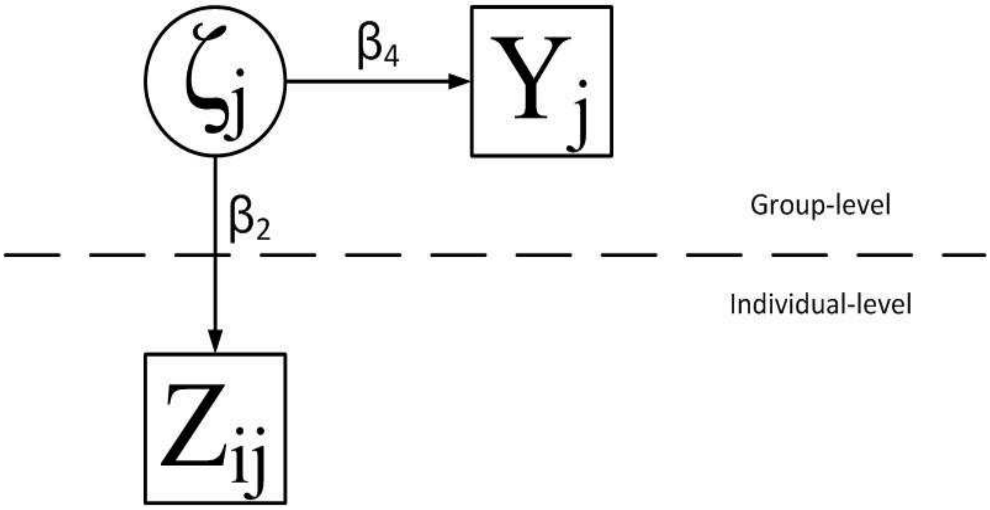

Figure 1: 1-2 Model

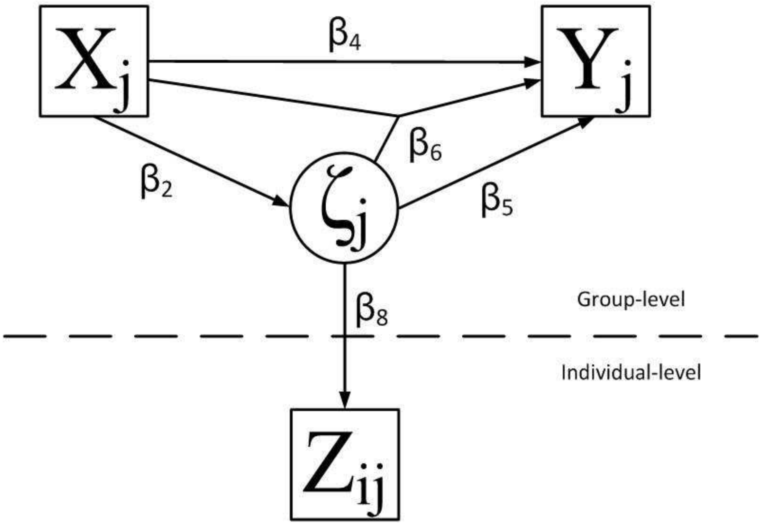

Figure 2: 2-1-2 Model

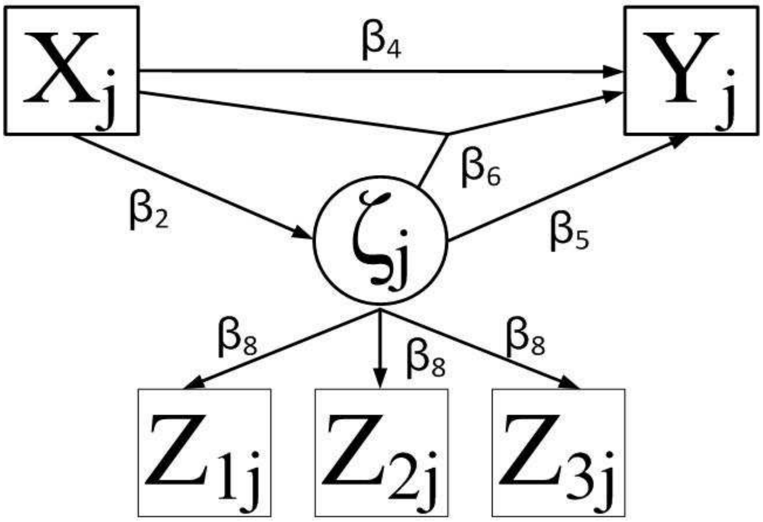

Figure 3: Persons as variables approach