STRUCTURAL EQUATION MODELING
2024,VOL. 31, NO. 2
ISSN: 1070-5511 | E-ISSN: 1532-8007

OPEN ACCESS

Latent Class Analysis with Measurement Invariance Testing: Simulation Study to Compare Overall Likelihood Ratio vs Residual Fit Statistics Based Model Selection

Zsuzsa Bakk

Leiden University

ABSTRACT

A standard assumption of latent class (LC) analysis is conditional independence, that is the items of the LC are independent of the covariates given the LCs. Several approaches have been proposed for identifying violations of this assumption. The recently proposed likelihood ratio approach is compared to residual statistics (bivariate residuals [BVR] and expected parameter change [EPC] statistics) for identifying nonuniform direct effect of covariates on the items of the LC model. The simulation study results show that the likelihood ratio (LR) test correctly identifies direct effects more often than the BVR statistics, showing comparable results to the EPC statistic in many situations- this at the price of having also a higher false positive rate than BVR. A real data example illustrates the use of the three procedures. Overall the combined use of residual statistics and LR testing is recommended for applied research.

KEYWORDS

BVR; EPC interest; latent class analysis; LR based MIMIC test; non uniform direct effects; residual statistics

1. Introduction

Latent class analysis (LCA) is a widely used approach in the social sciences for creating a grouping based on a set of items when the grouping is unknown. The approach is part of a broader family of models, known as latent variable models (LVMs)—a modeling framework where a complex concept is measured by a set of observed item variables. The approach when used with continuous observed and latent variable (LV) is known as factor analysis, or as item response models (IRT) when the items are categorical, and the LV is assumed to be continuous. LCA is the term used for models where both the observed and latent variable is modeled as categorical. The approach is often used with large scale comparative datasets (Da Costa & Dias, 2015; Ruelens & Nicaise, 2020; Van Horn et al., 2008; van Mierlo et al., 2005). In that context the "fairness" of measurement, namely that the LV measures the same underlying concept across cultures is of utmost importance.

In LVM it is common to separate a model into two components, namely the measurement and the structural model. The measurement model refers to the relationship between the LV and the items measuring the LV (for example different symptoms of depression). The structural model on the other hand refers to the antecedents and consequences or with other wording independent variables/covariates and dependent variables related to the LV in a larger model.

Measurement inequivalence, non-invariance or differential item functioning (DIF) are terms often used interchangeably in latent variable modeling literature. They refer to the situation where the measurement model is different at different levels of a (set of) covariate(s).That is when the observed covariate influences items independently of the latent variable or in interaction with the latent variable. A well-known example is when an item is easier for male/female respondents (for example a sensitive question about abortion rights can have a different acceptance rate per gender).

The standard modeling approach in LCA literature is to first establish the measurement model that is the class enumeration without covariates (Nylund-Gibson et al., 2019; Nylund-Gibson & Masyn, 2016; Vermunt, 2010). After class enumeration the structural model can be built using either a one-step or a stepwise estimation approach (Asparouhov & Muthén, 2014; Vermunt, 2010). Ideally the first step of building the structural model would be to check for the presence of direct effects (DEs) of covariates (Janssen et al., 2019) and model the direct effects accordingly (Vermunt & Magidson, 2021). Yet in practice this intermediary step of establishing measurement invariance is often ignored (D'Urso et al., 2022; Masyn, 2017).

Kankaras and Moors (2011) and Eid et al. (2003) are among the first authors to discuss the topic of measurement invariance in detail in the context of multigroup comparisons in LC models. The authors focus on establishing measurement equivalence across countries, following procedures for multigroup analysis. The procedure proposed by Eid et al. (2003) uses Likelihood ratio tests to compare multigroup models with different levels of equality restrictions. The least restrictive model assumes only the number of latent classes to be equal, while the most restrictive model assumes the conditional response probabilities on all items to be equal in all classes (full measurement invariance) or at least in some of the classes (partial measurement equivalence). A similar approach is also taken by Kankaras and Moors (2011). Readers interested in measurement invariance testing in the context of multigroup analysis can consult a very useful example comparing solidarity items across European countries (Kankaraš & Moors, 2009). A limitation of these approaches is that they are specifically developed for detecting measurement invariance in a multigroup context, so not appropriate with continuous covariates. Nevertheless ignoring direct effects of non-categorical covariates can bias the parameters of interest of the structural model (Asparouhov & Muthèn, 2014; Di Mari & Bakk, 2018; Janssen et al., 2019).

There are two main approaches that can detect direct effects in situations where the covariate is not (exclusively) categorical, namely the MIMIC approach (Masyn, 2017) and the approaches that focus on finding misfit using residual statistics (Oberski, 2014) that I present below.

The approach recently proposed by Masyn (2017) follows a step by step procedure of comparing models using likelihood ratio (LR) tests, motivated by the MIMIC literature. The approach starts by comparing a model with no direct effect (DE) of the covariate on any of the items to a model including all possible DE on all items. If the later model fits better based on the LR test, the approach proceeds to a stepwise selection procedure to identify the best fitting intermediary model. This selection is based on fitting a set of models with the goal of finding the model that includes as many direct effects as necessary (using the LR test as selection criteria) while maintaining parsimony. An applied example is available in Tsaousis et al. (2020). While the approach follows a rigorous step by step procedure of model selection, until now it has not been tested in a simulation study to investigate it's power to detect direct effects correctly.

The approaches for detecting local misfit are also mostly borrowed from structural equation modeling and IRT literature (Glas, 1999). Usually they are used after an overall fit statistic (for example the χ^2 test of overall model fit) shows lack of fit. Using residual statistics it is possible to identify where the misfit lies by examining the residual association among all observed variables while controlling for part of the variance that is explained by the model. Once the relevant residual associations are identified they can be added to the model, and overall fit statistics can be used to check the improvement in model fit. There are two main residual statistics used in LCA, namely expected parameter change (EPC) and bivariate residuals (BVR) (Oberski et al., 2013, 2015; Vermunt & Magidson, 2016). A key difference between the two are that while the EPC score statistics follow a chi squared distribution with defined degrees of freedom, the BVRs do not.

The EPC statistics was first introduced in the context of LCA as a test to identify local residual association between items of the LC model (Oberski et al., 2013), and later extended for testing for measurement invariance (Oberski et al., 2015). The simulation experiment with regard to local misfit (Oberski et al., 2013) concluded that the rejection rate (false positive rate) in conditions where the measurement model was stronger tended to be somewhat higher than the nominal 5% level.[1] At the same time the residual statistics had low power to detect (even simple) misfit when the association was small to medium, especially so when the measurement model was strong while it had better power when the missfit was stronger. The lower power to detect smaller missfits with stronger measurement models vs weaker measurement models could be caused by picking up more "noise" with weaker measurement models. The simulation study that accompanied the introduction of EPC as a sensitivity test for identifying measurement invariance is more limited, most importantly the strength of the measurement model was not varied. Yet the simulations study showed that when the measurement model is strong enough and violation of uniform direct effect is medium EPC estimates bias correctly, while with a larger bias EPC tends to overestimate the bias (Oberski et al., 2015). An important limitation of the simulation experiment by Oberski et al. (2015) is that it does not include non-uniform direct effects, that are more difficult to identify. In both simulation experiments the EPC statistics outperformed the BVR statistic, yet I will still include the later in the simulation experiment due to it's simplicity and popularity.

Given the complexity associated with testing stepwise model selection approaches extensive simulation studies did not test the approach proposed by Masyn (2017). In this paper I aim to fill this gap, even if on small scale. I compare the first few steps of the LR based MIMIC approach to residual statistics based identification of covariate direct effects. I restrict myself to the identification of nonuniform or class specific direct effects, this form representing the more complex form of DE than the uniform or class independent direct effect. The rational for focusing exclusively on nonuniform or cluster specific DE is that this form of DE is the most complex, so the hardest to identify. Furthermore, the performance of none of the three approaches has been tested in this more complex setup yet.

The manuscript proceeds as follows: I first introduce the LC model with focus on the conditional independence assumption, and present model definitions that allow for different types of covariate direct effects on items. I than introduce the residual statistic and LR based approaches for testing for the presence of DEs. I compare the different approaches (LR, BVR and two implementations of EPC) via an extensive simulation study and show how the approaches

[1] The rejection rate was defined as the percentage of samples out of 200 replications per condition where the null hypothesis of no effect was rejected at the α level of 5% in conditions where the null hypothesis holds. The strength of the measurement model was manipulated by two sets of loadings (.5 and .8) and four levels of sample size. In all but the lowest sample size condition in the .8 loading condition, the rejection rate was somewhat higher than the nominal rate.

can be used in practice via a real data application that looks on gender roles.

2. The Latent Class Model

2.1. The Simple LC Measurement Model

Consider the vector of responses $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})$, where Y_{ik} denotes the response of individual i on one of the K categorical item variables, with $1 \leq k \leq K$ and $1 \leq i \leq N$. Latent class (LC) analysis assumes that respondents belong to one of the T categories ("latent classes") of an underlying categorical latent variable X which affects the responses (Goodman, 1974; Hagenaars, 1990; A. L. McCutcheon, 1987). The measurement model for $\mathbf{Y}_i$ can then be written as

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

where $p(X = t)$ is the (unconditional) probability of belonging to latent class t, and $p(\mathbf{Y}_i | X = t)$ the class-specific probability of a pattern of responses to the items. For the measurement model, I make the "local independence" assumption that the K item variables are independent within the latent classes, leading to a model of the form:

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

I refer to this as the basic latent class model. The number of classes T is selected by comparing the goodness of fit of models with different values of T using model selection tools such as the AIC and BIC statistics. The entropy of the model (see e.g. Magidson, 1981) that indicates how well the class membership can be predicted by the observed variables can be used as an additional tool to evaluate the LC solution. In most instances, the model is parameterized by using a set of logistic regression equations. Extending the model by a set of covariates $\mathbf{Z}$ affecting class membership leads to a model of a form:

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

The model is graphically represented on the left hand side of Figure 1. For simplicity of notation, I will refer to a single covariate Z, but the ideas presented generalize to the case of multiple covariates directly.

Usually the conditional class membership probabilities $P(X|Z)$ are parameterized using a multinomial logistic regression model of the form:

$$P(X = t | Z = z_i) = \frac{\exp(\alpha_t + \beta Z_i)}{1 + \sum_{t=2}^{T} \exp(\alpha_t + \beta Z_i)}. \tag{4}$$

Alternatively the model can also be parametrized using a log-linear formulation (Kankaras & Moors, 2011; Magidson & Vermunt, 2001; A. McCutcheon, 2002), a parametrization more common in the multigroup setting, that I will not discuss here in details.

The conditional item response probabilities can also be formulated using logit parametrization:

$$P(Y = t | X = t_i) = \frac{\exp(\alpha_t + \beta X_t)}{1 + \sum_{t=2}^{T} \exp(\alpha_t + \beta X_t)}. \tag{5}$$

The model defined in Equation (3) (and parametrized with logistic regression type of parametrization in Equations (4) and (5)) assumes that given the LC variable X there is no direct relationship between Z and the vector of items $\mathbf{Y}$—a fairly common assumption in LV modeling known as measurement invariance.

This assumption can be relaxed by allowing a direct effect between the two sets of variables:

$$p(\mathbf{Y}_i | Z_i) = \sum_{t=1}^{T} p(X = t | Z_i) \prod_{k=1}^{K} p(Y_{ik} | X = t, Z_i). \tag{6}$$

The direct effect (DE) can be uniform, that is constant across the LCs'. Using multinomial logistic regression, this can be parameterized for a specific Y as:

$$P(Y = y | Z = z_i) = \frac{\exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i)}{1 + \sum_{t=2}^{T} \exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i)}. \tag{7}$$

This uniform direct effect is presented in the middle graph of Figure 1, where the uniform effect is represented by the extra direct effect from Z on Y_1. Alternatively, the more complex nonuniform DE, or class dependent DE, can be parameterized as an additional interaction effect between X and Z:

$$P(Y = y | Z = z_i) = \frac{\exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i + \beta \mathbf{X_t} * Z_i)}{1 + \sum_{t=2}^{T} \exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i + \beta \mathbf{X_t} * Z_i)}. \tag{8}$$

using this interaction effect parameterization two tests quantifying the effect of Z on Y are performed: the test of uniform DE (with df = 1), and the interaction effect (with df = T − 1).

Equation (8) can also be equivalently parametrized as follows:

$$P(Y = y | Z = z_i) = \frac{\exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i | X)}{1 + \sum_{t=2}^{T} \exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i | X)}. \tag{9}$$

using this later conditional parametrization, a single test with df = T is performed to test the nonuniform DE of Z on a specific item Y_k.

This later model (Equations (8) and (9)) corresponds to the situation where the effect of Z is different in every class X, and represents the model of interest for the current study. This situation is graphically represented in the rightmost graph in Figure 1. Over and above of the direct effect of Z on Y_1 (middle graph), there is also an interaction effect between Z and X also known as a class specific or class dependent effect of Z on Y_k.

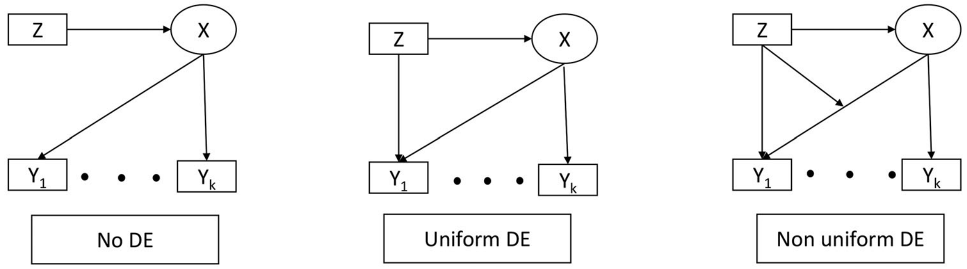

Figure 1. The LC model with covariates without DE and with uniform and nonuniform DE.

3. Identifying Direct Effects in LC Models

3.1. Residual Statistics

The overall fit of the LC model is evaluated based on statistics derived from the overall χ^2 test of the model:

$$\chi^2 = \sum_{r=1,2,3\ldots R} \frac{(n_r - \hat{\mu}_r)^2}{\hat{\mu}_r} \tag{10}$$

Where R is the number of unique response patterns in the dataset, n_r represents the number of observations for a unique response pattern r, and $\hat{\mu}_r = NP_r(Y_r|\theta = \hat{\theta})$ is the model-based expectation of n_r evaluated at the maximum likelihood solution of the model parameters, $\hat{\theta}$. When the model holds the statistic converges to a χ^2 distribution with the degrees of freedom given by the difference between the unique response patterns and model parameters. Measures like AIC, BIC are all based on this statistic. The problem with this overall fit measure is, that while it identifies lack of fit, the cause of misfit is unknown. Residual statistics can be used to inform about possible causes of misfit. In fact one can think of residual statistics as a type of Lagrange multiplier or Score test. There are three main type of statistics for detecting model fit, namely, the likelihood-ratio statistics that compare the log-likelihood values of nested models- usually a restricted and unrestricted model (The LR based Masyn approach for example), Wald statistics evaluate the decrease of the log-likelihood value at the maximum likelihood estimate if constraints were to be imposed, and Score or Lagrange multiplier statistics estimate the increase of the log-likelihood value if constraints are relaxed in the restricted model (the BVR and EPC statistics fall in this latter category) (Buse, 1982). In other words LR tests compare the fit of two models at their maximum likelihood estimate, Wald tests evaluate the fit of a parameter at it's ML estimate, while Score tests evaluate the possible change in a parameter were to be estimated instead of being set to 0.

The residual statistics look on the residual association between two variables after taking into account the association explained by the model. In LC analysis two type of residual statistics are used, Bivariate residuals (BVRs) and the Expected parameter change (EPC-interest).

3.1.1. Bivariate Residuals

The Bivariate residuals evaluate the residual association between each possible pair of observed variables using a χ^2 test with 1 degree of freedom. The statistics can be formally defined as:

$$BVR_{jj'} = 1/P \sum_{j} \sum_{j'} \frac{(n_{jj'} - En_{jj'})^2}{n_{jj'}} \tag{11}$$

where the $En_{jj'}$ for the covariate- item association is defined based on Equation (3) in the current case. That is based on Equation (3) $En_{jj'}$ is defined in such a way that given the LC variable X there is no association between Z and $\mathbf{Y}$. That is that the expected BVR between Z and any of the $\mathbf{Y_s}$ is 0. The $BVR_{jj'}$ statistic checks the difference between the expected and observed residual cell frequency between all possible items and the covariate Z.

It is important to note that the BVR does not test for the type of misfit, that is BVR cannot differentiate between uniform and nonuniform direct effects. It can only show the existence of residual bivariate association after controlling for the model, and it is up to the researchers to find the right form of the misfit for example by testing models with uniform and nonuniform DE and comparing fit using overall fit statistics like AIC, BIC and Wald test of the parameters of interest.

A problem with this statistic is, that while it assumes a χ^2 distribution, asymptotically this does not hold (Oberski et al., 2013; Vermunt & Magidson, 2016). While a bootstrap based approach was proposed by Oberski et al. (2013) to approximate the distribution of the BVR I do not include that in this simulation study, mainly because it is more time consuming and it is not common practice in applied research, yet I will show its use in the real data application.

3.1.2. Expected Parameter Change (EPC-Interest)

The expected parameter change (EPC) is also a score test (Vermunt & Magidson, 2016) first introduced by Rao (1948) in a general regression framework. In the context of LVM EPC was first introduced by Saris et al. (1987) as a tool of sensitivity analysis of fixed parameter estimates of structural equation models, and extended by Bentler and Chou (1992) who introduced the expected parameter change in a free parameter after freeing a fixed parameter, the EPC-interest. Furthermore the approach was introduced to item response models by Glas (1999). For LCA it was first introduced by Oberski et al. (2013) as a tool to test residual association between the items of the LC, and later extended to the situation of testing for measurement invariance by Oberski et al. (2015), where the authors derive how the

EPC-interest test can be generalized to the situation of categorical variables, where usually sets of parameters relate to particular variables using a set of logistic regression equations (Oberski et al., 2015). Based on Equations (7), (8), and (9) one can see that the test of measurement invariance often takes the form of restricting a set of parameters to 0. In the current case this refers to βZ_i for uniform DE, and $\beta X_t * Z$ for the nonuniform DE using the interaction based parametrization of Equation (8). For the conditional parametrization (Equation (9)) this refers to testing whether the set of logit parameters $\beta Z|X$ equal 0. For notational convenience let us consider a restriction on a vector of such logit coefficients as $\boldsymbol{\psi} = \mathbf{0}$. In a general form, the EPC-interest can then be formulated as:

$$EPC - interest = \mathbf{P}\left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}'}\right)(\boldsymbol{\psi} - \boldsymbol{\psi}') \tag{12}$$

where $\mathbf{P}$ is a matrix selecting the parameters of interest and $\boldsymbol{\theta}$ is the vector of free model parameters. As such EPC-interest can be seen as a linear approximation of the relationship between the free and fixed parameters of interest (Oberski et al., 2015), where the fixed parameters are the parameters measuring DE set to 0. The EPC-interest shows how much these parameters would change if they were freely estimated. Details of the derivation for EPC-interest are available in Oberski et al. (2015). In the rest of the article I will refer to the EPC-interest test based on Equation (8) that uses the interaction formulation as EPC.I, and for the conditional formulation based on Equation (9) as EPC.C.

The main difference between the EPC statistics and BVR is that the EPC statistics take into account the dependencies between the expected cell frequencies, thus using the correct asymptotic variance estimator, while the BVR does not (Oberski et al., 2013). Furthermore the EPC statistics follow asymptotically a χ^2 distribution, while the BVR does not. Another important difference is that while the BVR only tests overall residual misfit, the EPC interest is a more targeted test, it can separately test for the presence of uniform and nonuniform DE.

3.2. Likelihood Ratio Based Stepwise Multiple Item Multiple Cause (MIMIC) Modeling

The LR based MIMIC approach first proposed by Masyn (2017) is a multistage comparative approach, that I present shortly (full details of all steps are available in Masyn, 2017). The approach uses a logistic regression parametrization-similarly to the models defined above. It uses a stepwise procedure where models with different levels of complexity are compared with the goal to find the least restrictive well-fitting model.

The approach has a maximum of seven steps, depending on the results of each of the steps. For all but the sixth steps, a conditional LR test is used. The procedure continuous until the most restrictive model is found that does not perform significantly worse than the least restrictive model.

The main steps are presented below:

- Step 0: Class enumeration without covariates- standard best practice also with other approaches
- Step 1: Comparing model M1.0 with no DE to model M1.1 with all nonuniform DEs of the covariate on all of the items.
- Step 2: If in step 1 M1.1 fits better than M1.0 an item by item level testing of nonuniform DEs is performed.
- Step 3: Comparing the model with all significant nonuniform DEs (M3) with model M1.0 and M1.1. M3 should fit better than M1.0, but not worse than M1.1.
- Step 4: item level test of uniform DE for all items that proved to have a significant nonuniform DE in Step 2, and thus are included in M3. The goal is to find the most parsimonious way to model DEs. If the model with uniform DE does not fit significantly worse than the model with nonuniform DE the simpler model is taken forward for that item.
- Step 5: Compare the simplified model with uniform DEs (M5) for items for which this proved a good fit in step 4 with M3. Check if fit of M5 is not worse than that of M3.
- Step 6: Evaluate practical and substantive impact of DE's.
- Step 7: Evaluate the association between LC membership and the covariate controlling for the DEs.

In the current study I focus primarily on the identification of nonuniform DEs using the two types of residual statistics (BVR and EPC) and the likelihood ratio based MIMIC approach. As such I restrict myself to step 1 and 2 in the LR approach, focusing on the Research Question if a nonuniform DE is present is the LR approach identifying it's presence. I do not include steps 3–7 in the simulation experiment because of the increase in complexity in identifying the best fitting model for all the possible models to be evaluated. Since thus far the LR approach was not tested via a simulation experiment, this experiment represents an important first step in evaluating parts of the model selection process.

4. Simulation Study

4.1. Simulation Experiment Setup

I set up a simulation study to compare the residual statistics (BVR and EPC-interest) to the LR based MIMIC model selection procedure under situations where nonuniform DE is present. For estimating false positive rates I also added a condition where no DE is present as baseline. The research question driving the study is how well do the three different statistics do in identifying a nonuniform DE when that is present, and how well they do in rejecting models with DE when the data comes from a model with no DE.

I generated data from a LC model with six binary items measuring a three class model. The three classes are defined as follows: one class has high probability for a positive answer on all items, opposing a class with low probability on all items, and a medium class with high probability on the first three, and low on the last three items. The conditional item response probabilities were set to. 70, .80, and .90 for the positive answer corresponding to a low, medium and high class separation. This setup is based on the simulation study setup by Vermunt (2010), Bakk et al. (2013), and Di Mari and Bakk (2018), where the authors use a similar setup to obtain models corresponding to a low, medium and high class separation. In those simulation studies the class separation is shown to be an important factor for the bias in the effect of the covariate on the latent classes. I also manipulated the sample size (n = 250, 500, 1,000, 2,000), that is known to contribute to the quality of the measurement model together with the strength of the association between items and the LV (Bakk et al., 2013; Vermunt, 2010).

I generated data from a model with a single covariate Z. The effect of Z on X is defined using a logistic regression parametrization with dummy coding with the first class (the class with a high probability on all items) as reference category. The effect of Z on class 2 is set to $\beta Z_2 = -1$, and on class 3 as $\beta Z_3 = 1$. These values correspond to an odds ratio of 0.37 and 2.72, respectively, for the odds of being in class 2 and 3 with respect to class 1.

In all conditions, I set the class specific nonuniform direct effects on two items (Y_1 and Y_6), having three levels of severity of direct effect: weak, medium, and large. For generating the direct effect, I used the parametrization defined in Equation (9). For the weak DE condition, the nonuniform DE on Y_1 was set to the logit coefficient of.25 ($\beta Z_i|X_1 = \beta Z_i|X_3 = .25$) in class 1 and 3, and to no effect in class 2 ($\beta Z_i|X_2 = 0$). While the effect of Z on Y_6 was set to .25 in class 1 ($\beta Z_i|X_1 = 0.25$), and $-.25$ in class 3 ($\beta Z_i|X_3 = -0.25$), and no effect on class 2 ($\beta Z_i|X_2 = 0$).[2] In the medium and large DE effect size conditions the same pattern of effect sizes were set on Y_1 and Y_6 in class 1 and 3—using the conditional effect formulation of Equation (9), using a logit coefficient (in parenthesis odds ratio) of 0.50 (1.65) and −0.50 (0.61) in the medium, and 0.75 (2.12)/−0.75 (0.47) in the large DE condition.

The class sizes were also manipulated having a condition with equal class sizes, and a smaller simulation experiment with unequal class sizes. It should be mentioned that in the equal class size condition I generated data with LC sizes being set to 0.33 in class 1 and 2, and 0.34 in class 3—as such leading to approximately equal class sizes in the different replications. In the unequal class sizes condition I generated data from class sizes being set to: .20, .20, and .60, respectively.

I furthermore generated data also from a model with no DE. In all DE conditions (no, weak, medium and large) data was generated under the 12 different conditions of combination of sample size and conditional response probabilities for the equal class sizes condition. For the unequal class sizes condition data was generated only under the large DE condition.

Data was generated and LC models run in Latent Gold 6.0 (Vermunt & Magidson, 2013), with using R as a wrapper for the simulation study and for analyzing and organizing output. In total 100 replications were run per condition.

I analyzed the data with the three different approaches for testing for a nonuniform DE. Using the LR test I first fitted model M1.0 with no DE and Model M1.1 with all nonuniform DE. Following if the all DE model fitted better I continued to step 2, namely, comparing models item by item with and without a non-uniform DE. As such in step 2 in total 12 models are fitted (a model with and without DE for each of the six items). I stopped the procedure at step 2 for manageability of the experiment.

Using BVR and EPC I estimated only one model, namely the model described in Equation (3) with an effect of the covariate on the LC variable only, without modeling DEs. The residual statistics are used to evaluate the misfit of this model. I analyze the BVRs between the covariate and each of the items. I consider a misfit present if the BVR is larger than the critical value for a χ^2 distribution with df = 1. Using the EPC statistics the direct effects can be formulated either using Equation (8) (EPC.I) or (9) (EPC.C). For completeness I tested both model formulations. As such I evaluated the EPC statistic for both the parameters of the uniform (βZ) and nonuniform ($\beta Z * X$) DE for EPC.I, and in a separate run also the $\beta Z|X$ parameters as formulated in EPC.C. This means that in the interaction effect formulation there are two sets of parameters to be evaluated, namely the βZ with 1 degree of freedom and the $\beta Z * X$ with 2 degrees of freedom, while in the conditional formulation for the effect of $\beta Z|X$ there is a single test with 3 degrees of freedom per item.

4.2. Simulation Results

In all simulation conditions I assumed the number of latent classes to be known, and estimated all models with three classes, as in the population. While in most real data setting the number of classes are unknown to keep the focus of the experiment on the identification of DEs, I did not manipulate class selection. Before discussing the simulation results in detail first I reflect on the overall LR test comparing M1.0 to M1.1. This test in all datasets generated with DE rejected the hypothesis that M1.0 with no DE fits the data better than the all-DE M1.1 model. Given that there is no variability across conditions these results are not presented in more detail.

In Table 1, the results averaged across the four sample sizes per level of class separation and direct effect are presented. The columns that have T after the dot refer to the results averaged over Y_1 and Y_6 for the different statistics, while F refers to results averaged over items Y_2 to Y_5, without DE. The BVR statistics have a very low true positive rate, but also very low false positive, as such they did not prove to be discriminative in identifying the DEs.

The results with the EPC statistics are reported separately for the two formulations, namely the interaction effect formulation (Equation (8)) in columns EPC.I, and with the

[2] Note that a logit coefficient of .25 corresponds to an odds ratio of 1.28, and a logit coefficient of −0.25 of 0.78, respectively.

Table 1. Percentage of identified direct effects with the different statistics averaged across sample sizes for the different levels of direct effects and class separation conditions.

Class Separation	BVR.T	BVR.F	EPC.C.T	EPC.C.F	EPC.I.T	EPC.I.F	LR.T	LR.F
				Large DE equal class sizes				
High	0.18	0.00	0.98	0.16	0.81	0.12	0.97	0.18
Medium	0.22	0.00	0.86	0.20	0.68	0.19	0.83	0.29
Low	0.03	0.00	0.41	0.15	0.33	0.15	0.52	0.34
				Medium DE equal class sizes				
High	0.02	0.00	0.91	0.12	0.59	0.10	0.87	0.16
Medium	0.06	0.00	0.78	0.13	0.58	0.13	0.76	0.32
Low	0.01	0.00	0.55	0.16	0.37	0.15	0.59	0.40
				Small DE equal class sizes				
High	0.00	0.00	0.41	0.10	0.24	0.07	0.63	0.16
Medium	0.00	0.00	0.44	0.12	0.28	0.09	0.60	0.29
Low	0.00	0.00	0.37	0.15	0.25	0.12	0.42	0.27
				Large DE unequal class sizes				
High	0.86	0.00	0.97	0.16	0.96	0.20	1.00	0.13
Medium	0.19	0.00	0.76	0.39	0.48	0.32	0.76	0.35
Low	0.02	0.01	0.38	0.25	0.29	0.20	0.46	0.34

Note: T: refers to items 1 and 6 with DE, thus the % of samples where the true DE was found on average, and F to items 2 to 5 with no DE, % of samples with false positive

Table 2. The EPC$_{interest}$ statistics in the large DE equal class sizes condition broken down over conditions of class separation and sample size.

Class sep.	Sample	EPC.C.T	EPC.C.F	EPC.I.T.Y_1	EPC.I.T. Y_1^2	EPC.I.T. Y_6^2	EPC.I.F.
High	2,000	1.00	0.23	0.97	1.00	1.00	0.16
High	1,000	1.00	0.15	0.65	0.99	1.00	0.13
High	500	1.00	0.15	0.40	0.94	1.00	0.11
High	250	0.92	0.11	0.18	0.65	0.96	0.09
Medium	2,000	1.00	0.32	0.99	0.82	1.00	0.30
Medium	1,000	0.99	0.21	0.83	0.66	0.99	0.22
Medium	500	0.90	0.16	0.54	0.42	0.93	0.13
Medium	250	0.54	0.12	0.24	0.20	0.57	0.12
Low	2,000	0.66	0.21	0.52	0.23	0.95	0.22
Low	1,000	0.51	0.16	0.31	0.14	0.74	0.17
Low	500	0.32	0.12	0.21	0.15	0.32	0.10
Low	250	0.15	0.10	0.13	0.06	0.19	0.09

conditional formulation, in columns EPC.C (Equation (9)). The interaction formulation is doing worse on average than the conditional formulation in identifying the DEs. In conditions where the DE effect is large or medium and the class separation is high or medium the EPC.C is close to the nominal 95% rate, while the EPC.I is performing somewhat lower. However in conditions where the measurement model is weak, and the DE is smaller, the EPC has a low true positive rate, while in this conditions the false positive rate increases. These results are also comparable to those of Oberski et al. (2013). The LR test shows a very similar tendency to EPC, namely that it's capacity to correctly identify a DE strongly depends on the strength of the DE and of the measurement model, and the false positive rates are in all conditions above the nominal 5% rate, while the true positive rates are above the critical 95% rate only in conditions with large effect sizes and good measurement model. In the unequal class size conditions the three statistics show comparable tendencies, but all statistics have a somewhat worse performance, namely, the true positive rate is lower, while the false positive rate is higher than in the equal class size conditions.

In Table 2, I zoom in on the condition with large DE and equal class sizes to show the results separately at each level of class separation and sample size combination for the two types of EPC statistics, with conditional (EPC.C) and interaction (EPC.I) formulation. Using EPC.C the true positive rate is 90% and above up until the medium class separation 500 sample size condition, after which it drastically drops. This results are comparable for the interaction effect for Y_6. Using EPC.I. Y_6 was generated with DE in class 2 and 3, thus the DE should be identified only with the interaction effect. However in Y_1 DE was generated in the reference class 1 and in class 3, so for this item using EPC.I the DE is identified by the two separate tests: the main effect and interaction effect. In columns 5 and 6, it can be clearly seen that using the two separate tests the true positive rate is much lower, than for item 6 (column 7), or the averaged value for EPC.C (column 3). The false positive rate even in the high class separation and sample size conditions is above 15%, much higher than the desired 5% rate for both implementations of the EPC statistic.

In Figure 2, I report the EPC and itemwise LR statistics for the condition with no direct effect: as such no effect should be found. While BVR finds indeed no effects, using the EPC.I statistics the identified direct effects (averaged over all six items) are below 10% in all conditions, while with EPC.C the false positive rate is marginally higher especially in the low sample size and low class separation conditions. Using the LR statistics the false positive rates are higher. The overall LR test, identifies still in a large percentage of the samples support for some DE. Using the item

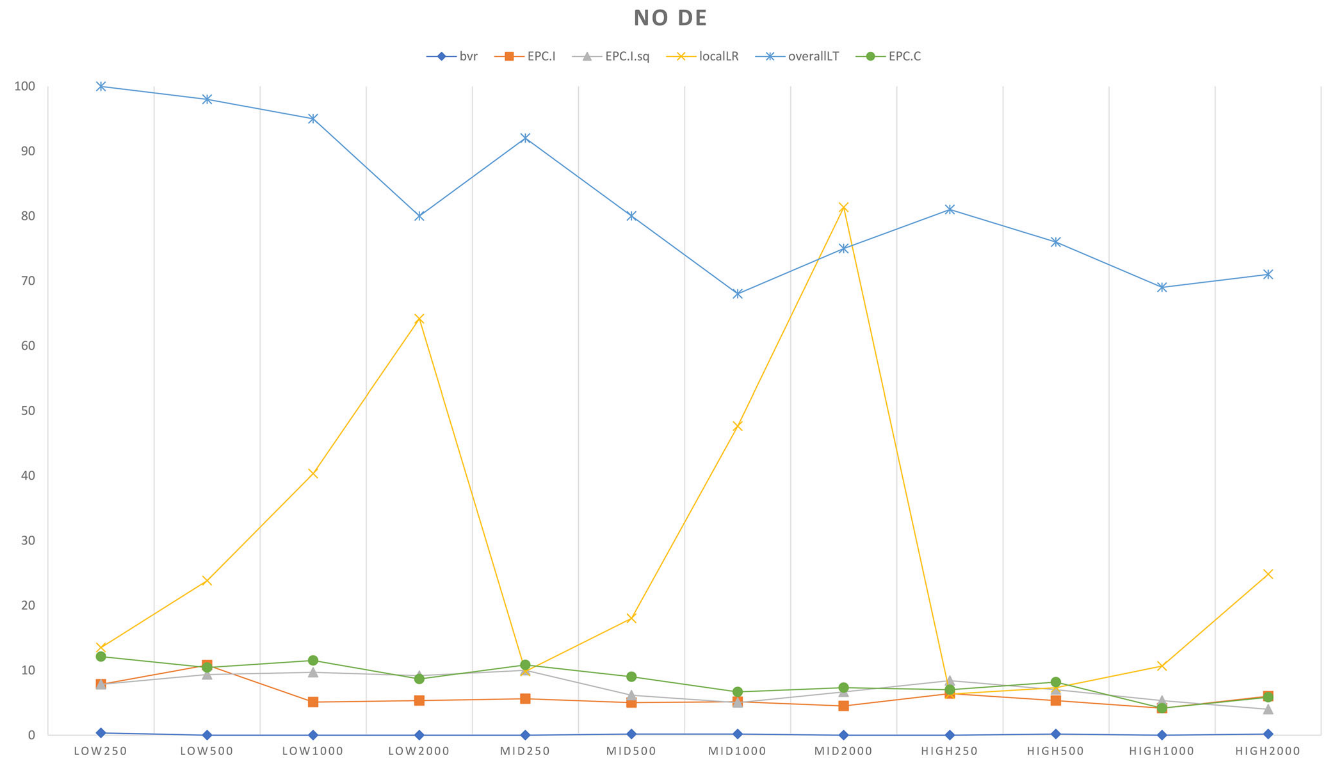

Figure 2. The percentage of samples where DE was identified using the EPC and LR approach for all simulation conditions without DE.

Table 3. Total number of replications out of 100 per combination of sample size by class separation (simulation conditions column represent class separation (low/medium/high) and sample size 250 to 2,000) for each DE condition where all the EPC statistics were estimated for all six items using EPC.I and EPC.C.

	Low DE		Medium DE		Large DE		No DE		Uneq. cl.	
Simulation conditions	EPC.I	EPC.C	EPC.I	EPC.C	EPC.I	EPC.C	EPC.I	EPC.C	EPC.I	EPC.C
Low 250	51	94	68	98	66	100	68	99	67	98
Low 500	70	99	64	97	80	99	95	99	71	97
Low 1,000	79	99	63	100	87	100	99	100	75	98
Low 2,000	94	100	70	100	89	100	91	100	79	100
Medium 250	91	100	77	99	67	99	100	100	69	100
Medium 500	99	100	87	100	87	100	100	100	92	100
Medium 1,000	100	100	97	100	91	100	98	100	88	100
Medium 2,000	100	100	99	100	100	100	100	100	99	100
High 250	98	100	96	100	91	100	100	100	94	100
High 500	100	100	98	100	95	100	100	100	94	100
High 1,000	100	100	99	100	99	100	100	100	99	100
High 2,000	100	100	99	100	100	100	100	100	99	100

Note: Uneq. cl refers to the large DE with unequal class sizes, the rest of the conditions are with equal class sizes.

level LR tests as sample size and separation between classes goes up the amount of noise picked up as direct effect is increasing, reaching up to 80% in the medium class separation 2,000 sample condition, while it only improves in the high class separation conditions.

An issue I observed across the simulation study is that in some more difficult conditions the EPC statistics are not reported for all items, especially for EPC.I. The reason for this is that when the piece of the information matrix related to the constraints is not full rank the statistic cannot be calculated. In Table 3, the number of samples per condition where the EPC.I and EPC.C statistic was available for all items is presented. It should be mentioned that for EPC.C the problem is less relevant because the score test is run on a single test with more degrees of freedom, rather than 2 separate test for the direct and interaction effect, so more information per test is available.

In the Online Appendix, the detailed results broken down over all combinations of simulation conditions are available. These results show that the main simulation conditions, namely class separation, sample size and strength of DE have a very strong effect on the performance of the three statistics. In all conditions BVR is consistently showing the weakest results, but the sensitivity and specificity of EPC-interest and LR test is also below the nominal rates in conditions where any of the 3 simulation conditions is not optimal (lower sample sizes, lower DEs, and weaker measurement models).

5. Real Data Application: Gender Steretypes

To illustrate the use of residual statistics and LR test in a real data setting I chose a dataset that is used in Latent

Table 4. LC model fit for models with one to five classes on the five items of gender role on the Heinen data.

	LL	BIC (LL)	AIC (LL)	AIC3 (LL)	L^2	df
1-Cluster	−3,319.69	6,674.54	6,649.38	6,654.38	708.98	26
2-Cluster	−2,999.15	6,075.67	6,020.30	6,031.30	67.90	20
3-Cluster	−2,972.16	6,063.89	5,978.32	5,995.32	13.92	14
4-Cluster	−2,970.14	6,102.04	5,986.27	6,009.27	9.88	8
5-Cluster	−2,967.58	6,139.14	5,993.17	6,022.17	4.77	2

Table 5. The three class measurement model showing the probability to agree with all items in each class and the class sizes.

	Disagree (.44)	Medium (.44)	Agree (.12)
Item 1	.23	.55	.87
Item 2	.04	.29	.90
Item 3	.11	.77	1.0
Item 4	.01	.05	.48
Item 5	.23	.73	.97

Gold as example data for measurement invariance in a multigroup context. The data was introduced by Heinen (1996) and next to Latent GOLD is also available as example dataset in the R package randomLCA (https://CRAN.R-project.org/package=randomLCA). The data was collected by Felling et al. (1987).

The dataset contains five dichotomous items measuring gender roles. The five items are:

- Q1 Women's liberation sets women against men.
- Q2 It's better for a wife not to have a job because that always poses problems in the household, especially if there are children.
- Q3 The most natural situation occurs when the man is the breadwinner and the woman runs the household and takes care of the children.
- Q4 It isn't really as important for a girl to get a good education as it is for a boy.
- Q5 A woman is better suited to raise small children than a man.

Table 4 compares models with one to five classes. There is agreement among AIC, AIC3, BIC in selecting the model with three classes as best fitting.

The model with three classes fitted the data very well (L^2=13.92, df = 14, $p=.46$, χ^2=14.76, df = 14, $p=.39$ entropy $R^2=0.58$) The three class model as presented in Table 5 is characterized by a class that agrees with all items (12% of respondents), a class that disagrees with all items (44% of respondents), and a middle class that shows a medium rate of agreement on most items (44%). Please note in interpreting the model that the items are negatively worded, thus disagreeing with items means a more progressive view on women rights.

After fitting the simple LC model I inspected the residual statistics for any sign of measurement invariance with regard to gender (see Table 6). I did this by fitting a model with gender as covariate with only indirect effects on items and asking for bivariate residuals and EPC statistics for all possible direct effects, using EPC.I and EPC.C. Using EPC statistics it is possible to directly test whether the DE is uniform or nonuniform, however with BVR it is only possible to test if there is a residual association between two variables- but not the type of association. All residual statistics point toward possible direct effect on item 3 of gender, EPC.I shows a possible significant uniform and nonuniform effect, and EPC.C is also highly significant. Furthermore EPC.I identifies also a possible uniform DE on item 4, and EPC.C also identifies a possible class dependent DE on item 4. The bootstrapped p value of the BVR statistic is also significant for items 3, 4, and 5. However using the rule of thumb value (BVR of 3) the violations on item 4 and 5 would not be considered substantively relevant violations of conditional independence.

Following I fitted a series of models trying to find the most parsimonious best fitting model based on the residual statistics. The EPC and BVR statistics are in agreement with regard to item 3, showing a possible direct effect, there is also some evidence for a possible DE on item 4, but the residual fit statistics show less support for this possible DE. As such I chose to fit 5 models where I added the possible direct effect of gender on item 3 (the item that has most evidence of DIF): namely a model with no DE, with uniform and nonuniform DE. Once the best model is selected on item 3, keeping this constant I added DE on item 4.

The model with the uniform DE on item 3 and no DE on item 4 shows the best fit according to all model fit indices (see Table 7). This model has a weak, but significant direct effect of gender on item 3 (Wald= 9.99, df = 1, $p=.002$) showing that females disagree more strongly with this item than males ($\beta= -1.27$, SE = 0.40).

It should be mentioned that the EPC statistic using the interaction formulation is not available for item 5, while using the conditional formulation it is calculated only for a subset of parameters for item 4 and 5 (df is 2 and 1, respectively).

Next the steps described by Masyn (2017) were deployed to the data. The results of the stepwise LR tests (each model compared to the previous less complex model) are presented in Table 8. Based on the first step (M1) it can be seen that the model with all nonuniform DE fits better than the no DE model. This means that I need to continue with step 2 where I fit a step 3 model (Vermunt (2010)) with nonuniform DE (in the models labeled M2.0.2y_{1to5} : *nonuniform*). I compare this more complex model with the models including no DE on the items (M2.01 y_{1to5} :no). The LR tests suggests in case of item 3 that the model with nonuniform DE fits better than the model without, while for the other items the model with no DE is better.

Following in step 3 (not reported in Table 8) I compare the model with item 3 with nonuniform DE with the No DE and all DE models. The expectation is that the model should fit better than the No DE, while fitting not worse

Table 6. BVR and EPC statistics gender by item.

Item	EPC uniform (df)	EPC nonuniform (df)	EPC conditional (df)	BVR	BVR bootstrap *p*
Item 1	0.99 (1)	3.08 (2)	3.10 (3)	0.52	.35
Item 2	0.30 (1)	3.65 (2)	4.71 (3)	0.07	.59
Item 3	19.06 (1)***	24.41 (2)***	24.97 (3)***	5.67***	.00
Item 4	4.60 (1)*	2.05 (2)	20.43 (2)***	2.98*	.03
Item 5	NA	NA	3.51 (1)	1.24***	.00

Note: For BVR, the bootstrap *p* value is provided. *$p = .05$, **$p = .01$, ***$p = .001$.

Table 7. Model fit indices of models for the three class model with DE of gender on item 3 and 4.

	LL	BIC(LL)	AIC(LL)	AIC3(LL) DE in item 3	Npar	L^2	df	Max. BVR
No DE	−2,952.99	6,039.62	5,943.98	5,962.98	19	83.47	43	5.67
Nonuniform DE	−2,940.96	6,036.66	5,925.93	5,947.93	22	59.42	40	1.67
Uniform DE	−2,942.27	6,025.21	5,924.54	5,944.54	20	62.03	42	1.83
			Modeling uniform DE on item 3 and testing DE for item 4					
Uniform DE	−2,941.01	6,029.73	5,924.02	5,945.02	21	59.51	41	1.76
Nonuniform DE	−2,940.81	6,043.40	5,927.63	5,950.63	23	59.12	39	1.74

Table 8. LR based MIMIC test of the real data application.

Model	LL	df	LR test
M1.0 no de	−2,952.99	43	
M1.1 all de	−2,927.92	28	0.00
step 2			
M2.0.2 y1:no	−2,812.851	7	
M2.0.1 y1:non unif	−2,811.87	10	0.58
M2.0.2 y2:no	−2,583.34	10	
M2.0.1 y2:non unif	−2,583.61	7	0.91
M2.0.2 y3:no	−2,669.83	7	
M2.0.1 y3:non unif	−2,660.88	10	0.00
M2.0.2 y4:no	−2,476.93	10	
M2.0.1 y4:non unif	−2,479.07	7	0.23
M2.0.2 y5:no	−2,713.13	7	
M2.0.1 y5:non unif	−2,711.77	10	0.44
step 4			
M4.0.2 y3:uni de	−2,660.9514	8	0.93

Table 9. The effect of gender on the latent classes and the uniform DE of gender on item 3.

LC	Female	Wald(df)
Disagree	0.84	45.88 (2)
Medium	−0.65	
Agree	−0.20	
Uniform DE		
Item 3 (agree)	0.64	9.99 (1)
Item 3 (disagree)	−0.64	

than the all DE model (Masyn, 2017). When compared to the No DE model one can indeed see that the more complex model fits better. The LL the model with item 3 with nonuniform DE is −2,940,96, with Npar = 22. Compared to the No DE model(Model M1.0, the LR test is 24,06, with df = 3, $p < .001$). The LR test of comparing this model to M1.1 is LR = 26.09, df = 12, $p = .01$. This result is counter intuitive as one would have expected the model to be not significantly worse than the model M1.1. This contradiction is not described in the original article, as such it is an interesting and relevant findings that needs further attention. I continue to step 4 comparing the model with item 3 with nonuniform DE with a simpler model with uniform DE. As the last line of Table 8 shows this comparison shows that the model with uniform DE does not fit worse than the model with nonuniform DE, thus the simpler model is selected, a decision aligned also with the selection using the residual statistics. In step 5 the final model identified in step 4 should be compared to model 3.0—the model with item 3 with nonuniform DE. This means fitting both models using a FIML estimate on approach. As the last comparison shows the simpler model does not fit worse, thus it can be accepted as the final model, and the comparative iterative process can stop here.

Given that in step 2 I identified only 1 non uniform DE in this example the amount of steps required to reach a final model with the LR test approach was relatively limited, yet it still requires estimating in total 15 models (compared to 4 models with the residual statistics approach).

In conclusion the same model was identified as best fitting model using both the residual statistics and the LR based MIMIC model. This model allows a uniform DE on item 3, the most conservative item. The effect of gender on the classes and on item 3 are presented in Table 9. Woman tend to agree more with this item than expected under the conditional independence assumption, warranting the inclusion of a direct effect of gender.

6. Discussion

In this paper I performed a simulation experiment to test the performance of two residual statistics (BVR and EPC test; Oberski et al., 2013) and the Likelihood Ratio based MIMIC procedure (Masyn, 2017) in identifying nonuniform direct effects of covariates on the items of a LC model.

While in other LV modeling frameworks the simultaneous use of residual fit measures and LR based model testing is mainstream in LC analysis these approaches are still relatively new. Furthermore thus far no comparison of the two type of approaches in a simulation setup was available. Given this lack in literature I conducted a simulation experiment focusing on the potential to identify nonuniform DE of the two main type of approaches- nonuniform DE being the most complex type of DE. Using the LR approach after step 2 a large variety of models need to be estimated/compared, as such focusing only on first steps (as did here) is more manageable in a simulation setup and already provides useful information about model characteristics.

The results of the simulation study show that when a nonuniform DE is present the LR based approach and

EPC.C have a higher power than the EPC.I and BVR to identify the DE. On the other hand the false positive rate is also higher with the LR based approach and EPC.C.

Furthermore a downside of the EPC statistic is that in more difficult conditions it can happen that the matrix corresponding to a certain Score test is not full rank, and the EPC statistic cannot be provided. These especially happened in smaller sample size/lower class separation conditions, but also for item 5 in the real data example. This problem occurs with both EPC.C and EPC.I, but it is more common with EPC.I.

In conditions where there is no DE the LR based approach has a higher false positive rate than the residual statistics, picking up more noise as signal from the data. In these conditions the BVR statistics proved to be the most conservative.

Overall based on the results of the simulation study I recommend using both residual statistics and the LR based approach simultaneously to inform model selection. From the residual statistics EPC.C is the most recommended approach to be used for identifying nonuniform DE, and the BVR the least recommended.

Residual statistics are usually used in literature as a tool to identify potential DE, followed by a modeling step, where the identified DEs are modeled and using model fit indices (such AIC, BIC, etc.) the best fitting model is selected. I exemplified this procedure in the real data example, where we can see that while the residual statistics pick up some noise also in item 4, once these effects are added to the model they do not improve model fit, while modeling the DE on item 3 proves to be important.

A limitation of the simulation experiment is that I kept the number of classes fixed, while class selection can also be biased by the presence of nonmodeled direct effects (Nylund-Gibson & Masyn, 2016). Even more pertinent, in the case of distal outcome models, over-extraction of classes is a known problem when the conditional distributional assumptions of the distal outcomes are violated (Bakk & Vermunt, 2016). The current recommendation in literature, because of the possible overextraction problems is to perform class enumeration using only the measurement model, and in a next step add the structural model, keeping the number of classes fixed.

A possible model selection strategy can be inspired by the strategy I implemented in the real data application. In this example I compared a series of models based on model misfit identified by the residual statistics, and chose the best fitting model between no DE, uniform and nonuniform DE on items 3 and 4 looking also on overall model fit. In total this meant comparing 5 models. Using the LR approach a series of 15 models were fitted, leading to the same conclusion in this case, namely that the model with uniform DE on item 3 is the best fit. Further research can look into how best to combine the LR based approach with using residual statistics to identify a well-fitting model in a relatively quicker manner than only with the LR based approach.

A counter intuitive finding in the real data example using the LR based MIMIC approach was that in step 3 of the model selection the model with nonuniform DE on the single item for which nonuniform DE was identified in step 2 fitted worse than the all DE M1.1 model. This means that the selection approach might be very sensitive to small misspecifications. Further research should look into how general the counterintuitive result found in the real data example with regard to comparing the models in step 3 (and possibly step 5) of the LR test is.

Further research can also investigate the problem of multiple testing with the LR test- the multiple testing without alpha correction can also be a reason for the higher acceptance rate than the nominal rate with this approach for the items that have no DE. A further downside of this approach is that it proposes a univariate approach for each Z of interest. Looking into how best the approach- possibly combined with residual statistics can be extended to a setting of simultaneous testing of DEs caused by multiple Z variables and possibly their interaction can be a lucrative next research step.

At the same time the residual statistics should also be used carefully. The simulation study clearly showed that the BVR statistics did not have a high power to identify nonuniform DE. Using a bootstrap SE estimation of the statistics might be better- for some results, see Oberski et al. (2013). I did not use the bootstrap approach in the simulation experiment to stay closer to applied settings, and even more importantly for keeping estimation time of the simulation study manageable. The EPC statistics on the other hand might not always be available for all the parameters that are set to 0 as such it's use being limited.

While with EPC it is possible to not only test whether a DE is present but also if the effect size of the DE is correctly estimated, that falls outside of the scope of the current study, that focused only on acceptance and rejection rate of overall significance test. Follow up research can investigate the bias in effect size estimates.

Finally it should be mentioned that the results of the current study are well aligned with what is known about the performance of EPC and BVR statistics for uniform DE (Oberski et al., 2013), namely, that an overestimation of false positive rates is common (in almost no conditions was this rate close to the nominal 5%), and that the performance of the residual statistics strongly depends not only on the effect size of the DE, but also on the quality of the measurement model. As such all the measures should be treated with some precautions and in applied research it is recommended to use the residual statistics combined with a stepwise model selection based on well-known overall fit measures like AIC and BIC. This procedure is exemplified in the real data application. All in all none of the tested approaches has an acceptable true positive and false positive rate across conditions, as such especially when the measurement model is weak, there is still a large degree of subjectivity in model selection and combination of various tools is recommended.

References

Asparouhov, T., & Muthèn, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 329–341. https://doi.org/10.1080/10705511.2014.915181

Bakk, Z., Tekle, F. T., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*, 272–311. https://doi.org/10.1177/0081175012470644

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 20–31. https://doi.org/10.1080/10705511.2014.955104

Bentler, P. M., & Chou, C.-P. (1992). Some new covariance structure model improvement statistics. *Sociological Methods & Research*, *21*, 259–282. https://doi.org/10.1177/0049124192021002006

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*, 153–157. http://www.jstor.org/stable/2683166

Da Costa, L. P., & Dias, J. G. (2015). What do Europeans believe to be the causes of poverty? a multilevel analysis of heterogeneity within and between countries. *Social Indicators Research*, *122*, 1–20. https://doi.org/10.1007/s11205-014-0672-0

Di Mari, R., & Bakk, Z. (2018). Mostly harmless direct effects: A comparison of different latent Markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 467–483. https://doi.org/10.1080/10705511.2017.1387860

D'Urso, E., D., M, E., V. A, M. M. A. L., Nuijten, M. B., De Roover, K., Wicherts, J. M. (2022). *The dire disregard of measurement invariance testing in psychological science*. https://osf.io/j72t4/?view_only=83cf802a792543419841d36cc885c702

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology*, *34*, 195–210. https://doi.org/10.1177/0022022102250427

Felling, J., Peters, A., & Schreuder, O. (1987). *Religion in Dutch society 85: Documentation of a national survey on religious and secular attitudes in 1985*. Steinmetz Archive.

Glas, C. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika*, *64*, 273–294. https://doi.org/10.1007/BF02294296

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology*, *79*, 1179–1259. https://doi.org/10.1086/225676

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

Heinen, A. (1996). *Latent class and discrete latent trait models: Similarities and differences (No. 6)*. Sage.

Janssen, J. H. M., van Laar, S., de Rooij, M. J., Kuha, J., & Bakk, Z. (2019). The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 280–290. https://doi.org/10.1080/10705511.2018.1541745

Kankaraš, M., & Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe: Insights from a multiple-group latent-class factor approach. *International Sociology*, *24*, 557–579. https://doi.org/10.1177/0268580909334502

Kankaras, M., & Moors, G. (2011). Measurement equivalence and extreme response bias in the comparison of attitudes across Europe: A multigroup latent-class factor approach. *Methodology*, *7*, 68–80. https://doi.org/10.1027/1614-2241/a000024

Magidson, J. (1981). Qualitative variance, entropy, and correlation ratios for nominal dependent variables. *Social Science Research*, *10*, 177–194. https://doi.org/10.1016/0049-089X(81)90003-X

Magidson, J., & Vermunt, J. (2001). Latent class factor and cluster models: Bi-plots and related graphical displays. *Sociological Methodology*, *31*, 223–264. https://doi.org/10.1111/0081-1750.00096

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 180–197. https://doi.org/10.1080/10705511.2016.1254049

McCutcheon, A. L. (2002). Basic concepts and procedures in singe- and multiple-group latent class analysis. In: A. L. McCutcheon (Ed.), *Applied latent class analysis* (pp. 56–88). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.003

McCutcheon, A. L. (1987). *Latent class analysis*. Sage.

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 967–985. https://doi.org/10.1080/10705511.2019.1590146

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 782–797. https://doi.org/10.1080/10705511.2016.1221313

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*, 45–60. https://doi.org/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. https://doi.org/10.1007/s11634-013-0146-2

Oberski, D. L., Vermunt, J. K., & Moors, G. B. D. (2015). Evaluating measurement invariance in categorical data latent variable models with the EPC-interest. *Political Analysis*, *23*, 550–563. https://doi.org/10.1093/pan/mpv020

Rao, C. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, *44*, 50–57.

Ruelens, A., & Nicaise, I. (2020). Investigating a typology of trust orientations towards national and European institutions: A person-centered approach. *Social Science Research*, *87*, 102414. https://doi.org/10.1016/j.ssresearch.2020.102414

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, *17*, 105–129. http://www.jstor.org/stable/271030 https://doi.org/10.2307/271030

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2020). Measurement invariance and differential item functioning across gender within a latent class analysis framework: Evidence from a high-stakes test for university admission in Saudi Arabia. *Frontiers in Psychology*, *11*, 622. https://doi.org/10.3389/fpsyg.2020.00622

Van Horn, M. L., Fagan, A. A., Jaki, T., Brown, E. C., Hawkins, J. D., Arthur, M. W., Abbott, R. D., & Catalano, R. F. (2008). Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research*, *43*, 289–326. https://doi.org/10.1080/00273170802034893

van Mierlo, H., Rutte, C. G., Kompier, M. A. J., & Doorewaard, H. A. C. M. (2005). Self-managing teamwork and psychological well-being: Review of a multilevel research domain. *Group & Organization Management*, *30*, 211–235. https://doi.org/10.1177/1059601103257989

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450–469. https://doi.org/10.1093/pan/mpq025

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

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for latent GOLD 5.1: Basic, advanced and syntax*. Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2021). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 356–364. https://doi.org/10.1080/10705511.2020.1818084