Power and type I error of local fit statistics in multilevel latent class analysis

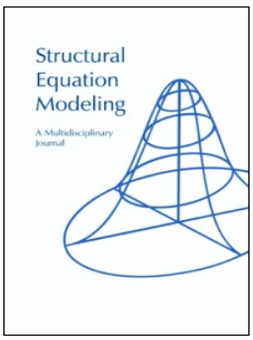
Nagelkerke, E. ; Oberski, D.L.; Vermunt, J.K.

Published in:
Structural Equation Modeling

DOI:
10.1080/10705511.2016.1250639

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record



Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

Power and Type I Error of Local Fit Statistics in Multilevel Latent Class Analysis

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

To cite this article: Erwin Nagelkerke, Daniel L. Oberski & Jeroen K. Vermunt (2017) Power and Type I Error of Local Fit Statistics in Multilevel Latent Class Analysis, Structural Equation Modeling: A Multidisciplinary Journal, 24:2, 216-229, DOI: 10.1080/10705511.2016.1250639

To link to this article: http://dx.doi.org/10.1080/10705511.2016.1250639

Published online: 21 Dec 2016.

Structural Equation Modeling: A Multidisciplinary Journal, 24: 216–229, 2017
ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2016.1250639

Routledge
Taylor & Francis Group

Power and Type I Error of Local Fit Statistics in Multilevel Latent Class Analysis

Erwin Nagelkerke,[1] Daniel L. Oberski,[2] and Jeroen K. Vermunt[1]
[1]*Tilburg University*
[2]*Utrecht University*

In the social and behavioral sciences, variables are often categorical and people are often nested in groups. Models for such data, such as multilevel logistic regression or the multilevel latent class model, should account for not only the categorical nature of the variables, but also the nested structure of the persons. To assess whether the model accomplishes this goal adequately, local fit measures for multilevel categorical data were recently introduced by Nagelkerke, Oberski, and Vermunt (2015). The BVR-group evaluates the variable–group fit, and the BVR-pair evaluates the person–person fit within groups. In this article, we evaluate the performance of these 2 measures for the multilevel latent class model (Vermunt, 2003). An extensive simulation study indicates that whenever multilevel latent class modeling itself is viable, Type I error is controlled and power is adequate for both fit statistics. Thus, the BVR-group and BVR-pair are useful measures to locate important sources of misfit in multilevel latent class analysis.

Keywords: bivariate residual, latent class analysis, local fit, multilevel

Latent class (LC) models can be used to search for classes of systematically similar respondents by considering their responses to a number of discrete indicator items. Analogous to many statistical methods, this model assumes the observations that are classified to be independent. However, dependence often does occur when respondents are observed in naturally occurring groups, leading to a violation of the assumption. When ignored, this dependence will bias the results (Park & Yu, 2015). The multilevel extension to the LC model provides a solution for such cases of nested categorical data (Vermunt, 2003) by taking the grouping into account. Additionally, and maybe more important, it does not only solve the statistical problem of dependent observations, but it substantively allows observed groups to also be classified based on their members (Vermunt, 2003, 2008), providing a simultaneous classification of individuals and groups.

The resulting classification of respondents and groups could be used as a predictor in subsequent analyses (e.g., Roosma, Van Oorschot, & Gelissen, 2015), or covariates can be added to the model to try and substantively explain the classes after an exploratory or confirmatory classification (e.g., Fagginger Auer, Hickendorff, Van Putten, Béguin, & Heiser, 2016; Tomczyk, Hanewinkel, & Isensee, 2015). Regardless of the approach, in both these cases the quality of the classification has a direct influence on the quality of the eventual outcomes of interest, and the fit of the measurement model should be carefully considered before continuing with further analyses.

Central to the model fit in multilevel LC analysis are two assumptions of conditional independence given the latent variables. On the lower level the assumption is that all dependence between items is captured by the latent variable, thus assuming conditional independence of the indicators given the LC variable. This assumption is identical to that of a regular LC model. On the higher level a similar assumption is made, where the observed group members are assumed conditionally independent given the higher level latent variable.

Correspondence should be addressed to Erwin Nagelkerke, Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg,The Netherlands E-mail:.e.nagelkerke@tilburguniversity.edu

In such relatively complex models, with assumptions on two levels and the distinct substantive goals of reproducing the overall and within-group responses, local fit statistics are of increased importance. Traditionally, global fit indexes such as Akaike's information criterion and Bayesian information criterion are used to examine whether all of the assumptions hold, relative to some measure of model complexity. However, this approach has the disadvantage that misspecifications that are small relative to the model complexity might in fact still be harmful to subsequent analyses of interest (Oberski, 2014). Furthermore, their use generally limits itself to the comparison of estimated models in practice. With a nigh infinite number of model specifications, the selected, best fitting model out of the estimated alternatives might very well contain misspecifications and assumption violations. For this reason, the global fit measures can best be supplemented with measures of local fit that examine the strength of evidence against individual model assumptions.

To examine local fit in models for multilevel categorical data, Nagelkerke, Oberski, and Vermunt (2015) recently proposed the BVR-group and BVR-pair measures. Both are in line with the bivariate residual (BVR) proposed by Vermunt and Magidson (2013b) that measures how well the item–item dependence is captured by a single-level LC model. The two multilevel fit measures are comparable, but test how well the model captures the group–item dependence and person–person dependence related to the higher level of the model. All three, the BVR-group, BVR-pair, and BVR, take the form of a Pearson residual, but despite this resemblance they do not follow an asymptotic chi-square distribution. p values can nonetheless be obtained relatively easily by means of a parametric bootstrap (Oberski, Van Kollenburg, & Vermunt, 2013).

The two higher level residuals respectively aim to detect misfit related to the conditional independence assumption and substantively correct reproduction of the data. The BVR-group signals residual dependence between observed group membership and indicator items. When such residual dependence exists it is an indication of the model not fitting one or more of the groups correctly, implying that the model does not fully capture the between-group differences. The BVR-pair signals residual dependence between persons that are members of the same group. This residual dependence is also indicative of the model not correctly capturing the nested structure of the data, but here the focus is on the within-group similarities of the group members.

In Nagelkerke et al. (2015) only a limited simulation study is provided, however, and little is currently known about the properties of the two statistics. With an extensive simulation study we here aim to more thoroughly investigate the power and Type I error of the bootstrapped BVR-pair and BVR-group. Of primary interest is whether and under what conditions the two statistics have enough power to detect several types of misspecification of the multilevel LC model. The misspecifications of the model that are considered are closely related to the two assumptions of conditional independence added by the multilevel extension to the LC model; that is, the assumption that the members of an observed cluster in the data are independent conditional on the higher level latent variable to achieve a group-level classification, and the assumption that observed group membership and the individual responses are conditionally independent to correctly reproduce the observed responses within the observed groups (Vermunt, 2003).

It should be noted that the context of the study is confined to multilevel LC analysis for which the statistics are originally developed, but that they can be obtained for any method that models nested categorical data, such as multilevel IRT. The two residuals namely aim to test for the correct modeling of within-group similarities and between-group differences by contrasting the observed and expected frequencies. Whether these expected frequencies are obtained from a multilevel LC model or an alternative method does not impact the way in which the eventual values are obtained.

The remainder of this article is structured as follows. In the following section the multilevel LC model is briefly introduced. Next the BVR-group and BVR-pair statistics are described, after which the design of the simulation study, including the bootstrap procedure are discussed. The results of the simulation study and the conclusions that can be drawn in terms of Type I error and power are presented in the final two sections.[1]

MULTILEVEL LATENT CLASS MODEL

The multilevel LC model is described using two equations. Both strongly resemble the expression of a regular LC model that classifies individuals based on the probabilities of their responses. The equation for the lower level of a multilevel model does exactly the same, but to take into account the nested structure of the data the response probabilities are made conditional on the group-class membership. To classify the groups and obtain this group-class membership, the higher level equation describes the marginal probabilities of the combined response patterns of the group members of observed groups; that is, it describes the vector of response patterns that is obtained by combining all members of a group (Vermunt, 2003, 2008).

Let the lower level latent variable be denoted as η_{ij}, classifying units in C LCs, with one class referred to as c. The higher level latent variable is denoted ζ_j, with G group-level LCs, one of which is denoted g. Here the response of individual i in group j to item k is denoted y_{ijk}, with a total of J groups, all having n_j members summing to N, and K items having R_k categories. The vector of responses of individual i in group j to all K items is denoted y_{ij}, with r referring to a particular answer pattern and r_k referring to

[1] Appendices and additional resources can be found online at the Open Science Framework, at the permanent URL: osf.io/23mp2.

one particular response to item k. Assuming conditional independence, the lower level of the model is expressed as:

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

When the conditioning on the group-level latent variable is removed, Equation 1 is identical to that of a regular LC model. Without this conditioning the probability of observing a certain pattern of responses r is the sum over the unconditional probability of class membership multiplied by the product of all conditional probabilities of observing the separate responses r_k. In turn, conditioning all these terms on the group-level classes $(\zeta_j = g)$ allows the classification of groups.

Given the lower level expression, the higher level now describes the classification of groups based on their members. Here the vector of all response patterns of units within group j is denoted as y_j, with s denoting a particular combination of response patterns. The conditional independence assumption on this level relates to the units within groups, where not the responses to single items, but the entire response patterns of group members are assumed independent (Vermunt, 2003). The upper level can then be expressed as:

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

This second equation likewise resembles that of a regular LC model, but now the full vector y_j of individual response patterns s in group j is described as a combination of the size, or prevalence, of group-level LC g, and the conditional probabilities of observing the combination of individual answer patterns r.

Equations 1 and 2 describe the most general form of the multilevel LC model, in which both the class sizes and the response probabilities are allowed to vary across group-level LCs. This general form is hardly ever used, because of the difficulty interpreting group clusters with a completely different lower level structure. There are two common ways to constrain the model. Setting $Pr(\eta_{ij} = c|\zeta_j = g) = Pr(\eta_{ij} = c)$ fixes the class membership on the lower level to be independent of that on the higher level, but allows the response probabilities to be estimated freely. The second and most common constraint $Pr(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) = Pr(y_{ijk} = r_k|\eta_{ij} = g)$ inversely fixes the response probabilities on the lower level to be independent of the higher level class membership, but allows the class sizes to be estimated freely. The latter constraint leads to the model that simultaneously classifies respondents and the groups in which they are nested (Lukočienė, Varriale, & Vermunt, 2010).

MULTILEVEL LOCAL FIT STATISTICS

The idea behind the two fit statistics for the higher level of the multilevel LC model is relatively straightforward. Given that LC analysis is concerned with categorical indicators, and both the substantive goal of the model as well as the assumptions it makes can be reduced to adhering to a conditional independence assumption, a test comparable to a chi-square test is an intuitive solution. Both statistics can then compare the dependencies captured by the model to the dependencies present in the data. Because the asymptotic distribution is unknown, in the following the Type I error and power are considered for the bootstrap of the measures.

BVR-GROUP RESIDUAL

The BVR-group is concerned with the average model fit across groups, and quantifies the covariance between observed groups and items that is not captured by the model. When such residual covariance exists, the observed group membership still affects the response probabilities of group members, implying that between-group differences are not fully captured by the group-level latent variable; that is, the BVR-group tests whether the observed response frequencies within the observed groups are adequately reproduced by the model of interest.

The expectation under a well-fitting model is, of course, that the expected and observed response frequencies are close to identical. For the within-group frequencies, this implies that, given the model, the indicator variables should be conditionally independent of observed group membership. To test this, a Pearson residual can be obtained by cross-tabulating the observed and expected frequencies within all groups. This residual is then indicative of all the uncaptured variation caused by observed group membership.

The expected frequency, denoted as m_{jr}, can be obtained from the model as the individual probability of giving a certain response $Pr(y_{ijk} = r_k)$ and summing this probability over the group members:

$$Pr(y_{ijk} = r_k) = \sum_{g=1}^{G} Pr(y_{ijk} = r_k|\zeta_j = g) Pr(\zeta_j = g|\boldsymbol{y}_j = s), \quad (3)$$

with

$$Pr(y_{ijk} = r_k|\zeta_j = g) = \sum_{c=1}^{C} Pr(y_{ijk} = r_k|\eta_{ij} = c, \zeta_j = g) Pr(\eta_{ij} = c|\zeta_j = g). \quad (4)$$

Then,

$$m_{jr} = \sum_{i=1}^{n_j} Pr(y_{ijk} = r_k). \quad (5)$$

These equations can be simplified in a model without covariates to multiplying the probability of a response with the number of group members, rather than the more general sum over n_j in Equation 5. The observed response frequencies, denoted n_{jr}, are a simple count of the responses given by the group members. The BVR-group then equals:

$$BVRgroup_{.k} = \frac{1}{(R_k - 1)(J - 1)} \sum_{j=1}^{J} \sum_{r=1}^{R_k} \frac{(n_{jr} - m_{jr})^2}{m_{jr}}. \quad (6)$$

As shown in Equation 6, a separate residual is computed for each group and each response category, all of which are subsequently summed over the J groups, and R_k categories. Additionally, the resulting statistic is divided by $(R_k - 1)(J - 1)$, which is the number of nonredundant parameters in the cross-table, standardizing the BVR-group so it is not affected by the number of groups in the data and the number of categories of the variable.

Because the focus is mainly on item-specific misfit, the BVR-group is here obtained per item. However, by removing the sum over J groups the statistic can be obtained per group to inspect whether misfit originates from the model not fitting specific groups. Moreover, by not summing over the R_k categories it can be obtained per response category, which could be useful when extreme responses are a plausible cause of misfit.

BVR-PAIR RESIDUAL

On the higher level of a multilevel LC model the assumption is made that given the group-level latent variable the response patterns of nested units are conditionally independent. That is, the full response patterns r of all n_j group members in Equation 2 are assumed to be conditionally independent. The BVR-pair tests for violations of this assumption. When there is residual dependence among the members of observed groups, the within-group similarity is not correctly reproduced by the group-level latent variable. In other words, the nested structure of the data is not fully captured by the model.

Because the assumption on this level does not relate to the items, but to the units, the group members need to be related to one another. This is done by creating all possible pairs of units within an observed group to obtain the pairwise response frequencies. When the assumption that all dependence between the units is captured by the model holds, the expected and observed frequencies would again be in agreement. Here, by considering the pairwise frequencies, this would be indicative of the response of unit i and i', rather than item k and k', being locally independent.

The expected frequency of a pair of responses is obtained using the joint probability of unit i giving response r, and unit i' giving response r' to item k:

$$Pr(y_{ijk} = r_k, y_{i'jk} = r'_k) = \sum_{g=1}^{G} Pr(y_{ijk} = r_k|\zeta_j = g) Pr(y_{i'jk} = r'_k|\zeta_j = g) Pr(\zeta_j = g|\mathbf{y}_j = \mathbf{s}). \quad (7)$$

After that, the expected frequency $m_{krr'}$ can be obtained by multiplying with the number of possible pairs within the group:

$$m_{krr'} = \sum_{j=1}^{J} (n_j(n_j - 1)/2) Pr(y_{ijk} = r_k, y_{i'jk} = r'_k). \quad (8)$$

Essentially, the probabilities of all possible combinations of the discrete responses to a single item are obtained per group and multiplied by the number of possible pairs of members in a group.

Obtaining the observed frequency ($n_{krr'}$) can be thought of as creating a cross-table for each pair. This table would identify the combined response of unit i and i' to item k, as only one cell would have a value of one. Subsequently summing these tables over all pairs results in the pairwise frequency for that particular item (for an illustration see Nagelkerke et al., 2015).

Important to note here is that in a multilevel LC model the ordering of the responses does not matter for the probability of a pair. For example, two units responding to a dichotomous item forming a yes–no pair have a probability that is identical to a no–yes pair. Yet, in practice the observed frequencies for such pairs will almost always differ depending on how the data set is ordered. Therefore, patterns with the same, but differently ordered responses are summed when obtaining the BVR-pair statistic:

$$BVRpair = \frac{J}{N} \frac{1}{R_k(R_k - 1)/2} \left[\sum_{r}^{R_k} \sum_{r'>r}^{R'_k} \frac{((n_{krr'} + n_{kr'r}) - (m_{krr'} + m_{kr'r}))^2}{m_{krr'} + m_{kr'r}} + \sum_{r} \frac{(n_{krr} - m_{krr})^2}{m_{krr}} \right]. \quad (9)$$

To arrive at the BVR-pair, the raw residual is divided by the number of nonredundant parameters in the table. Given the symmetry on the off-diagonals, this is a division by $R_k(R_k - 1)/2$. Additionally, because the theoretical maximum value increases as a triangular sequence with n_j, the statistic is divided by the average group size, simply to reduce the absolute values.

Because of this triangular increase, one more problem needs to be solved when the groups are of different sizes, as it causes units in larger groups to be in far more pairs than those in smaller groups. As a result, the observed and expected

marginal frequencies can differ, where $(n_{krr} - m_{krr}) \neq (m_{krr'} + m_{kr'r})$. Such a difference affects the values of the BVR-pair, although not being indicative of residual dependence. To avoid the influence of these marginal differences on the BVR-pair, iterative proportional fitting is used to update the table with expected pairwise frequencies so that it retains its cross-product ratios (Bishop, Fienberg, & Holland, 1975), but has the observed marginal frequencies (again, for an illustration see Nagelkerke et al., 2015).

Note that the computational complexity of obtaining the BVR-pair is primarily determined by the number of items, possible responses, and group classes. For the BVR-group this would be the number of possible responses and groups. The sample size in terms of the number of observations and groups only affects frequency counts and thus adds little time to the required computations. However, because the residuals do not follow a known asymptotic distribution, a bootstrap is required to obtain p values. This, of course, does increase the computational times, and might make obtaining the results computationally intensive for truly big data sets. For $N = 62{,}500$ the average time for 250 bootstraps in this study was approximately 9 minutes on a 4 × 3.30 Ghz processor.

SIMULATION DESIGN

The misfit that the two residual statistics aim to capture are the model not fitting observed groups, causing residual conditional dependence between group membership and indicator items, and the model not capturing all within-group dependence between units, causing a residual dependence between pairs of observations. These types of misfit can be remedied in the multilevel LC model by either allowing a direct effect from the group-level LC variable on one or more of the indicators, or adding additional group-level LCs. To test the power of detecting such misfit this logic is reversed, whereby a population model is assumed containing, for instance, a direct effect and analyzing these data with a misfitting model excluding that particular parameter. To investigate the power and Type I error of the two residuals, a Monte Carlo simulation is used to evaluate a range of different models, with differing types of misspecification.

VARIABLES AND FACTORS

The power in LC models themselves is primarily dependent on two mutually influencing factors, namely the amount of information and entropy, or class separation. The former is what affects the power of any statistical test, and depends on commonly studied factors such as the sample size, the size of observed groups, and the number of observed items. In LC analysis an important additional aspect is how distinctly different the LCs are, which is affected by the number of classes and the effect sizes of the parameters.

The factors that are varied and affect the structure of the sample are:

- Number of observed groups: 50, 100, or 250 groups.
- Number of observed group members: 10, 50, or 250 group members.
- Number of indicator items: 6 or 10 items.

The factors that are varied and can be thought of as model specific are:

- Number of lower level classes: two or three classes.
- Number of higher level classes: two or three classes.
- Log-linear effect from the lower level latent variable to the indicator: 0.424 or 0.693 (conditional response probabilities).
- Log-linear effect from the higher level latent variable on one or two indicators: 0.000 (no direct effect), 0.201 or 0.511.
- Log-linear effect from the higher level latent variable on the lower level latent variable (see Table 1).

The log-linear parameters are effect coded, leading to conditional probabilities of 0.7 or 0.8 in one lower level class, and the complement of 0.3 or 0.2 in the other. In conditions with three classes, half of the items in the middle class have a conditional probability of 0.7 or 0.8, and the complement for the other half of the items. For examples of the conditional probabilities in the different population models, see Appendix A at osf.io/23mp2. All intercept values are kept at zero, which implies equal class sizes. Of course, by crossing the number of groups and their members, different sample sizes are obtained, namely 500, 1,000, 2,500, 5,000, 12,500, 25,000, and 62,500.

The power to detect misfit is considered for eight types of misspecification, as well as for the correctly specified models to estimate the Type I error.

The misspecifications considered are:

- A missing class on the lower level.
- A missing class on the higher level.
- A missing direct effect (weak and strong).
- A missing direct effect when there are two direct effects (weak and strong).
- A missing class on the higher level and a missing direct effect.

TABLE 1
Logit Parameters for the Higher Level: Effects of the First Group-Level Class on the Two or Three Lower Level Classes

	2 Lower Level Classes			*3 Lower Level Classes*		
	Logit 1	*Logit 2*	*Logit 3*	*Logit 1*	*Logit 2*	*Logit 3*
2 group class (W)	0.424	–0.424	—	0.196	0.014	–0.209
Group class (W)	0.424	0.000	–0.424	–0.514	1.027	–0.514
Group class (S)	0.693	–0.693	—	0.928	0.341	–1.269
Group class (S)	0.693	0.000	–0.693	–0.693	1.386	–0.693

Note. For examples of the resulting conditional probabilities see Appendix A.

DESIGN OF EXPERIMENTS

It needs to be taken into account that the model itself is relatively complex, and that the residuals require a parametric bootstrap. This leads to many model reestimations because the bootstrap needs to be performed for each Monte Carlo replication. To reduce the computational intensity and keep the study feasible, a smaller design than full factorial was chosen, whereby the higher order interactions between the variables are deliberately left confounded (see, e.g., Lundstedt et al., 1998). The idea is identical to a fractional factorial design, or I^{k-p} design, but because the variables of interest have different numbers of levels the setup does not result in a true fraction of the full factorial. Using SAS JMP (see, e.g., Montgomery, 2012) a design consisting of 422 conditions was generated that has no aliasing for the main effects, nor for the second- and third-order interactions in the full set of conditions. This way, only one fifth of the computations are needed. The compromise is that higher order interactions cannot be estimated, although generally four variable interaction effects and up are of limited practical value. It must be noted that these are interactions on the variable level, which means that the limitations occur on the factor level, where certain combinations are not taken into account. For example, all low N conditions have an observed group size of 10.

MONTE CARLO AND BOOTSTRAP

The Monte Carlo simulation is conducted using a combination of R (R Development Core Team, 2015) and LatentGOLD 5.0 (Vermunt & Magidson, 2013a), whereby R is used to generate syntax and postprocess the results. Based on the desired population model, LatentGOLD is used to generate a data set, which is subsequently analyzed with either a correctly or misspecified estimation model. To obtain the p value for the BVR-pair and BVR-group statistics a bootstrap is conducted using the maximum likelihood values that follow from the estimation.

The bootstrap data are obtained by sampling group-class membership based on the class prevalences, class membership conditional on the sampled group-class membership, and finally the responses conditional on both the sampled memberships. The p values for the BVR statistics are then obtained by computing the proportion of bootstrap samples in which the residuals are larger than in the original model. This process of generating data, analyzing the data, and performing the bootstrap is repeated for the desired number of Monte Carlo replications. The proportion of significant p values of the total number of replications then is indicative of the power. For the null models, both the number of bootstrap samples and Monte Carlo replications are set to 250. For the misspecified models both are set to 500 for the large majority of models, with the exception of several conditions with a very large N and weak class separation that are computationally extremely intensive.

RESULTS

First the results for the null models are discussed, because estimations of power cannot be interpreted when the nominal alpha levels are incorrect.

Type I Error

Table 2 depicts the average proportion of significant BVR values at the $\alpha = .05$ level for the first indicator variable when a direct effect is present, and the third when there is not, where the mean is computed over all conditions that satisfy a particular factor. Note that this reverses the interpretation of the numbers in Table 2, where all values lower than .05 are too liberal, because there are too few significant values indicating misfit. The reason for depicting the third indicator is that, when present, direct effects from the group-level latent variable on an indicator are on indicators one, two, or both. L here refers to the number of conditions that the average is based on, because not all factors occur equally often due to the study design.

Overall, the BVR-group and BVR-pair statistics are very close to the nominal alpha level, regardless of the condition over which the mean is computed. The BVR-group, however, is slightly too liberal, especially in the conditions with a smaller N. This might in large part be due to the statistic taking the form of a χ^2 test, which becomes more conservative as sparseness increases, also when a parametric bootstrap is used (Langeheine, Pannekoek, & Van De Pol, 1996; Von Davier, 1997). That is, the χ^2 test is too conservative in that the null hypothesis that there is no misfit is not rejected, making the BVR-group too liberal. In these conditions the number of groups is set to 50 or 100 with only 10 members, leading to relatively sparse frequency tables. This is in line with the BVR-pair not showing any problems, as it is obtained on an $R \times R$ rather than an $R \times J$ table, in addition to the number of pairs being far larger than the number of observations.

The left side of Table 2 depicts the Type I error for the first indicator item and only for conditions in which a direct effect on the indicator is present. A direct effect being present causes slightly more variation, but the overall results are still good in terms of the Type I error rate. The most problematic cases are clearly those where little information per group is available, especially when there are many small groups. This can, for example, be seen from the conditions $N = 500$ and $N = 1{,}000$, both of which have 10 observed cases per group. Again, this can largely be attributed to sparseness.

Inspecting the BVR that tests the local independence between items (not reported) on the lower level of the model does not indicate any problems with the model itself either. Where it might have been possible that strong group-level dependencies affect the fit or the fit statistics on the lower level, there is no evidence of this occurring.

TABLE 2
BVR-Group and BVR-Pair Mean and Standard Deviation of the Proportion of Significant Bootstraps per Main Factor for an Indicator Item With and Without Direct Effects (Type I Error)

	With a Direct Effect on the Item					*Without Direct Effect on the Item*				
	L	*BVR-Group*	*SD*	*BVR-Pair*	*SD*	*L*	*BVR-Group*	*SD*	*BVR-Pair*	*SD*
Classes = 2	159	.040	.020	.047	.015	216	.049	.014	.050	.013
Classes = 3	156	.048	.015	.050	.013	206	.048	.016	.051	.014
Group Cl. = 2	158	.047	.015	.051	.012	208	.050	.014	.051	.014
Group Cl. = 3	157	.041	.020	.047	.015	214	.048	.016	.050	.014
Items = 6	160	.044	.018	.048	.014	215	.048	.016	.049	.014
Items = 10	155	.044	.018	.049	.015	207	.049	.014	.051	.013
Groups = 50	103	.044	.018	.049	.015	137	.049	.016	.050	.013
Groups = 100	103	.042	.017	.046	.013	142	.048	.015	.050	.013
Groups = 250	109	.046	.020	.051	.015	143	.048	.014	.051	.015
Members = 10	106	.034	.021	.046	.016	143	.044	.016	.051	.012
Members = 50	103	.048	.015	.050	.013	139	.050	.013	.051	.014
Members = 250	106	.050	.013	.050	.014	140	.052	.014	.049	.014
N = 500	35	.036	.020	.048	.018	48	.047	.017	.049	.011
N = 1,000	33	.029	.017	.041	.012	47	.042	.016	.051	.012
N = 2,500	73	.042	.021	.050	.014	95	.047	.015	.051	.014
N = 5,000	34	.045	.014	.048	.012	46	.050	.013	.050	.013
N = 12,500	67	.049	.015	.049	.013	88	.050	.014	.051	.015
N = 25,000	36	.051	.012	.049	.012	49	.053	.013	.050	.013
N = 62,500	37	.051	.014	.053	.015	49	.051	.013	.049	.017
Class sep. = low	155	.045	.019	.050	.015	205	.049	.015	.050	.013
Class sep. = high	160	.043	.017	.047	.012	217	.048	.014	.051	.014
Group sep. = low	156	.043	.019	.049	.015	209	.048	.015	.051	.014
Group sep. = high	159	.045	.017	.048	.013	213	.049	.014	.050	.013
Overall	315	.044	.018	.049	.014	422	.049	.015	.050	.014

Power to Detect Ignored Nesting

The most fundamental type of misspecification considered in the simulation study follows from specifying a model with too few classes on the group level when only two are present in population. This results in a model that ignores the nested structure of the data altogether. As can be expected, the parameter estimates and LC solution in this situation are strongly biased, both in the parametric (Kaplan & Keller, 2011) and nonparametric (Park & Yu, 2015) multilevel LC model.

Table 3. depicts the power to detect the presence of an additional group-level class when only one is specified in the analysis, which is identical to specifying a regular LC model. The full conditions are presented here, because splitting on all factors would result in a largely empty table, whereas confounding any of the factors would not provide the full picture. All conditions with a larger sample size are omitted, as the power equals one.

Preferably the BVR-group and BVR-pair values should be significant for each separate indicator item when detecting a missing class. The dependence that is not captured by the model is namely affecting all of the indicators. However, it is not necessarily the case that none of the group-level dependence is modeled on the lower level, and the fit of some of the indicators might well be acceptable. Vice versa, if only one or two of the indicator items were identified as not being reproduced correctly by the model, the conclusion of a missing class would probably not be drawn, and model improvements would focus primarily around these specific items. Therefore, what is reported in the table are three proportions for the BVR-group, namely the power when only looking at the first item, the proportion of Monte Carlo replications where at least one out of the K residuals is significant, and the proportion of replications where 50% or more of the BVR-group values are significant (so for 3 or 5 out of 6 or 10 indicator items). For conciseness, the BVR-pair values are not reported, because they show an identical pattern, albeit slightly less powerful.

The power of the BVR-group to detect that something is wrong when completely ignoring the nested structure of the data, while there are two group-level classes, is close to one in practically all situations. Judging from the second to last column of Table 3. only in two extreme situations the combined power over all indicator items drops below .90. In these two cases class separation on the group level is almost nonexistent, as shown in Table A2 and Table A3, with an estimated entropy of 0.317 and 0.323, respectively. Combined with the small sample size and associated uncertainty about the classification, the dependence can actually be modeled without a group-level class. The nested structure

TABLE 3
Power to Detect Ignoring the Nested Structure: The Last Three Columns Indicate the Power to Reject Fit for Item 1, at Least One Item, and at Least Half of the Items

Sample			*Class Separation*				*BVR-Group*		
N	*Groups*	*Group Size*	*Items*	*Level 1*	*Level 2*	*C*	*Item 1*	*Minimum 1*	*50%*
500	50	10	10	L	L	3	0.062	0.476	0.002
500	50	10	6	H	L	3	0.114	0.464	0.024
500	50	10	6	L	H	3	0.654	0.992	0.830
500	50	10	10	H	L	2	0.780	1.000	0.906
1,000	100	10	10	L	L	2	0.468	0.996	0.582
1,000	100	10	6	L	H	2	0.958	1.000	1.000
1,000	100	10	10	H	H	2	1.000	1.000	1.000
2,500	50	50	10	H	H	3	1.000	1.000	1.000
2,500	50	50	10	L	H	2	1.000	1.000	1.000
2,500	250	10	10	L	H	3	0.998	1.000	1.000
2,500	250	10	6	H	H	3	1.000	1.000	1.000
5,000	100	50	10	L	L	3	0.276	0.914	0.098
5,000	100	50	6	H	L	3	0.656	0.986	0.862
5,000	100	50	6	H	L	2	1.000	1.000	1.000

in these situations is only detected with a truly large sample (power equals one in the omitted conditions with $N \geq 12{,}500$).

However, when misfit on any one of the items is detected, it is not necessarily the case that misfit is found for all separate items. Generally more than half of the items will be reported as problematic, but two remarkable discrepancies are the first $N = 1{,}000$ and $N = 5{,}000$ conditions. At least one BVR-group value is significant for these conditions, but rarely more than half indicate misfit. Inspecting these two conditions further the average number of significant BVR-group values over all the Monte Carlo replications are 4.91 and 2.50 out of 10, so it is still likely that misfit in multiple indicators is detected for these two cases, although it might be too few to point to an unmodeled group-level class.

In practice, this means that the BVR-group will detect the nested structure in more typical situations where N is not too small and class separation at the group level is not too low. Situations with small N and an extremely low class separation at the group level should already be cause for concern in the sense that there might not be a nested structure strong enough to model. In all other situations, at least one of the BVR-group values will generally be significant with an $N \geq 1{,}000$. Given that this is a situation where one (identical to no) group-level class is modeled, there is no other way to address this dependence than adding group-level classes. The exact number of significant BVR-group values is less relevant in this respect, but will be returned to in the next section.

Power to Detect a Missing Group-Level Class

A logical next step to consider is the situation in which too few, rather than no, group-level classes are specified. From Table 4 it is evident that the power to detect a third group-level population class as missing when two are specified is markedly lower. Inspecting the conditions more closely, the power of the BVR-group is acceptable in conditions with larger separation between the classes. Note here that separation on the lower level also directly affects separation on the higher level, as can be seen by the conditional probabilities in the population models, illustrated by the group-level classes in Table A4 and A5. The stronger dependence between group membership and the responses of its members in turn leads to a higher residual dependence when not modeled correctly.

In case the classes are not as strongly separated, more information is required to detect that the population might contain an additional class. However, this is not achieved by simply having a larger sample, but requires the sample size at either level to be sufficient. That is, enough information needs to be available on both the higher and lower level to detect residual dependence on the higher level. This is not too surprising given the model specification, whereby observed groups are essentially classified based on the lower level class membership of their members. Although a similar sample size recommendation for multilevel LC analysis is not readily available, the consistently high power in conditions where group size is 50 is in line with previous research on multilevel logistic regression (Moineddin, Matheson, & Flora Glazier, 2007).

To further clarify the mutual effect between the number of groups and their size several additional conditions were considered. The marked $N = 2{,}500$ conditions in Table 4 are identical to the conditions with an N of 1,000. Comparing these four conditions to the lower N ones clearly shows that increasing the number of groups when they are very small barely increases the power to detect the correct nested structure, whereas the sample size more than doubles. The $N = 500$ conditions show that conversely increasing the

TABLE 4
Power to Detect a Missing Group-Level Class: The Last Three Columns Indicate the Power to Reject Fit for Item 1, at Least One Item, and at Least Half of the Items

Sample			*Class Separation*				*BVR-group*		
N	*Groups*	*Group Size*	*Items*	*Level 1*	*Level 2*	*C*	*Item 1*	*Minimum 1*	*50%*
500	50	10	6	L	H	2	0.062	0.238	0.002
500	50	10	6	H	H	3	0.568	0.998	0.962
500	10	50	6	L	H	2	0.104	0.504	0.020
500	10	50	6	H	H	3	0.476	0.972	0.956
1,000	100	10	6	H	L	2	0.044	0.224	0.004
1,000	100	10	10	L	H	2	0.046	0.434	0.000
1,000	100	10	6	L	L	3	0.220	0.806	0.120
1,000	100	10	10	H	H	3	0.560	1.000	1.000
2,500	250	10	6	L	L	2	0.040	0.242	0.004
2,500	250	10	10	H	L	2	0.052	0.414	0.000
2,500	250	10	6	H	H	2	0.110	0.476	0.016
2,500	250	10	6	L	L	3	0.440	0.984	0.540
2,500	250	10	10	L	H	2	0.064	0.528	0.000
2,500	250	10	6	H	L	2	0.036	0.248	0.004
2,500	250	10	10	H	H	3	0.556	1.000	1.000
2,500	50	50	6	H	L	2	0.202	0.728	0.124
2,500	50	50	6	L	H	3	0.530	1.000	1.000
2,500	50	50	10	H	H	3	0.542	1.000	1.000
2,500	50	50	10	L	L	3	0.554	1.000	0.998
5,000	100	50	6	L	L	2	0.106	0.498	0.018
5,000	100	50	10	H	L	2	0.326	0.958	0.274
5,000	100	50	6	H	L	3	0.496	1.000	1.000
5,000	100	50	6	H	H	2	0.982	1.000	1.000

group size for a small number of groups does not increase the power in a similar fashion. Whether this is due to too little power of the multilevel LC model to detect the true structure, or the power of the BVR-group to detect the failure of modeling, the true structure is hard to disentangle and both might be occurring.

A final remark on Table 4 is that the BVR-group residual is generally more powerful with three, compared to two lower level classes, even when the higher level classes are further apart in terms of conditional response probabilities. This is a general trend, which can best be explained in terms of the population data. When the group members belong to a higher number of distinct classes, the classification of the groups is automatically more fine-grained as well. That is, there is a more diverse composition of the group members in terms of the lower level class that they belong to. This diversity will create a larger effect of observed group membership on the probability to give a certain response, and hence, failing to model the effect will create a larger residual. Related to this, note that the number of indicator items in the condition is not further discussed, because it causes no systematic differences in the power estimates.

The model here turns out to be quite good at redistributing the residual dependence. The practical implication of these findings is that for weakly defined classes or samples with small groups, residual dependence is not picked up by one particular item, or a large majority of the items. Although this implies that groups should have around 50 members, it might not actually be extremely problematic in terms of model adjustments. The dependence is truly redistributed and generally ends up in one or two items that do show problems. When these items are addressed, for example, by allowing a direct effect between the item and the group-level latent variable, it will not resolve the problem and other indicators will show residual dependence (for an example see the application in Nagelkerke et al., 2015). This will either cause many BVR-group values to start indicating problems, or iteratively cause a few to show problems until an additional group-level class is the best solution in terms of parsimony. Of course, addressing problematic items blindly to merely reduce the residual dependence does lead to capitalization on chance, and will most likely not result in finding the population model. Given the results, a good exploratory approach would be to use global fit statistic or information criteria to determine the number of classes, attempting to resolve any residual dependence with theoretically sensible parameters, and if the dependence returns in other indicators to increase the number of classes.

Power to Detect Missing Effects

A second general type of misspecification concerns a missing direct effect from the group-level latent variable to one of the indicators. This model mimics the situation in which observed group membership is not conditionally independent

from the indicators, and the univariate item distributions are not properly reproduced by the model. Here the ideal outcome is reversed from the detection of a missing class in terms of the residuals, where the BVR-group and BVR-pair should only detect misfit in the item to which the direct effect pertains.

In Table 5 the power of detecting a weak missing direct effect is presented; that is, an effect that causes a small residual dependence between observed group-membership and the first indicator item. With a few exceptions, the BVR-pair has notably higher power to detect the misspecification. A quick summary of the results is that power increases with sample size and is generally higher for larger, rather than more, groups. The latter is also confirmed by inspecting several additional conditions with 10 groups with 50 members, otherwise identical to the $N = 500$ conditions presented, which all have slightly higher, but still insufficient power. An extra set of conditions is also used for the effect of having more indicator items, which increases power slightly. However, having four additional indicators primarily increases lower level class separation, which in turn only substantially affects group-level class separation when the group-level effects are strong. That is, it primarily increases power in already high-power conditions, and has a limited effect on low-power conditions.

For the other factors, the results are somewhat paradoxical. First, it seems that in small sample conditions a stronger separation of the classes generally leads to lower power. However, this is an artifact of the importance of the direct effect to separate the classes. When the effect is highly important for class separation (i.e., creates a large discrepancy between the entropy of the model and the population) it is picked up in conditions with weakly separated classes as it creates very large residual dependencies. Furthermore, in conditions with more classes the power is generally lower. The reverse at first seems more likely, as there is more information on the correct specification. However, more classes simply make it easier to model dependencies, as there are a lot more parameters that can be used to compensate for the missing direct effect.

It should be noted that the direct effect here is an effect coded logit of 0.201, which creates only very minor changes in conditional probabilities. The power to detect a missing direct effect with a stronger effect of 0.511, presented in Table B3, quickly approaches one for all conditions with an $N \geq 1{,}000$. Only the conditions with two group-level and three lower level classes remain an exception, but this is due to class separation being very low. See, for example, Tables A2 and A3, where it is debatable whether there is a nested structure at all.

In Table 6 the results are averaged for the different sample sizes. The average power seems relatively low, but this is due to a few conditions resulting in a power close to zero to detect the weak effect that is missing (see also Table 5). The last four columns give some insight into how precise the residuals are able to identify the problematic variable, as they should preferably not identify other indicators as causing misfit. The BVR-group here does surprisingly well, especially when considering that a direct effect from the group-level latent variable to any of the indicators affects the LC solution (see, e.g., Table A6). When the direct effect in the population is strong enough, excluding it from the model will affect the conditional probabilities for all items in both the lower and higher level classes. In such a case, one group-level class, and thus the members of the observed groups that are classified into that class, will systematically resemble one another more, causing the residual to report uncaptured dependence. This can readily be seen from the BVR-pair value for a strong direct effect and large N. Here the power is large enough to identify the additional dependence that is created between members of the same group by excluding a direct effect from the model, as the BVR-pair residual has a power of close to one to identify both the first and second indicator as problematic.

Yet, this does not occur as persistently as expected. In most of these conditions the BVR-group does not identify the second item as causing misfit up to a certain point. As explained, there is true uncaptured dependence in all indicators due to a missing direct effect, so it can be expected that as the amount of information to identify that dependence increases, such as having $N = 62{,}500$, it is indeed detected. Also, it cannot be expected that these residuals then remain equal to the nominal alpha level. Nonetheless, even with a power to detect the direct effect on the first indicator item of 0.9, mistakenly identifying the second indicator as problematic only occurs in less than 30% of the replications.

Finally, in Table 7 the average power of the more powerful BVR-pair is shown for conditions where one direct effect is missing, but two are present in the population. Comparing the power to that of the BVR-pair for Log(0.5) effects in Table 6, it is clearly harder to detect this misspecification. Similarly comparing Item 3 in Table 7 to Item 2 in Table 6, the false detection rates go up slightly, which is not surprising given the stronger dependencies throughout the data. In large sample studies it is even the case that the BVR-pair values almost always indicate significant misfit on more than half of the indicator items, which could lead to the conclusion that there are too few group-level classes. This might, however, not be extremely problematic, as it is unlikely that adding a group-level class will be able to fully resolve the residual dependence problem, and misfit will still be indicated for the first item. Furthermore, the absolute value of the BVR-pair, rather than its p value, is larger by quite a margin in the majority of cases (42 out of 51). For a selection of single conditions from these averages including absolute values see Table B4.

With respect to the practical use of the BVR-group and BVR-pair, the power differences in the two different types of misspecification could prove informative and can be used

TABLE 5
Power to Detect the Absence of a Weak Direct Effect From the Group-Level Latent Variable on the First Indicator Variable

Sample			*Class Separation*					*Group-Level Entropy*		*Lower Level Entropy*		*Power*	
N	*Groups*	N_j	*Items*	*Level 1*	*Level 2*	*C*	*G*	*Pop.*	*Model*	*Pop.*	*Model*	*BVR-Group*	*BVR-Pair*
500	50	10	6	L	L	2	2	0.646	0.511	0.543	0.540	0.340	0.550
500	50	10	10	L	H	3	3	0.873	0.868	0.649	0.647	0.042	0.088
500	50	10	10	H	H	3	2	0.960	0.953	0.817	0.814	0.108	0.194
500	50	10	6	H	H	3	3	0.935	0.927	0.755	0.748	0.040	0.104
500	50	10	10	H	H	2	3	0.590	0.570	0.943	0.943	0.228	0.478
1,000	100	10	10	L	L	2	2	0.684	0.585	0.704	0.703	0.680	0.918
1,000	100	10	6	L	L	3	3	0.575	0.564	0.416	0.412	0.030	0.076
1,000	100	10	6	L	H	2	2	0.858	0.818	0.613	0.619	0.090	0.574
1,000	100	10	10	H	H	2	3	0.588	0.570	0.943	0.943	0.358	0.736
1,000	100	10	6	H	H	3	2	0.913	0.915	0.677	0.675	0.084	0.222
2,500	250	10	10	L	H	2	3	0.536	0.496	0.716	0.718	0.732	0.980
2,500	250	10	6	L	H	3	2	0.728	0.721	0.419	0.412	0.040	0.102
2,500	250	10	6	H	L	2	3	0.357	0.314	0.826	0.827	0.922	0.970
2,500	250	10	10	H	L	3	3	0.874	0.871	0.830	0.828	0.084	0.614
2,500	250	10	10	H	L	2	2	0.721	0.665	0.940	0.940	0.938	1.000
2,500	250	10	10	H	L	3	2	0.258	0.150	0.787	0.787	0.058	0.064
2,500	50	50	10	L	L	3	3	0.996	0.995	0.602	0.596	0.286	0.490
2,500	50	50	6	L	L	3	2	0.639	0.279	0.333	0.331	0.728	0.674
2,500	50	50	6	L	L	2	3	0.757	0.665	0.535	0.540	0.900	1.000
2,500	50	50	10	L	L	2	2	0.993	0.983	0.709	0.714	0.910	1.000
2,500	50	50	6	H	L	2	2	0.995	0.991	0.834	0.839	0.582	1.000
2,500	50	50	10	H	H	2	2	1.000	1.000	0.949	0.949	0.378	1.000
2,500	50	50	6	H	L	3	3	0.999	0.999	0.712	0.707	0.166	0.776
2,500	50	50	10	H	H	3	3	1.000	1.000	0.868	0.866	0.154	0.826
5,000	50	10	10	L	H	2	3	0.946	0.927	0.727	0.731	0.854	1.000
5,000	50	10	6	L	L	2	2	0.990	0.972	0.554	0.565	0.708	1.000
5,000	50	10	10	L	L	3	2	0.719	0.373	0.481	0.479	0.998	0.946
5,000	50	10	6	H	H	2	3	0.959	0.951	0.845	0.849	0.650	1.000
5,000	50	10	10	H	H	3	2	1.000	1.000	0.819	0.817	0.462	0.998
5,000	50	10	6	H	L	3	3	0.999	0.999	0.712	0.708	0.286	0.972
12,500	250	50	6	L	H	2	2	0.999	0.998	0.626	0.640	0.176	1.000
12,500	250	50	6	L	H	3	3	0.997	0.997	0.596	0.587	0.134	0.698
12,500	250	50	6	L	H	3	2	0.999	0.999	0.571	0.564	0.720	0.980
12,500	250	50	10	H	L	2	3	0.816	0.775	0.938	0.939	1.000	1.000
12,500	250	50	6	H	L	3	2	0.669	0.438	0.625	0.625	0.802	0.102

TABLE 6
Average Power to Detect the Absence of a Direct Effect From the Group-Level Latent Variable on the First Indicator Variable, by Sample Size

					Item 1				*Item 2*			
					BVR-Group		*BVR-Pair*		*BVR-Group*		*BVR-Pair*	
N	*Groups*	*Size*	*L 0.2*	*L 0.5*	*Log 0.2*	*Log 0.5*	*Log 0.2*	*Log 0.5*	*Log 0.2*	*Log 0.5*	*Log 0.2*	*Log 0.5*
500	50	10	5	7	0.152	0.355	0.283	0.475	0.042	0.050	0.049	0.063
1,000	100	10	5	6	0.250	0.849	0.505	0.916	0.044	0.056	0.038	0.090
2,500	250	10	6	6	0.462	0.692	0.622	0.736	0.043	0.069	0.051	0.147
2,500	50	50	8	5	0.513	0.702	0.846	0.936	0.059	0.090	0.136	0.242
5,000	100	50	6	5	0.660	0.942	0.986	0.948	0.066	0.152	0.073	0.389
12,500	250	50	5	6	0.566	0.842	0.756	0.986	0.116	0.282	0.077	0.523
12,500	250	50	5	5	0.974	0.814	1.000	0.996	0.052	0.276	0.158	0.852
25,000	100	250	6	6	0.950	1.000	1.000	0.980	0.075	0.608	0.195	0.671
62,500	250	250	6	7	0.990	1.000	1.000	1.000	0.223	0.780	0.410	0.941

Note. L refers to the number of conditions that the average is based on. Due to the design not all factors occur equally often.

TABLE 7
Average Power of the BVR-Pair to Detect the Absence of a Direct Effect on Item 1 When Two Are Present, by Sample Size

			Log 0.2 Missing					*Log 0.5 Missing*				
N	*Groups*	*Size*	*L*	*Item 1*	*Item 2*	*Item 3*	*50%*	*L*	*Item 1*	*Item 2*	*Item 3*	*50%*
500	50	10	7	0.243	0.045	0.059	0.004	4	0.538	0.055	0.114	0.020
1,000	100	10	5	0.391	0.050	0.071	0.015	6	0.626	0.043	0.115	0.012
2,500	250	10	7	0.651	0.042	0.089	0.017	6	0.659	0.043	0.155	0.095
2,500	50	50	6	0.691	0.052	0.120	0.027	6	0.999	0.050	0.233	0.170
5,000	100	50	6	0.925	0.051	0.216	0.080	6	0.755	0.048	0.548	0.216
12,500	250	50	6	1.000	0.058	0.332	0.150	6	0.996	0.047	0.726	0.559
12,500	250	50	6	0.999	0.043	0.445	0.349	6	0.992	0.059	0.777	0.527
25,000	100	250	6	1.000	0.054	0.673	0.406	6	1.000	0.057	0.934	0.750
62,500	250	250	6	0.936	0.045	0.839	0.588	5	0.992	0.050	0.942	0.924

Note. The remaining effect is log(0.511) on Item 2 in all conditions.

to identify potential model improvements. Where the BVR-group generally has a higher power to detect a missing group-level class, the BVR-pair is better able to detect missing direct effects. Because a missing class has been shown to sometimes cause only one or two BVR-group values to be significant, the conclusion could be drawn that only one or two items are problematic, rather than that an entire group-level class is missing. However, when only one item is problematic, it is more likely that either the BVR-group and BVR-pair are both significant or only the BVR-pair is significant. If there is a missing class it is more likely that either both or only the BVR-group is significant. So, when only one of the two measures shows residual dependence, this could be indicative of what the cause of the problem is. Of course, the wording here is deliberate in that one is more likely than the other, but not necessarily always the case.

Determining the Misspecified Level

Given the mutual influence of the lower and higher level classes, class separation, and sample size, the BVR-group and BVR-pair residuals might also indicate group-level misfit, when the true problem is too few lower level classes. Table 8 gives the values for the regular BVR and the BVR-group residuals when the population consists of three lower level classes and only two are present in the estimation model. Note that the last column for the BVR values depicts the proportion of replications where one third of the BVR values are significant rather than half, thus 5 out of 15 or 15 out of 45 item pairs showing residual covariance.

It is clear that the BVR detects residual dependence between indicator items as soon as the information on the lower level classes is sufficient, either by having a large enough sample size, or by having well-defined and separated classes. Unfortunately the lower level residual dependence is also detected by the higher level residuals, due to the way in which they are obtained. Ideally the latter would not occur and misfit would solely be detected on the lower level.

However, as noted by Lukočienė et al. (2010) the most fruitful strategy in fitting multilevel LC models is assuring good fit of the lower level before making adjustments to the higher level. This is also in line with studies concerning per-level fit in multilevel analysis (see, e.g., Yuan & Bentler, 2007), where misspecification on the higher level does not systematically affect the lower level fit when the levels are considered separately. Therefore, the BVR-group and regular BVR values are contrasted for conditions with a missing higher level class to those with a missing lower level class in Table 8. In doing so it becomes clear that, although the BVR-group does report misfit when the source of that misfit originates on the lower level, the reverse does not occur. That is, the regular BVR values are very close to nominal alpha when the misfit originates on the higher level (see Table B5 for the exact values), still allowing the location of the misfit to be identifiable. Furthermore, the average proportion of significant BVR values over all replications (not reported) is similarly close to 0.05, verifying that significant values are solely due to Type I errors.

CONCLUSION

Inspecting the properties of the two recently developed local fit statistics BVR-group and BVR-pair shows that they work as intended in detecting different types of misfit that cause residual dependence in a multilevel LC model. They allow the level of misfit to be determined, are generally capable of identifying the problematic items, and in combination with global fit statistics and the regular bivariate residual for the lower level allow comprehensive testing and inspection of the main assumptions and substantive goals of the model.

Nonetheless, there are several issues that should be noted. First, in situations where the measures fail to detect

TABLE 8
Power of the BVR-Group and Lower Level BVR to Detect a Missing Lower Level Class

Sample			*Class Separation*			*BVR-Group*			*BVR*		
N	*Groups*	*Group Size*	*Level 1*	*Level 2*	*G*	*Item 1*	*Minimum 1*	*50%*	*Item 1*	*Minimum 1*	*33%*
500	50	10	L	H	2	0.040	0.246	0.002	0.130	0.908	0.100
500	50	10	L	H	3	0.328	0.970	0.126	0.248	0.998	0.002
500	50	10	H	L	2	0.056	0.440	0.002	0.610	1.000	0.732
1,000	100	10	L	H	3	0.384	0.978	0.530	0.082	0.994	0.198
1,000	100	10	L	L	2	0.046	0.394	0.000	0.414	1.000	0.262
1,000	100	10	H	H	3	0.562	1.000	0.906	0.316	1.000	0.896
1,000	100	10	H	H	2	0.080	0.472	0.002	0.844	1.000	0.996
2,500	250	10	L	L	2	0.046	0.276	0.002	0.440	1.000	0.976
2,500	250	10	H	L	3	0.538	1.000	1.000	0.734	1.000	0.996
2,500	50	50	L	H	3	0.544	1.000	1.000	0.152	1.000	0.456
2,500	50	50	H	L	3	0.526	1.000	1.000	0.666	1.000	0.930
2,500	50	50	H	H	2	0.068	0.300	0.006	0.880	1.000	1.000
5,000	100	50	L	H	2	0.046	0.382	0.000	0.970	1.000	1.000
5,000	100	50	H	H	3	0.596	1.000	1.000	0.550	1.000	0.952

the residual dependencies, this could have two different causes. In cases where there is a fairly large sample on both levels, but classes are not clearly separated in terms of conditional probabilities, the residuals themselves lack power. This is not surprising, but should be kept in mind. Both the BVR-group and BVR-pair, analogous to many other fit statistics, merely test for discrepancies between model predicted and sample observed frequencies. In situations where the classes in the population are very hard to distinguish, it is likely that existing dependencies can be modeled with fewer than the true number of classes and parameters. This implies that the problem is limited in that parameter bias and classification errors in these situations will be low. However, when a weakly defined class is highly relevant from a theoretical perspective, a substantive problem will remain. In turn this does mean that the residuals can be used in an exploratory setting to see whether the nested structure needs to be taken into account.

In a few, rather exceptional, situations, class separation is primarily determined by large between-group difference on only one item. The model is then able to sufficiently approach the observed frequencies while misspecified, as it can redistribute the dependence throughout the classes. This implies that not detecting misfit does not guarantee correct parameter estimation, which brings us to an important point that cannot be stressed enough. As with any residual modification index, and despite the residuals working as intended when the data are sufficient for multilevel LC analysis, they should not be used blindly. As already discussed in Nagelkerke et al. (2015), simply trying to reduce the residuals by addressing the area of the model they report to be problematic will lead to capitalization on chance, and will hardly ever result in finding the true population model. The residuals as they are applied here only identify the indicator items that are generally problematic. Because the different areas of the model are intertwined, they cannot point to a given solution, as any conditional dependence might be modeled in many different ways.

For practical use the general conclusion is that the residuals do provide relevant information and can help to improve model fit, but should be used in conjunction with other available measures. Also, it should be kept in mind that these are indeed residuals that detect unmodeled dependence. The briefest summary would be that if significant values are found, something is wrong in terms of capturing dependencies. By using the BVR-group and BVR-pair residuals in conjunction with global fit measures, the regular BVR, and plausible alternative models, it is possible to determine at which hierarchical level misfit occurs, identify which indicator items prove problematic, and in most cases also point at the most parsimonious way to model the uncaptured dependence. If no significant BVR-group and BVR-pair values are found, one can be sure that the nested structure of the data is captured adequately by the model. Yet, although this is a valid conclusion, it does not always imply that the specified model agrees with the true data generating process, meaning that evaluating and comparing alternative models might still be valuable; that is, a better fitting or substantively more sensible solution can still be found when no misfit is detected.

Finally, despite this being an extensive simulation study, several factors, such as different class sizes or the addition of covariates to the model, have not been taken into account here due to the already high computational intensiveness of the current conditions. Furthermore, the relation between the detection of misfit and actual bias in parameter estimation has not been investigated, and is a valuable avenue for future research because currently little is known about the relation between these types of misspecifications and parameter estimation.

Still, for the extensive number of factors that were considered, the overall conclusion is that the measures work as intended, provided that the data are sufficient for multilevel LC analysis to

be viable. Although definitely requiring further research, these results also bolster our expectation that they will work for other analyses dealing with discrete nested data as well.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Fagginger Auer, M. F., Hickendorff, M., Van Putten, C. M., Béguin, A. A., & Heiser, W. J. (2016). Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education*, *29*, 144–159. doi:10.1080/08957347.2016.1138959

Kaplan, D., & Keller, B. (2011). A note on cluster effects in latent class analysis. *Structural Equation Modeling*, *18*, 525–536. doi:10.1080/10705511.2011.607071

Langeheine, R., Pannekoek, J., & Van De Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, *24*, 492–516. doi:10.1177/0049124196024004004

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, *40*, 247–283. doi:10.1111/j.1467-9531.2010.01231.x

Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, A., Pettersen, J., & Bergman, R. (1998). Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems*, *4*, 23–40. doi:10.1016/S0169-7439(98)00065-3

Moineddin, R., Matheson, I., & Flora Glazier, R. H. (2007). A simulation study on the sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, 1–10. doi:10.1186/1471-2288-7-34

Montgomery, D. C. (2012). *Design and analysis of experiments* (8th ed.). Hoboken, NJ: Wiley.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2015). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*. Advance online publication. doi:10.1177/0081175015581379

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*, 45–60. doi:10.1093/pan/mpt014

Oberski, D. L., Van Kollenburg, G. H., & Vermunt, J. K. (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*, 267–279. doi:10.1007/s11634-013-0146-2

Park, J., & Yu, H. T. (2015). The impact of ignoring the level of nesting structure in nonparametric multilevel latent class models. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164415618240

R Development Core Team. (2015). *R v3.3: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roosma, F., Van Oorschot, W., & Gelissen, J. (2015). A just distribution of burdens? Attitudes towards the social distribution of taxes in 26 welfare states *International Journal of Public Opinion Research*. Advance online publication. doi:10.1093/ijpor/edv020

Tomczyk, S., Hanewinkel, R., & Isensee, B. (2015). Multiple substance use patterns in adolescents: A multilevel latent class analysis. *Drug and Alcohol Dependence*, *155*, 208–214. doi:10.1016/j.drugalcdep.2015.07.016

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239. doi:10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*, 33–51. doi:10.1177/0962280207081238

Vermunt, J. K., & Magidson, J. (2013a). *LatentGOLD 5.0*. Belmont, MA: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2013b). *Technical guide for Latent GOLD 5.0: Basic, advanced and syntax*. Belmont, MA: Statistical Innovations.

Von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research*, *2*, 29–48.

Yuan, K. H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*, 53–82. doi:10.1111/j.1467-9531.2007.00182.x