Goodness-of-Fit of Multilevel Latent Class Models for Categorical Data

Erwin Nagelkerke, Daniel L. Oberski, Jeroen K. Vermunt

Tilburg University

Address for correspondence

Erwin Nagelkerke
Department of Methodology and Statistics
Tilburg University
P.O. Box 90153, 5000 LE, Tilburg
The Netherlands
Office P1.174a

Tel.: +31 13 4664030

E-mail: e.nagelkerke@tilburguniversity.edu

GOODNESS-OF-FIT OF MULTILEVEL LATENT CLASS MODELS FOR CATEGORICAL DATA

1. INTRODUCTION

Latent class (LC) analysis is mostly used to detect and develop a latent, unobserved, classification of subjects based on multiple observed categorical characteristics. The usefulness of this application in many scientific fields combined with favorable properties, such as the ability to handle multiple dependent variables and measurement error, have recently caused a growing interest in LC analysis. This in turn has resulted in the development of several extensions to the regular model in an attempt to relax assumptions and make the method more widely applicable. An important extension that has gathered quite some attention is the multilevel LC model (e.g. Muthén and Asparouhov 2011, Vermunt 2003, Vermunt 2008).

Substantively the major benefit of this multilevel extension is that it allows simultaneous classification of groups and individuals. The regular LC model may either be used to distinguish typologies of the units under study that are systematically similar (e.g. Harrell et al. 2012), or find the most common characteristics of predetermined classes (e.g. Laudy et al. 2005, Finch and Bronk 2011). The multilevel extension now makes it possible for nested categorical data in which a natural grouping is observed to also classify the groups based on the similarity of their members. For inherently nested data, such as pupils in a school, a separate classification of both the pupils as well as the schools can be obtained (e.g. Bennink et al. forthcoming, Mutz and Daniel 2012).

Additionally, the multilevel approach solves the statistical problem of dependent observations. Analogous to a multitude of statistical methods, LC analysis assumes that the units under study are independent of one another. However, this assumption does not hold when observing cases nested within a certain grouping, whether it are persons that belong to a particular group or repeated measures that belong to the same unit (Hox 2010, Snijders and Bosker 2012). An earlier solution to this dependency problem is the multiple-group approach (Clogg and Goodman 1984), but it requires all parameters to be estimated separately for all groups, causing the method to lose its value when a large number of groups is observed.

Compared to the regular LC model, the multilevel LC model thus has additional substantive applications, and offers a solution for categorical data in which there is dependency between observations. However, testing whether or not the model is correctly specified and actually captures all the dependency is currently not possible in its own right, as inspecting model fit is limited to global tests, such as the chi-square (X^2) or log-likelihood-ratio (L^2), and model comparisons through information criteria, such as the Bayesian (BIC) and Akaike Information Criterion (AIC). Although these tests and criteria can identify a well-fitting model, or the best fitting out of a series of alternative models, their global nature limits the control they provide. Especially when models become increasingly complex, the information available on the cause of better or worse fit becomes obscured. This, in turn, does not only hinder the search for possible model improvements, but also limits substantive understanding of the data.

In order to gain insight, understand the result of model adjustments, and detect specific misfit or violations of assumptions, these global criteria should ideally be supplemented with local fit statistics that single out and test one particular area of the model. In a regular LC model, such local fit measures exist in the form of the bivariate residual (BVR) (Vermunt and Magidson 2005, see also Mavridis, Moustaki and Knott 2007) and a score-test approach that leads to modification indices (Glas 1999, Oberski, Van Kollenburg and Vermunt 2013). Both test the local independence assumption that is central to the LC model and evaluate the degree to which the model captures the association between all pairs of observed variables. As such these measures indicate why one model fits better or worse, pinpoint violations of the local independence assumption, and facilitate the search for model improvements. For the multilevel LC model, however, there currently are no local fit statistics that give these insights on the group level.

Here we propose two complementary diagnostic measures that enhance exactly these abilities to detect a particular type of model misfit and increase the understanding of the fitted model for multilevel LC analysis. Both take the form of a Pearson residual and relate to the higher level of a multilevel LC model. The first residual, BVR_{group}, relates to the item distributions, and can be considered a between group measure. It can be used to evaluate the difference in responses between groups, and detect misfit that originates from the model not fitting particular groups as well as others. The second residual, BVR_{pair}, is a within group measure in the sense that it can be used to evaluate the degree of similarity amongst cases within a group, and is indicative of misfit that originates from any leftover dependency amongst the units within groups.

The remainder of this paper is structured as follows. In section 2 the multilevel LC model is introduced. In section 3 the problems with model fit statistics are discussed more elaborately, the existing BVR is discussed, and the proposed residuals are introduced. In section 4 the use of the residuals as local fit measures is demonstrated by applying them to a data example.

2. THE MULTILEVEL LC MODEL

The multilevel LC model can be expressed using two equations; one for the lower level denoting the conditional probability of all responses given by a unit, and one for the higher level marginal probability of all response patterns per group (Lukočiené, Varriale and Vermunt 2010, Vermunt 2003). The expression for the lower level is essentially that of a LC model, but in the case of a multilevel structure is made conditional on the latent class membership of the group (Vermunt 2003, Vermunt 2008).

Let the response of individual i in group j on item k be denoted as y_{ijk}, with a total of J groups, each having n_j individual members summing to N, and a total of K items, each having R_k categories. All responses to the K items of person i in group j are denoted as the vector $\boldsymbol{y}_{ij}$ with $\boldsymbol{r}$ referring to one particular answer pattern and r_k referring to a particular response to item k. The latent variable η_{ij} that classifies the units within groups has C latent classes and the latent variable ζ_j that classifies the groups has G latent classes, with c and g referring to one of these classes. Assuming conditional independence, the lower level of the multilevel LC model is expressed as:

$$\Pr(\boldsymbol{y}_{ij} = \boldsymbol{r} | \zeta_j = g) = \sum_{c=1}^{C} \Pr(\eta_{ij} = c | \zeta_j = g) \prod_{k=1}^{K} \Pr(y_{ijk} = r_k | \eta_{ij} = c, \zeta_j = g). \quad (1)$$

Removing the conditioning on the group-level latent variable (ζ_j) from Equation 1 results in the standard LC model, in which the probability of observing a particular response pattern $\boldsymbol{r}$ is a combination of the prevalence of latent class c on the latent variable η_{ij} and the probabilities of observing the combination of the responses r_k conditional on the unit's class membership. In the multilevel LC model, all these terms are made conditional on the latent class membership of the group a unit belongs to ($\zeta_j = g$), such that groups can be classified along G latent classes and the probability of an individual response pattern is affected by the group-level class membership.

The expression for the higher level of the model then denotes the marginal probability of all response patterns of individuals within group j as $\boldsymbol{y}_j$, with $\boldsymbol{s}$ denoting a particular combination of response patterns. Here an assumption of independence is required as well, but now the full response patterns of individuals rather than the responses to one item should be independent, that is,

$$\Pr(\boldsymbol{y}_j = \boldsymbol{s}) = \sum_{g=1}^{G} \Pr(\zeta_j = g) \prod_{i=1}^{n_j} \Pr(\boldsymbol{y}_{ij} = \boldsymbol{r} | \zeta_j = g). \quad (2)$$

The probability of observing the vector $\boldsymbol{y}_j$ of all individual response patterns $\boldsymbol{s}$ in group j is a combination of the prevalence, or size, of a particular group-level latent class g on the latent variable ζ_j and the probabilities of observing the combination of the individual answer patterns $\boldsymbol{r}$ conditional on the latent class membership of the group.

It should be noted that these two expressions result in a model where both the lower-level class prevalence, as well as the response probabilities can differ between all higher-level classes. Although a multitude of constraints are possible, two are most commonly used in practice, the first of which leads to the most used model that simultaneously classifies higher- and lower-level units. The first constraint $\Pr\left(y_{ijk} = r_k \middle| \eta_{ij} = c, \zeta_j = g\right) = \Pr\left(y_{ijk} = r_k \middle| \eta_{ij} = c\right)$ causes the response probabilities on the lower level to be independent of the higher-level class membership, but the class sizes to be estimated freely (Lukočiené et al. 2010, Vermunt 2003, Vermunt 2008). The second possibility is to constrain the model by setting $\Pr\left(\eta_{ij} = c \middle| \zeta_j = g\right) = \Pr(\eta_{ij} = c)$, causing the response probabilities to be estimated freely, but the lower-level class membership to be independent of higher-level class membership (Lukočiené et al. 2010, Vermunt 2004).

3. GOODNESS-OF-FIT

In this multilevel LC model, there are several key issues relating to model fit. There are the two central assumptions, namely the local independence of item responses on the lower level and the conditional independence of response patterns of individuals on the higher level, and there are the goals of correctly reproducing the item distributions or observed frequencies for both the individual observations as well as for the groups. These latter goals relate to arriving at a correct classification on both levels, and obtaining the conditional probabilities of interest depending on the substantive goal and specification of the model (Goodman 2002). Improving the fit of this model can be achieved in a multitude of ways that improve the quality of the prediction, or relax an assumption. A latent class or group-level latent class can be added, for example. Or, when keeping the same number of classes, a covariance between any combination of observed variables may be modeled, as well as any direct effect from the group-level latent variable to an observed variable.

Unfortunately, despite these different sources of misfit and the many ways to adjust the model, there is little information available as to where model misfit originates and what the effects are of model adjustments. Currently only the local independence assumption on the lower level of the model, the independence of responses conditional on the latent variable, can be inspected through the BVR. The analogous assumption on the higher level, the independence of response patterns conditional on the group-level latent variable, the quality with which the model describes the individual responses, and the degree to which the model correctly describes the groups can only be assessed jointly through global statistics. That is, the fit of the model as a whole is considered, rather than any of the individual aspects of the model.

As a result local misfit may go unnoticed, since even when a model shows adequate global fit the model may still be misspecified. In such cases, a type of local misfit averages out with other, correctly specified, areas of the model. This problem is reinforced when using information criteria, such as the Bayesian (BIC) and Akaike Information Criterion (AIC), which only compare estimated models. As long as all estimated models in such cases violate one or more assumptions, selecting the best one will still result in using a model that does not fit the data correctly. Ultimately this may lead to a wrong classification and wrong substantive conclusions.

To address this problem two local fit statistics for multilevel LC models are proposed in the following, which aim to test specific areas of the model individually. The first tests the reproduction of univariate item distributions in all the groups, and provides a partial test of how well the higher level of the model fits the data. The second is aimed at testing the conditional independence of response patterns, and in combination with the BVR allows a test of two central assumptions of the model. Both provide information on the location and extent of misfit.

3.1. Bivariate Residual

To show how the proposed statistics fit the LC framework, and for the sake of completeness, the existing BVR is briefly introduced. Vermunt and Magidson (2013) construct the BVR to test the assumption of local independence for all pairs of observed variables in a regular LC model, but it can be applied identically to the lower level of a multilevel model. The BVR assesses the difference between the observed frequencies ($n_{rr'}$) and the model expected frequencies ($m_{rr'}$) in the two-way cross-tabulation of items k and k' by a Pearson statistic divided by its number of degrees of freedom (see also Bartholomew and Leung 2002, Vasdekis, Cagnong and Moustaki 2012), that is,

$$BVR_{kk'} = \frac{1}{(R_k - 1)(R_{k'} - 1)} \sum_{r=1}^{R_k} \sum_{r'=1}^{R_{k'}} \frac{(n_{rr'} - m_{rr'})^2}{m_{rr'}}. \tag{3}$$

The expected frequencies follow from the LC model, which assumes conditional independence of item responses given latent class membership. More specifically, they are obtained by multiplying the class-specific probabilities of the response r on item k and response r' on item k', and summing these over the latent classes using the class membership probabilities as weight. For a LC model without a multilevel structure:

$$m_{rr'} = \sum_{i=1}^{N} \sum_{c=1}^{C} \Pr(y_{ik} = r_k | \eta_i = c) \Pr(y_{ik'} = r_{k'} | \eta_i = c) \Pr(\eta_i = c | \boldsymbol{y}_i = r). \qquad (4)$$

When no values are missing the same $m_{rr'}$ can be obtained by using $\Pr(\eta_i = c)$ as weight instead of $\Pr(\eta_i = c| \boldsymbol{y}_i = \boldsymbol{r})$, and multiplying by N rather than summing over N, since $\Pr(\eta_i = c)$ equals the average $\Pr(\eta_i = c| \boldsymbol{y}_i = \boldsymbol{r})$ for the complete sample. However, in the case of missing values the observed frequencies only contain the cases for which both variables are observed. To obtain the corresponding expected frequencies, the class membership probabilities should be based on this subsample. That is, using $\Pr(\eta_i = c)$ is not appropriate, and the frequency should be obtained by summing over the cases with both variables observed, using $\Pr(\eta_i = c| \boldsymbol{y}_i = \boldsymbol{r})$ as weight.

The above formulation for $m_{rr'}$ can easily be generalized to be applicable in a multilevel LC analysis. The sum over latent classes must then contain the joint posterior probability of the lower and higher level latent variables and the sum over individuals must be over both groups and individuals within groups:

$$m_{rr'} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{g=1}^{G} \sum_{c=1}^{C} \Pr(y_{ijk} = r_k \big| \eta_{ij} = c, \zeta_j = g \big) \tag{5}$$

$$\Pr(y_{ijk'} = r_{k'} \big| \eta_{ij} = c, \zeta_j = g \big) \Pr(\eta_{ij} = c, \zeta_j = g | \, \boldsymbol{y}_j = s).$$

Any deviation between the observed and the predicted frequency, which assumes local independence of items given latent class membership, is now contained in the residual.

3.2. Group-Variable Residual

In order to further deconstruct global misfit, we here propose a group-variable residual ($\mathrm{BVR}_{\mathrm{group}}$). As can be seen from Equation 2, the response vector $\boldsymbol{y}_j$ containing all individual response patterns is a function of the size of the group-level class and the individual answer patterns. This implies that, among others, the univariate response frequencies within each group should be modeled correctly for the latent class solution to be correct. Because the observed group membership can be understood as a nominal covariate in a multilevel LC model, the BVR can be adapted to assess the response to a nominal dependent variable and group membership:

$$BVR_{group.k} = \frac{1}{(R_k - 1)(J - 1)} \sum_{j=1}^{J} \sum_{r=1}^{R_k} \frac{\big(n_{jr} - m_{jr}\big)^2}{m_{jr}} \tag{6}$$

The observed frequency here is simply the number of units in group j with response r_k. The expected frequencies m_{jr} can be obtained from the individual probabilities $\Pr\big(y_{ijk} = r_k\big)$:

$$\Pr\left(y_{ijk} = r_k\right) = \sum_{g=1}^{G} \Pr\left(y_{ijk} = r_k \middle| \zeta_j = g\right) \Pr\left(\zeta_j = g \middle| \boldsymbol{y}_j = s\right), \tag{7}$$

where

$$\Pr\left(y_{ijk} = r_k \middle| \zeta_j = g\right) = \sum_{c=1}^{C} \Pr\left(y_{ijk} = r_k \middle| \eta_{ij} = c, \zeta_j = g\right) \Pr\left(\eta_{ij} = c \middle| \zeta_j = g\right). \tag{8}$$

Then

$$m_{jr} = \sum_{i=1}^{n_j} \Pr(y_{ijk} = r_k) = \sum_{i=1}^{n_j} \sum_{g=1}^{G} \Pr\left(y_{ijk} = r_k \middle| \zeta_j = g\right) \Pr\left(\zeta_j = g \middle| \boldsymbol{y}_j = s\right). \tag{9}$$

Thus, the probability of a particular response is summed over all group members to obtain its frequency within the group, and is itself a function of the group-class response probabilities and the group-class membership probabilities. It should be noted that for the class-membership on the group-level the posterior probability $\Pr\left(\zeta_j = g \middle| \boldsymbol{y}_j = s\right)$ is used . Because the interest lies in testing the group by variable relationships and aggregating these over the groups, all available information on the groups should be used, as contained in the posterior.

The statistic itself is computed for all groups separately, and summed over the groups to test the assumption of correct model fit in each of the groups. This sum is additionally divided by $(R_k - 1)(J - 1)$. The $\text{BVR}_{\text{group}}$ now equals the average contribution to the residual per degree of freedom. That is, the dimension of the matrix to which Equation 6 is applied is $R_k \times J$, resulting in $(R_k - 1)(J - 1)$ non-redundant parameters. Correcting for both R_k and J standardizes the $\text{BVR}_{\text{group}}$ such that it is not affected by the number of groups, nor the number of categories on the variable.

As can be seen from Equations 7 through 9, a special case exists when the nested structure of the data is ignored by estimating the multilevel LC model with only one group-level class. The results are identical to omitting the group-level latent variable altogether, and assures the $\mathrm{BVR}_{\mathrm{group}}$ is independent from the number of lower-level classes in order to obtain its baseline value, which is substantively indicative of the between-group heterogeneity, or the between-group variance. For this model, the residual is then broadly comparable to the empirical Bayes estimates as used in linear multilevel models. Although their common use is testing the normality assumption on the higher level, they can also be used to construct influence diagnostics (Snijders and Berkhof 2008) and as such are indicative of misfit.

3.3. Paired-Case Residual

In a multilevel LC model, the higher level has a local independence assumption similar to that of the lower level. Where the assumption in Equation 1 is that the responses r_k are independent for all the K items per individual, in Equation 2 the response patterns $\boldsymbol{r}$ are assumed independent for all the n_j individuals per group. However, to capture this dependency amongst units within a group the responses of the individual members should be related to one another. This cannot be done as straightforwardly as is the case for the dependency between item pairs. Where the response frequencies for the latter can be cross tabulated directly, the cross tabulation of dependency amongst units requires all units within a group to be related. An intuitive approach to do so is to create all pairs of units within every group and obtain the pairwise response frequencies. The expected and observed response frequencies can then again be compared:

$$BVR_{pair} = \frac{J}{N}\frac{1}{R_k(R_k-1)/2}\left[\sum_{r}^{R_k}\sum_{r'>r}^{R_{k'}}\frac{((n_{krr'}+n_{kk'r})-(m_{krr'}+m_{kr'r}))^2}{m_{krr'}+m_{kr'r}} + \sum_{r}\frac{(n_{krr}-m_{krr})^2}{m_{krr}}\right]. \tag{10}$$

To illustrate, consider a group containing five observations, with responses to one of multiple variables as in Table 1. The residual can be understood as considering the combined responses r and r' of cases i and i' to item k as one element. To obtain the observed frequencies a square contingency table of which the order is equal to the number of categories on the variable of interest can then be made per pair. The cell that identifies the actual answer pattern of that pair of cases has a frequency of one and all else equals zero.

Table 1. Illustration of obtaining the observed pairwise response frequencies

Data

Obs	Var	Group
A	0	1
B	0	1
C	1	1
D	0	1
E	1	1
F	0	2
G	1	2
H	…	…

		B	
		0	1
A	0	1	0
	1	0	0

		C	
		0	1
A	0	0	1
	1	0	0

		D	
		0	1
A	0	1	0
	1	0	0

		E	
		0	1
A	0	0	1
	1	0	0

		C	
		0	1
B	0	0	1
	1	0	0

		D	
		0	1
B	0	1	0
	1	0	0

		E	
		0	1
B	0	0	1
	1	0	0

		D	
		0	1
C	0	0	0
	1	1	0

		E	
		0	1
C	0	0	0
	1	0	1

		E	
		0	1
D	0	0	1
	1	0	0

The corresponding predicted probability of a certain pair of responses follows from the combined probability of person i giving response r and person i' giving response r' conditional on the group-level class:

$$\Pr(y_{ijk} = r_k, y_{i'jk} = r'_k) = \sum_{g=1}^{G} \Pr(y_{ijk} = r_k | \zeta_j = g) \Pr(y_{i'jk} = r'_k | \zeta_j = g) \Pr(\zeta_j = g | \mathbf{y}_j = s), \quad (11)$$

where $\Pr(y_{ijk} = r_k | \zeta_j = g)$ can be obtained by Equation 8. Because these probabilities are only conditional on the group-level latent variable in a model without covariates, they are identical for identical patterns, and the order of the responses is interchangeable. That is, within a group only the probabilities on the diagonal and either the upper or lower off-diagonal need to be obtained. Aggregating these probabilities to arrive at the expected frequencies then can be done by multiplying the probability of a pair with the number of pairs $n_j(n_j - 1)/2$:

$$m_{krr'} = \sum_{j=1}^{J} (n_j(n_j - 1)/2) \Pr(y_{ijk} = r_{k'} y_{i'jk} = r'_k) \quad (12)$$

Again, as is done for the BVR_{group}, the posterior probability is used in Equation 11 to obtain this estimated frequency. In this case the main reason is that this weighting is more appropriate in cases where groups are of different size, and thus containing different numbers of pairs per group. As can be seen from Equations 10 and 12, in comparison to Equations 6 and 9, the BVR_{pair} is not obtained for each group separately and only subsequently summed over the groups, but the aggregation already occurs when computing the expected frequencies. By weighting by the posterior probability

$\Pr(\zeta_j = g|\boldsymbol{y}_j = s)$ the expected frequencies account in the best manner for unequal group sizes. With equal group sizes, using posterior or unconditional class membership probabilities will give the same expected frequencies.

The observed frequency of pairs can now be obtained by summing the pairwise tables from Table 1, as is done in Table 2. The probability of a pair follows from Equation 11, and the expected frequency from Equation 12. For the illustration, the probabilities from the first model in the application section are used.

Here, the structure of Equation 10 also becomes clear. Note that because the order of the observations within a group is arbitrary, observing a 0-1 pair is in fact the same as observing a 1-0 pair. This is why the symmetric off-diagonal elements of the table are combined in the first summation in Equation 10. The latter part of Equation 10 adds the discrepancy between the observed and expected frequencies on the diagonal.

Table 2. Illustration of obtaining the pairwise residual contribution per answer pattern

		Observed				Probability				Expected				Residual Contr.	
		i'				r'				i'				i'	
		0	1			0	1			0	1			0	1
i	0	3	5	r	0	.415	.225	i	0	4.152	2.249	i	0	0.320	0.056
	1	1	1		1	.225	.135		1	2.249	1.351		1	-	0.091

$$BVR_{pair} = \frac{1}{6}\frac{1}{2(2-1)/2}(0.320 + 0.056 + 0.091) = 0.079$$

To finally arrive at the $\mathrm{BVR}_{\mathrm{pair}}$ the resulting residual is divided in such a way that the statistic equals the contribution to the residual per degree of freedom, in this case $R_k(R_k - 1)/2$ given the symmetry on the off-diagonals. Additionally the raw residual is divided by the average group size to avoid extremely large values, which are likely to occur since the theoretical maximum value of the statistic increases as a triangular sequence with n_j.

Unfortunately, the univariate marginal values for the resulting tables are not reproduced correctly when groups differ in size, in which case $(n_{krr'} + n_{kr'r}) \neq (m_{krr'} + m_{kr'r})$, which is also the case in the illustration. The cause is simply that an observation in a larger group is in more pairs than an observation in a smaller group. Differences between the observed (n) and expected frequency (m) would then not only reflect the degree to which the model captures dependency between cases, but the residual would also partly reflect the difference in the univariate distribution. This changes the interpretation of the BVR_{pair}, and unnecessarily so because the univariate distributions are always correctly reproduced by the model.

Therefore a number of iterative proportional fitting (IPF) cycles is used to equate the reproduced and observed marginal frequencies and reduce the BVR_{pair} to zero when there is no residual dependency. The pairwise contingency table is made symmetrical first, such that answer patterns that only differ in respect to the order of the responses have the same frequency. As mentioned, the probability and thus the expected frequency of a certain pair of responses is identical regardless of order, but this is not necessarily the way in which they are observed.

In the IPF procedure the cells in the expected frequency table are adjusted so that its marginals match the observed marginals. The subsequent iterations alternate between row and column adjustments where each cell is multiplied by the ratio between the observed and the expected row (column) marginal. This process converges to a table with marginals equal to the observed marginal frequencies whilst retaining the cross-product ratios within the table (Bishop, Fienberg and Holland 1975).

Table 3. Illustration of Iterative Proportional Fitting cycles

Observed

		i' = 0	*i'* = 1	
i	0	3	(5+1)/2	6
	1	(1+5)/2	1	4
		6	4	10

Expected

		i' = 0	*i'* = 1	
i	0	4.152	2.249	6.401
	1	2.249	1.351	3.599
		6.401	3.599	10

IPF Cycle 1 – Row[a]

		i' = 0	*i'* = 1	
i	0	3.892	2.108	*6*
	1	2.499	1.501	*4*
		6.391	3.609	-

IPF Cycle 1 – Column[b]

		i' = 0	*i'* = 1	
i	0	3.654	2.336	5.990
	1	2.346	1.664	4.010
		6	*4*	-

IPF Cycle 2 - Row

		i' = 0	*i'* = 1	
i	0	3.660	2.340	*6*
	1	2.340	1.660	*4*
		6.000	3.999	-

IPF Cycle 2 - Column

		i' = 0	*i'* = 1	
i	0	3.660	2.340	6
	1	2.340	1.660	4
		6	*4*	10

a. Row operation: Multiply cell with the ratio between the observed and expected row marginal; cell(observed row / expected row)
b. Column operation: Multiply cell with the ratio between the observed and expected column marginal; cell(observed column / expected column)

In the IPF procedure the cells in the expected frequency table are adjusted so that its marginals match the observed marginals. The subsequent iterations alternate between row and column adjustments where each cell is multiplied by the ratio between the observed and the expected row (column) marginal. This process converges to a table with marginals equal to the observed marginal frequencies whilst retaining the cross-product ratios within the table (Bishop, Fienberg and Holland 1975).

The resulting BVR_{pair} statistic reduces to zero when the model captures all the dependency amongst cases within a group. Identical to the BVR_{group} its baseline value can be obtained by estimating the model where the nesting of the data is ignored by modeling only one group-level class. The statistic is broadly comparable to the residual intra-class correlation (ICC) in mixed models, which is the degree of dependency that is not captured by the model when controlling for the independent variables (Snijders and Bosker 2012). The BVR_{pair} is similarly related to the uncaptured dependency, and indicative of the homogeneity within groups that is ignored when the nested structure of the data is not, or only partially reproduced.

3.4. Bootstrap

The BVR, BVR_{group} and BVR_{pair} residuals are all obtained identically to Pearson residuals. However, for the BVR it is known that it does not follow the chi-square distribution, and the same is expected to be true for the two proposed measures. To still obtain p-values for the residuals, a parametric bootstrap can be used (Langeheine, Pannekoek and Van de Pol 1996), which is known to work for the BVR (Oberski et al. 2013). Based on the maximum likelihood estimate, the bootstrap in this instance samples group-class membership, class membership conditional on group-class membership, and the responses conditional on the membership of both. This results in alternative data sets with the same structure as the original to which the model is fitted. For each of these refitted models the BVR values are obtained. The estimated p-value then is the proportion of replicated models in which the BVR residuals are larger than in the original model

(Vermunt and Magidson 2013). As such the BVR_{group} and BVR_{pair} are not compared to an asymptotic distribution, but rather to an empirical distribution constructed by simulation. The bootstrap p values can be used for hypothesis testing, that is, for determining whether or not potential assumption violations are statistically significant.

4. APPLICATION

To demonstrate their use and usefulness, the BVR_{group} and BVR_{pair} will be applied to a real data example. The data in this application were collected to investigate the relationship between task variety and the psychological well-being of employees. Since both team-level and individual task characteristics are expected to be of importance, Vermunt (2003) aggregates the data by classifying both teams and individuals through multilevel LC analysis. In a later stage these classes can then be related to an outcome such as well-being. For instance, to detect class characteristics that affect the outcome for better or worse (e.g. Mutz, Bornmann and Daniel 2013, Shim and Finch 2014).

However, when the LC model is incorrectly specified or violates assumptions, not only is there a possibility of wrongly classifying the teams and employees, the relationship between an outcome and the classification is similarly unsound. This first step of classification then clearly is an important one, since a wrong classification may result in wrong substantive conclusions on the actual goal of the study. Here the classification will be re-examined using the proposed BVR_{group} and BVR_{pair} statistics to demonstrate their use. After excluding all cases with missing values and two teams with only one member, the data contains 848 cases in 86 teams, and is similar to that used by

Vermunt (2003, Vermunt 2005), as collected by Van Mierlo (2003). For all employees the perception of task variety in their job was measured with five categorical items of which the four categories are collapsed to make them dichotomous. The variable measuring task repetitiveness is coded inversely to the other variables such that a higher score reflects lower repetition and all scores are substantively in accordance. All models are estimated in LatentGOLD 5.0. The survey wording and LatentGOLD syntax are both provided in the appendix. The data set itself is included in LatentGOLD as example data.

Since the BIC is currently the main criterion for model selection, selecting the best fitting from a series of alternative models, Table 4 depicts the BIC values for 29 models with differing numbers of classes. All of these models assume conditional independence between the five items, contain one latent variable on both levels (η and ζ), and only allow an indirect effect of the group-level latent variable ζ on the items through the lower level latent variable η (see also Vermunt 2003). It should be noted that these BIC values are computed using the number of groups as the sample size, rather than the number of cases, as this is found to be the more appropriate sample size to determine the number of classes in multilevel LC models (Lukočiené et al. 2010, Lukočienė and Vermunt 2010).

Based on these values the model with two group-level and three low-level classes would be the best fitting, resulting in the profile as depicted in Table 5. On the lower level, the largest of the three classes is one where people report high levels of task variation and creativity. The second class is one in which people report having repetitive, uncreative, and unvaried tasks. The third is a class with highly creative tasks, yet quite unvaried and repetitive. On the group level, the classes are less distinguished in their overall profile. Members of teams in the first group-level class are most likely to belong to the first individual-level class, and of the second higher level class to the second lower-level class. Overall then the team profile of the first group-level class mostly is that of diverse, varied, and challenging tasks, whereas the second class have more repetitive tasks that allow less creativity.

Table 4. BIC Values for 29 models assuming local independence of items and indirect effects of the group-level latent variable[a] (N equals number of groups)

		Lower-level Classes				
		2	**3**	**4**	**5**	**6**
Group-level Classes	**1**	4820	4818	4837	4861	.[b]
	2	4786	***4785***	4799	4819	4844
	3	4794	4795	4794	4814	4836
	4	4802	4806	4809	4826	4850
	5	4811	4818	4822	4839	4865
	6	4820	4831	4838	4857	4880

a. Constraint: $\Pr(y_{ijk} = r_k | \eta_{ij} = c, \zeta_j = g) = \Pr(y_{ijk} = r_k | \eta_{ij} = c)$
b. Unidentified

However, the two problems laid out in section three would arise when this model would be accepted solely based on the BIC value. Firstly, the BIC identifies the best alternative out of the models presented, but it does not guarantee that no assumptions are violated, that is, that the model picks up all relevant aspects in the data. If this is not the case, the classification described in Table 5 could be faulty, and any further analysis to relate this classification to outcomes may also be affected negatively. Secondly, many alternative models can be specified, other than those with differing numbers of classes.

Table 5. Latent class profile for the 3-class, 2-group-class model. BIC=4785.3

	GClass 1 Diverse	GClass 2 Uniform	Class 1 Diverse	Class 2 Structure	Class 3 Creative	Overall
Non-Repetitive	.428	.279	.515	.125	.225	.385
Creative	.631	.382	.707	.065	.914	.558
Diverse	.792	.480	.961	.146	.483	.700
Capacity	.730	.578	.837	.439	.350	.685
Variation	.754	.461	.964	.192	.000[a]	.668
Class 1	.752	.371				
Class 2	.150	.537				
Class 3	.098	.092				
Prevalence	.707	.293	.640	.263	.097	

a. Boundary solution

In all the estimated models, conditional independence of the observed items is assumed, which can be relaxed by allowing one or more covariances between the observed variables. Furthermore, the effect of the group-level latent class on the observed variables is assumed to be fully mediated by lower-level class membership. This too can be relaxed by allowing direct effects from the higher-level latent variable on any of the items. The prohibitive difficulty of improving the model through trial and error, or even considering the option of estimating all possible models, now quickly becomes clear. When keeping the number of classes constant, there are 1024 different combinations of allowable covariances, and for each of these combinations another 32 possible combinations of direct effects. If the possibility of equating certain parameters to one another is also considered, this model can be adjusted in 17 factorial different ways.

To illustrate how the local fit measures may largely resolve the problem of identifying misfit without the need to estimate many additional models, the residual measures for the model with the lowest BIC are presented in Table 6 with bootstrapped p-values for all BVR measures in parentheses. The regular BVR indicates that the variable measuring the diversity of a person's job shows some residual dependency with the variable measuring job variation, which substantively should come as no surprise. On the higher level, the BVR_{group} and BVR_{pair} also show assumption violations, whereby the repetitive and creative variables both show dependency between cases that is not captured by the model, as well as an incorrectly reproduced item distribution between the groups. So, even though it is the best alternative out of thirty models, the three individual-level, two group-level class model violates the three tested assumptions to some extent.

Table 6. BVR, BVR_{group} & BVR_{pair} residuals for the 3-class, 2-group-class model. BIC=4785.3

	Non-Rep.	Creative	Diverse	Capacities	Variation
Creative	0.763 (.242)				
Diverse	0.248 (.282)	0.028 (.442)			
Capacities	0.183 (.570)	0.359 (.308)	0.504 (.106)		
Variation	0.010 (.706)	0.036 (.272)	0.153 (.016)	0.011 (.790)	
BVR-group	1.586 (.000)	1.051 (.058)	0.788 (.164)	1.072 (.132)	0.816 (.316)
BVR-pair	1.740 (.000)	0.570 (.028)	0.123 (.296)	0.366 (.098)	0.000 (.974)

From Table 4 it can be concluded that improving this model is not achieved by increasing the number of classes. Inspecting the BVR measures for these models leads to the same conclusion, as a combination of problems on both levels of the model persists when increasing either the number of classes on the lower level, the higher level, or both.

Thus, to improve this model, a solution other than increasing the number of classes is required. Starting model improvements on the lower level of the model is often the most fruitful, as it is more likely that group-level dependency is introduced by having a wrong specification on the lower level than the reverse (Lukočiené et al. 2010). This is due to the higher-level classification being partly determined by the classes on the lower level, as can be seen in Equation 2.

Substantively, the significant dependency between the self-reported variation and diversity of work is sensible and including a covariance between these two variables seems justified. As shown in Table 7, adding this covariance removes any problematic bivariate dependency on the lower level of the model.

Table 7. Residuals for the 3-class, 2-group-class model, covariance between Variation and Diverse. BIC = 4783.2

	Non-Rep.	Creative	Diverse	Capacities	Variation
Creative	0.101 (.642)				
Diverse	0.602 (.104)	0.022 (.514)			
Capacities	0.871 (.184)	0.001 (.938)	0.178 (.264)		
Variation	0.062 (.400)	0.042 (.316)	0.000 (1.00)	0.028 (.670)	
BVR-group	1.576 (.000)	0.973 (.140)	0.776 (.264)	1.037 (.194)	0.842 (.312)
BVR-pair	1.523 (.000)	0.294 (.130)	0.128 (.296)	0.256 (.138)	0.011 (.780)

Considering the BVR_{group} and BVR_{pair} statistics, the logical next step is to add a direct effect from the group-level latent variable on the repetitive variable. Such a direct effect is the most parsimonious solution in an attempt to capture more dependency and improve within-group model fit regarding the repetitive variable, adding only one parameter. Substantively too there is evidence that the differences in repetitive work between teams reflect on that of the individual tasks (Van Mierlo 2003).

After adding this effect problems arise in all five variables as depicted in Table 8, causing the model to no longer describe the within-team item distributions correctly, nor does it adequately capture the dependency between cases. Yet, despite the large shift on the group-level of the model, the lower level does not show any problems. The interpretations of the individual level classes (not reported) also do not change, indicating that the problems are largely the result of a failure to capture team differences correctly. Given that there are problems with all five variables on the group level of the model, adding an additional group-level class is the best option here.

Table 8. Residuals for the 3-class, 2-group-class model, covariance between Variation and Diverse and direct effect from group-level latent variable on Repetitive. BIC = 4777.1

	Non-Rep.	Creative	Diverse	Capacities	Variation
Creative	0.004 (.922)				
Diverse	0.737 (.082)	0.068 (.204)			
Capacities	0.962 (.180)	0.026 (.732)	0.046 (.670)		
Variation	0.019 (.664)	0.034 (.212)	0.000 (1.00)	0.090 (.432)	
BVR-group	1.544 (.000)	1.405 (.000)	1.356 (.000)	1.194 (.040)	1.125 (.048)
BVR-pair	1.657 (.000)	0.930 (.006)	1.325 (.002)	0.458 (.048)	0.280 (.070)

Adding a third group-level class indeed solves most problems on the higher level of the model, as can be seen from Table 9. In this model the covariance between the variation and diverse variable, as well as the direct effect on the repetitive variable are retained. As a final adaptation a direct effect from the group-level latent variable on the creative variable is added, following the BVR_{group} value, and the reasoning that the structure of a team and the overall packet of tasks it realizes may have a direct effect on the creativity an employee has in accomplishing their share of the teamwork.

Table 9. Residuals for the 3-class, 3-group-class model, with covariance between Variation and Diverse and direct effect from group-level latent variable on Repetitive. BIC = 4768.9

	Non-Rep.	Creative	Diverse	Capacities	Variation
Creative	0.073 (.720)				
Diverse	0.315 (.214)	0.054 (.362)			
Capacities	0.620 (.274)	0.170 (.536)	0.003 (.880)		
Variation	0.046 (.378)	0.114 (.154)	0.000 (1.00)	0.053 (.546)	
BVR-group	1.041 (.046)	1.185 (.012)	0.843 (.316)	1.150 (.054)	0.931 (.290)
BVR-pair	0.138 (.214)	0.589 (.020)	0.051 (.496)	0.326 (.118)	0.092 (.454)

In Table 10 the BVR, BVR_{group}, and BVR_{pair} residuals for the final model are presented. Further attempts to make this model more parsimonious, result in models where significant residuals are reintroduced.

Table 10. Residuals for the three class, three group-class model, with covariance between Variation and Diverse and direct effects from group-level latent variable on Repetitive and Creative. BIC = 4775.3

	Non-Rep.	Creative	Diverse	Capacities	Variation
Creative	0.001 (.950)				
Diverse	0.530 (.108)	0.085 (.192)			
Capacities	0.837 (.186)	0.005 (.858)	0.090 (.454)		
Variation	0.003 (.890)	0.048 (.238)	0.000 (1.00)	0.023 (.716)	
BVR-group	0.771 (.260)	0.739 (.452)	0.927 (.112)	1.083 (.136)	0.914 (.216)
BVR-pair	0.016 (.628)	0.011 (.696)	0.202 (.174)	0.280 (.150)	0.014 (.728)

The profile of this final model is presented in Table 11. Comparing these results to those in Table 5, it becomes clear that the individual-level classification is practically identical to that obtained in the model with two group-level latent classes, and three individual-level latent classes. On the group-level the additions to the model, an extra latent class and two direct effects, led to splitting up the large first class from the initial solution. The second group-level class in this model is similar to the second class in the model presented in Table 5. The first class from Table 5, however, is split up into two classes. These two classes are rather similar when compared to each other, as well as when compared to the class from the first model, but with a large difference in degree of task repetition reported by the team members.

Table 11. Profile for the three class, three group-class model, with covariance between Variation and Diverse and direct effects on Repetitive and Creative. BIC = 4775.3

	GClass 1 Repetitive	GClass 2 Defined	GClass 3 Non-Rep.	Class 1 Diverse	Class 2 Structure	Class 3 Creative	Overall
Non-Rep.	.301	.316	.613	.554	.130	.233	.400
Creative	.660	.348	.674	.731	.077	.844	.557
Diverse	.822	.521	.754	.953	.209	.526	.698
Capacity	.753	.590	.707	.851	.444	.342	.683
Variation	.786	.506	.704	.962	.263	.000[a]	.665
Class 1	.784	.529	.678				
Class 2	.122	.382	.195				
Class 3	.095	.090	.127				
Prevalence	.352	.345	.302	.613	.284	.103	

a. Boundary solution.

The results from Table 11 clearly show the difficulty in capturing team differences using team-level classes, as the first and third class only differ with respect to the degree of task repetition. Given that the group-level classes in the initial model are only affecting the indicators indirectly through the lower-level latent class, such a relatively small difference between teams may become obscured between other characteristics that the teams do have in common. That is, detecting these specific characteristic on the team-level in a model without direct effects from the team-level latent variable also requires more classes on the lower level. Such an addition of latent classes on either level is not warranted when inspecting the BIC values for these models, which are known to favor model parsimony. However, through the proposed BVR_{group} and BVR_{pair} this lack of a direct effect between the group-level latent class and the repetitiveness variable could be detected, as well as the subsequent need for an additional class on the group level.

Maybe more importantly, due to the improved fit and the possibility to test assumptions, the model arrives at different substantive results. In this instance, the added group-level class causes a separation primarily based on task repetitiveness. Given that the interest lies on relating the classes to well-being as an outcome, the results may differ between the original model as depicted in Table 5, and the better fitting model arrived at in Table 11. When, for example, task repetitiveness on the team level is detrimental to employee well-being it would have been hard to distinguish as an important factor in the model with two group-level classes. It would however be visible in the model with three group-level classes where a comparison between group-level classes one and three would identify repetitiveness as an important factor.

Using the residuals as additional guidance now results in a model with substantial better fit that would likely not have been found when only relying on the BIC or comparable criteria. Both the proposed BVR_{group} and BVR_{pair}, in combination with the BVR, allow detection of the initial assumption violations, and identify not only which part of the model, but also which specific parameters may prove problematic. Misfit can be pinpointed and tested, allowing for far more informed and directed model adjustments, which may lead to different, more thoroughly tested, substantive results.

5. DISCUSSION

Several problems exist when solely using global fit statistics or information criteria for model selection in multilevel LC analysis. Due to the lack of local fit statistics potential model misfit may go unnoticed, and there is no information available as to how a model might be adjusted and improved. Therefore two new local fit statistics, the BVR_{group} and BVR_{pair}, are proposed, which test individual areas of the model and as such help in determining which areas of the model are problematic and how a model can best be improved. In conjunction with the standard BVR, they also allow the two local independence assumptions central to multilevel LC models to be inspected and tested. Computation of both the BVR_{group} and BVR_{pair} is already implemented in the user-friendly LatentGOLD 5.0 software package.

By using the BVR_{group} and BVR_{pair} as additional guidance to test and improve a multilevel LC model, it is shown that they enhance the ease with which fruitful model adjustments can be found. The model obtained by relying on the two residuals has better global fit and is known to better adhere to the local independence assumptions. The usefulness of the residuals is further emphasized by the change in substantive results between the initial model selected through the BIC, and the latter model as improved through the use of the proposed statistics. That is, the misfit that is detected in this instance is not a mere misspecification against which the model is robust, but actually distorts model based conclusions.

In this case then the important sources of misfit that affect the results have been picked up by the two residuals. Still, this paper serves as an introduction and a more indepth simulation study is lacking. Such a future study would not so much focus on the type I error rates, as the process of p-value bootstrapping is identical to that of the BVR for which it has been extensively tested (Oberski et al. 2013). Rather, it would focus on the consistency with which misspecification is detected under different circumstances and in more complex models, such as models incorporating covariates.

An additional extension that does require future work is to develop a similar residual for LC models for longitudinal data, where dependencies can be assumed to take on the form for autocorrelation structures. Furthermore the use of the BVR_{group} and BVR_{pair} may be studied for different methods and models that could also benefit from these statistics (see e.g. Varriale and Vermunt 2012). Where they are originally aimed at testing the local fit of multilevel LC models, they can be applied to all cases where categorical multilevel data is used, since the observed frequencies would be identical, and only the expected frequencies would need to be obtained from the alternative approach.

REFERENCES

Bartholomew, David, J., and Shing On Leung. 2002. “A goodness of fit test for sparse 2^p contingency tables.” *British Journal of Mathematical and Statistical Psychology* 55(1):1-15. doi: 10.1348/000711002159617.

Bennink, Margot, Marcel A. Croon, Jos Keuning, and Jeroen K. Vermunt. Forthcoming. “Measuring Student Ability, Classifying Schools, and Detecting Item-Bias at School-Level Based on Student-Level Dichotomous Attainment Items.” *Journal of Educational and Behavioral Statistics.*

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Clogg, Clifford C., and Leo A. Goodman. 1984. “Latent Structure Analysis of a Set of Multidimensional Contingency Tables.” *Journal of the American Statistical Association* 79(388):762-71. doi: 10.2307/2288706.

Finch, W. Holmes, and Kendall C. Bronk. 2011. “Conducting Confirmatory Latent Class Analysis Using Mplus.” *Structural Equation Modeling* 18(1):132-151. doi: 10.1080/10705511.2011.532732.

Glas, Cees A. W. 1999. “Modification Indices for the 2-Pl and the Nominal Response Model.” *Psychometrika* 64(3):273-94. doi: 10.1007/BF02294296.

Goodman, Leo A. 2002. "Latent Class Analysis: The Empirical Study of Latent Types, Latent Variables, and Latent Structures." Pp. 3-56 in *Applied Latent Class Analysis*, edited by J. A. Hagenaars and A. L. McCutcheon. Cambridge, UK: Cambridge University Press.

Harrell, Paul T., Brent E. Mancha, Hanno Petras, Rebecca C. Trenz and, William W. Latimer. 2012. “Latent Classes of Heroin and Cocaine Users Predict Unique HIV/HCV Risk Factors.” *Drug and Alcohol Dependence* 122(3):220-27. doi: 10.1016/j.drugalcdep.2011.10.001.

Hox, Joop J. 2010. *Multilevel Analysis: Techniques and Applications*. New York, NY: Routledge.

Langeheine, Rolf, Jeroen Pannekoek, and Frank van de Pol. 1996. "Bootstrapping Goodness-of-Fit Measures in Categorical Data Analysis." *Sociological Methods & Research* 24(4):492-516. doi: 10.1177/0049124196024004004.

Laudy, Olav, Mark Zoccolillo, Raymond H. Baillargeon, Jan Boom, Richard E. Tremblay, and Herbert J. A. Hoijtink. 2005. “Applications of Confirmatory Latent Class Analysis in Developmental Psychology.” *European Journal of Developmental Psychology* 2(1):1-15. doi: 10.1080/17405620444000193.

Lukočiené, Olga, Roberta Varriale, and Jeroen K. Vermunt. 2010. "The Simultaneous Decision(s) About the Number of Lower- and Higher-Level Classes in Multilevel Latent Class Analysis." *Sociological Methodology* 40(1):247-83. doi: 10.1111/j.1467-9531.2010.01231.x.

Lukočienė, Olga, and Jeroen K. Vermunt. 2010. "Determining the Number of Components in Mixture Models for Hierarchical Data." Pp. 241-49 in *Advances in Data Analysis, Data Handling and Business Intelligence*, edited by A. Fink, B. Lausen, W. Seidel and A. Ultsch. Heidelberg, DE: Springer. doi: 10.1007/978-3-642-01044-6.

Mavridis, Dimitris, Irini Moustaki, and Martin Knott. 2007. “Goodness-of-Fit Measures for Latent Variable Models for Binary Data.” Pp. 135-161 in *Handbook of Latent Variable and Related Models*, edited by S. Lee. Amsterdam, NL: Elsevier. doi: 10.1016/S1871-0301(06)01007-9.

Mierlo, van, Heleen. 2003. *Self-Managing Teamwork and Psychological Well-Being*. PhD dissertation, Department of Technology Management, Eindhoven University of Technology, Eindhoven, NL.

Muthén, Bengt O., and Tihomir Asparouhov. 2011. “Multilevel regression mixture analysis.” Pp. 639-57 in *Handbook of Advanced Multilevel Analysis*, edited by J. J. Hox and J. K. Roberts. New York, NY: Taylor and Francis.

Mutz, Rüdiger, Lutz Bornmann, and Hans-Dieter Daniel. 2013. “Types of research output profiles: A multilevel latent class analysis of the Austrian Science Fund’s final project report data.” *Research Evaluation* 22(2):118-133. doi: 10.1093/reseval/rvs038.

Mutz, Rüdiger, and Hans-Dieter Daniel. 2012. “University and Student Segmentation: Multilevel Latent Class Analysis of Students’ Attitudes Towards Research Methods and Statistics.” *British Journal of Educational Psychology* 83(2):280-304. doi: 10.1111/j.2044-8279.2011.02062.x.

Oberski, Daniel L., Geert H. van Kollenburg, and Jeroen K. Vermunt. 2013. "A Monte Carlo Evaluation of Three Methods to Detect Local Dependence in Binary Data Latent Class Models." *Advances in Data Analysis and Classification* 7(3):267-79. doi: 10.1007/s11634-013-0146-2.

Shim, Sungok S., and W. Holmes Finch. 2014. “Academic and Social Achievement Goals in Early Adolescents’ Adjustment: A Latent Class Approach.” *Learning and Individual Differences* 30(1):98-105. doi: 10.1016/j.lindif.2013.10.015.

Snijders, Tom A. B., and Johannes Berkhof. 2008. "Diagnostic Checks for Multilevel Models." Pp. 141-75 in *Handbook of Multilevel Analysis*, edited by J. de Leeuw and E. Meijer. New York, NY: Springer.

Snijders, Tom A. B., and Roel J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Los Angeles, CA: Sage.

Varriale, Roberta, and Jeroen K. Vermunt. 2012. “Multilevel Mixture Factor Models.” *Multivariate Behavioral Research* 47(2):247-275. doi: 10.1080/00273171.2012.658337.

Vasdekis, Vassilis, Silvia Cagnone, and Irini Moustaki. 2012. “A composite likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika* 77(3):425-441. doi: 10.1007/s11336-012-9264-6

Vermunt, Jeroen K. 2003. “Multilevel Latent Class Models.” *Sociological Methodology* 33(1):213-39. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x.

Vermunt, Jeroen K. 2004. "An EM Algorithm for the Estimation of Parametric and Nonparametric Hierarchical Nonlinear Models." *Statistica Neerlandica* 58(2):220-33. doi: 10.1046/j.0039-0402.2003.00257.x.

Vermunt, J. K. 2005. "Mixed-Effects Logistic Regression Models for Indirectly Observed Outcome Variables." *Multivariate Behavioral Research* 40(3):281-301. doi: 10.1207/s15327906mbr4003_1.

Vermunt, Jeroen K. 2008. "Latent Class and Finite Mixture Models for Multilevel Data Sets." *Statistical Methods in Medical Research* 17(1):33-51. doi: 10.1177/0962280207081238.

Vermunt, Jeroen K., and Jay Magidson. 2005. *Technical Guide for Latent Gold 4.0: Basic and Advanced.* Belmont, MA: Statistical Innovations Inc.

Vermunt, Jeroen K., and Jay Magidson. 2013. *Technical Guide for Latent Gold 5.0: Basic and Advanced*. Belmont, MA: Statistical Innovations Inc.

APPENDIX A. LATENTGOLD SYNTAX

The full syntax is given for the first model, only the equations and latent variables change thereafter. The high number of starting value sets and EM and NR iterations were not required for these models, but used for convenience as not to have to revisit and tweak the values.

First model, 3-class and 2-group-class without covariances or direct effects:

```
options
   maxthreads=all;
   algorithm
      tolerance=1e-100 emtolerance=0,0001 emiterations=250000
      nriterations=5000;
   startvalues
      seed=0 sets=500 tolerance=1e-005 iterations=500;
   bayes
      categorical=0 variances=0 latent=0 poisson=0;
   montecarlo
      allchi2 seed=0 sets=0 replicates=500 tolerance=1e-008;
   quadrature
      nodes=10;
   missing
      excludeall;
   output
      parameters=effect  betaopts=wl standarderrors profile
      probmeans=posterior bivariateresiduals estimatedvalues=model
      iterationdetails;

variables
   groupid
      team;
   dependent
      w_rep nominal, w_cre nominal, w_div nominal, w_cap nominal, w_var
      nominal;
   latent
      GClass  group nominal 2,
      Cluster nominal 3;

equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
```

3-class, 2-group-class model, covariance between Variation and Diverse:

```
equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

3-class, 2-group-class model, covariance between Variation and Diverse and direct effect from group-level latent variable on Repetitive:

```
equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster + GClass;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

3-class, 3-group-class model, covariance between Variation and Diverse and direct effect from group-level latent variable on Repetitive:

```
Variables
   …
   latent
      GClass  group nominal 3,
      Cluster nominal 3;

equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster + GClass;
   w_cre <- 1 + Cluster;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

3-class, 3-group-class model, covariance between Variation and Diverse and direct effects from group-level latent variable on Repetitive and Creative:

```
equations
   GClass <- 1;
   Cluster <- 1 | GClass;
   w_rep <- 1 + Cluster + GClass;
   w_cre <- 1 + Cluster + GClass;
   w_div <- 1 + Cluster;
   w_cap <- 1 + Cluster;
   w_var <- 1 + Cluster;
   w_var <-> w_div;
```

APPENDIX B. SURVEY QUESTIONS

The survey questions are part of the Questionnaire on the Experience and Assessment of Work [NL: Vragenlijst beleving en beoordeling van de arbeid (VBBA)].

Repetition	-	In your work, do you repeatedly have to do the same things?
Creativity	-	Does your work require creativity?
Diversity	-	Is your work varied?
Capacity	-	Does your work sufficiently require all your skills and capacities?
Variety	-	Do your have enough variety in your work?

Veldhoven, van, Marc, Theodorus F. Meijman, Jacobus P. J. Broersen, and R. J. Fortuin. 1997. *Handleiding VBBA: Onderzoek naar de beleving van psychosociale arbeidsbelasting en werkstress met behulp van de vragenlijst beleving en beoordeling van arbeid*. [VBBA manual: An investigation of perceptions of psychosocial workload and work stress by means of the Dutch Questionnaire on the Experience and Evaluation of Work]. Amsterdam, NL: SKB.

See also http://www.marcvanveldhoven.com/ques.html