Measurement Invariance in Two-Step Latent Class Analysis: A Comparison of Residual Statistics to Identify Direct Effects

Jeroen Janssen

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: August 2017

Student number: 1833464

Supervisor: Dr. Zsuzsa Bakk

Acknowledgments

During this project, I have been given the opportunity to develop myself, both as a statistician and as a researcher. I have learned so much, from programming a simulation study to the theory behind latent class models and latent variable modeling in general. I would like to express my sincere gratitude towards my supervisor, dr. Zsuzsa Bakk, for giving me this opportunity, for helping me get through this process, always as patient and helpful , even when things did not go exactly as planned.

Second of all, besides my parents, family and friends in general, I would like to thank my good friend and roommate Maya, who has helped and supported me throughout the whole spectrum of feelings and moods I have gone through, from overly enthousiastic all the way to not feeling motivated anymore.

Thank you.

Abstract

Introduction: Problems with traditional latent class analysis approaches occur when a direct effect from the external variable to the indicators is present. An alternative approach, a two-step limited information maximum likelihood (LIML) method is developed by (Bakk & Kuha, 2017), possibly not having those problems anymore. Two residual statistics, the bivariate residual (BVR) and the expected parameter change (EPC) are used in a simulation study to test both the power and α-levels of the statistics to see whether they are able to pick out the direct effects. The LIML approach is compared to the one-step full information (FIML) method. *Method*: A simulation study was conducted where variations were made in (1) the number of direct effects, (2) the sample size, and (3) the separation between classes. For every simulation, it was checked whether the right direct effect(s) where picked out - if any, and whether the EPC or BVR where significant. Furthermore, a logistic regression analysis was conducted to see what factors are important in determining the statistics' performance. *Results*: Results show a general increase in performance (power) of both BVR and EPC if sample size and separation increase, even as some interaction effect resulting in differences between the two methods, with FIML generally performing better as compared to LIML. α-levels where nowhere at the nominal 5% level, resulting in either overfitting (EPC) or underfitting (BVR in FIML) effects. *Discussion*: The current research has brought possible problems with the LIML approach to the surface. The methods used in this research and latent class research in general are not in this form applicable for two-step estimation, meaning that alternative approaches have to be developed.

KEYWORDS: latent class analysis, limited information maximum likelihood, bivariate residual, expected parameter change, measurement invariance.

Table of Contents

List of Tables

List of Figures

1 Introduction

Imagine that you would like to predict political party affiliation based on people's response on various statements (e.g., using the *Stemwijzer* in The Netherlands), or predict the type of diabetes a patient has based on different blood measures. One could try to use a latent variable model to answer these questions, as the interest is in an underlying construct rather than an observed variable. A classic latent variable model will not work here, however, since our latent variable is categorical instead of continuous. Therefore, a new set of techniques was developed for these type of research questions, called latent class models or finite mixture models.

Latent class analysis (LCA) refers to a set of models used to classify subjects in different latent classes based on their response on a set of categorical indicator variables. For this reason, it can in a way be viewed as a categorical answer to factor analysis (McCutcheon, 1987). The LC model is used in various fields. Examples include estimating the risk factors for complex disorders in human genetic studies (Pickles et al., 1995), use classification of risk patterns to improve understanding of opioid drug users (Monga et al., 2007) and identify subgroups in poker players to improve intervention methods (Dufour, Brunelle, & Roy, 2013).

The classic LC model consists of both a measurement and a structural part. The measurement part entails the relationship between the latent class variable and the set of indicators. The structural part then relates this measurement model to external variables, either covariates, distal outcomes or both. In terms of estimation of these models, there are two traditional options, the one-step and the three-step method.

In the one-step approach (e.g., Hagenaars, 1993; McCutcheon, 1987; Vermunt, 2010), also called the Full Information Maximum Likelihood (FIML) approach, the structural and measurement part are both modeled at the same time in a single step. Although computationally efficient as no information is lost, this approach is rarely used in practice for a number of reasons. First of all, since the measurement and structural models can be seen as two separate aspects of the LC model, intuitively, it would make more sense to model these two parts separately Furthermore, a disadvantage of FIML is that small changes in the structural model lead to the forced re-estimation of both measurement and structural model, and possible changes in definition and estimation, and thus changes in interpretation of the measurement part and the latent classes (Bakk & Kuha, 2017).

In the three-step approach (Vermunt, 2010; Asparouhov & Muthén, 2014) on the other hand, the two parts are modeled separately. In a first step, the measurement part is estimated. Then, in the subsequent second step, the subjects are classified based on their estimated posterior class probabilities, obtained from step one. In the third step, this classification variable is related to external vari-

ables. Although this three-step approach may seem more appealing on a conceptual level because of the stepwise procedure, there are some possible problems related to this method. In the second step, the subjects are assigned to one of the classes with a probability of 1.0. However, in most of the cases, the posterior probability of belonging to a certain class is smaller than one, resulting in a classification error. Although there are several methods developed to correct for this classification error (Bolck, Croon, & Hagenaars, 2004; Asparouhov & Muthén, 2014; Bakk, Tekle, & Vermunt, 2013; Vermunt, 2010), it is known that in most cases, the classification step is but a starting point from which the actual analysis of interest can be conducted, i.e., relating the latent classes to external variables. Using these correction methods then seems a little overcomplicated if the classification is not actually used for interpretation. For this reason, among others, a more recent alternative approach is developed by Bakk and Kuha (2017), called the two-step LC model. This approach differs from its three-step alternative in the sense that there is no step related to classification, while still a stepwise logic is followed for computational and conceptual convenience. The authors show that this two-step LC model, also called a Limited Information Maximum Likelihood (LIML) model, is more efficient than three-step models, and almost as efficient as FIML, if the assumptions hold (Bakk & Kuha, 2017). The reason for this is the following: when using stepwise procedures, information is lost over the different steps, due to various reasons. In three-step models it is due to the use of the classification variable rather than the actual indicators in the third step, while in LIML the first-step parameters are fixed when estimating the second step (as will be explained later). This causes a loss of information and thus a loss of efficiency in estimation.

One of the key assumptions of latent class models is *local independence*, meaning that the relationship between the indicators is accounted for by the latent variable (McCutcheon, 1987; Hagenaars & McCutcheon, 2002); Stated otherwise, the indicator variables are assumed to be independent given the latent class membership. This same assumption is also applied for the indicators and the variables in the structural model. When this assumption is violated, the results of the LC model can be severely biased (Asparouhov & Muthén, 2014). For example, when a cross-cultural study is conducted, it might be the case that the questions (indicators) are interpreted differently across cultures (the external variable). This would indicate a direct causal relationship between the external variable and the indicators, referred to as differential item functioning (DIF; see for example Osterlind & Everson, 2009) in IRT literature, or, alternatively, *measurement non-invariance*. For a formal description and definition, see Masyn (2017). When this is the case, for example the three-step model is known to produce biased estimators (Asparouhov & Muthén, 2014), since the coefficient value belonging to the direct effect is enclosed in the relationship between the latent class and the external variable, making this relationship overestimated. This states the importance of knowing

whether there are direct effects present in the model, and directly shows one of the disadvantages of the three-step model. In the third step, the external variables are related to the classification variable, which is assumed to hold all the information from the indicators. However, the original information coming from these indicators is lost, making it unable for this model to conduct a posterior check for direct effects.

Therefore, if a check for measurement invariance is to be conducted, three-step models cannot be used. It makes sense then to use FIML, as no information is lost in the modeling process. However, as opposed to bias-adjusted three-step models, FIML is rarely used in practice, because of the computational feasibility and instability if the model changes (Bakk & Kuha, 2017). For the newly developed LIML models, as well as the FIML models, it is therefore still unknown how these models perform with respect to measurement invariance in latent class analysis.

In terms of model modification based on measurement non-invariance, it is even argued by Kuha and Moustaki (2015) that in general multi-group latent variable models, in some situations, ignoring this nonequivalence is a better approach than modeling it. However, the authors did not investigate LC models in the concerned article. The current project will follow a similar strategy, namely checking the models' performance when the direct effects (i.e., measurement non-invariance) are not modeled.

Testing for measurement invariance is common practice in general SEM, and in multi-group latent variable models in particular, since measurement invariance is about the interpretation of the measurement in different groups. The article by Kim, Cao, Wang, and Nguyen (2017) gives an overview of different techniques that are used in latent variable literature. The article mainly focused on the actual testing of DIF with posterior model fitting, which is only one of the possibilities. Another possibility in latent variable modeling is the use of global and local fit statistics to check for DIF. For a brief overview, see Van der Schoot, Ligtig, and Hox (2012).

A third possibility is checking statistics that use the residuals of the fitted model to see whether any significant association is left unmodeled. Although less frequently used in general SEMs, residual association checks are more common in LC literature (e.g., Oberski, Vermunt, & Moors, 2015; Oberski, van Kollenburg, & Vermunt, 2013; Nagelkerke, Oberski, & Vermunt, 2017).

One of the statistics that can be used is the bivariate residual (BVR; Vermunt & Magidson, 2005), indicating the amount of residual association left between two variables after the model is fitted. BVR-statistics between indicators and covariates can tell us something about possible direct effects in the model. Another example is the use of a score statistic, for instance the expected parameter change (EPC; Oberski & Vermunt, 2014). The score-based EPC indicates the amount by which a parameter would change if it would be freed rather than fixed.

This EPC statistics is commonly used in for example econometrics, where it is called a Lagrange multiplier (e.g., Breusch & Pagan, 1980). In the field of structural equation modeling, this statistic is related to the Modification Index (MI) used in that field. For a description of MI and differences as compared to EPC), see Whittaker (2012).

Although residual statistics are used to investigate measurement invariance in FIML by Oberski et al. (2013), the performance of these statistics is not yet tested in the newly developed two-step LC model, which will be one of the main goals of the current article.

The current project will focus on the detection of DIF (i.e., the detection of direct effects) in two-step and one-step LC models by using BVR and EPC residual statistics. The study can then be divided in two slightly different parts. First of all, we will investigate the residual statistics' type I error probability by examining its ability to detect the absence of direct effects when in fact there are none present (RQ1). The second part will focus on the statistics' power, meaning its ability to find the correct direct effects when there are one or more present in the model (RQ2).

Based on previous literature, some expectations can be formulated. First of all, since both BVR and EPC are large-sample statistics (Oberski et al., 2013), it can be expected that the statistics perform better in larger compared to smaller simulated samples. Since the one-step approach is a full information method, as compared to the limited-information two-step model, the one-step can be expected to give better statistics' performance, due to statistical efficiency. Lastly, the BVR is treated as following a chi-squared distribution with one degree of freedom, although in fact it is known that this is not true (e.g., Oberski et al., 2013). Therefore, the EPC might perform better compared to the BVR.

The remainder of this thesis is structured as followed. First of all, the two-step model as developed by Bakk and Kuha (2017) is described, together with a definition and description of the used residual statistics. Next, the simulation setup and its results are discussed. The article concludes with a (general) conclusion with implications for current and future research.

2 The Two-Step Latent Class Model

2.1 Step 1: The Measurement Model

The first step of the analysis consists of a simple latent class model to define the measurement part of the model. Suppose we have K categorical indicators. Let y_{ik} be the response of person i on indicator k, with $k \in \{1, ..., K\}$, and let $\mathbf{Y}_i$ denote the full response pattern of person i. Then define a latent variable X consisting of T different classes, such that $t \in \{1, ..., T\}$. A latent class model to

define $P(\mathbf{Y}_i)$ can be defined as (e.g., McCutcheon, 1987):

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

Typically, given class membership, categorical responses are assumed to be independent. That is, the probability of a response pattern given class membership can be defined as the product of the item-specific response probabilities, or,

$$P(\mathbf{Y}_i|X = t) = \prod_{k=1}^{K} P(Y_{ik}|X = t) = \prod_{k=1}^{K}\prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)}, \tag{2}$$

where $\pi_{ktr} = P(Y_{ik} = r|X = t)$ and $I(Y_{ik} = r)$ is an indicator variable that equals 1 if the response on indicator k for subject i equals r, and 0 else. Then, if we denote $\xi_t = P(X = t)$, the T class probabilities and the $(K - 1)KT$ item-specific response probabilities $\{\pi_{ktr}\}$ can be combined in a parameter vector $\boldsymbol{\theta}_1 = [\boldsymbol{\xi}, \boldsymbol{\pi}]$. Substituting Equation (2) in Equation (1) and assuming independence of observations, this parameter vector can be estimated by maximizing the first-step log-likelihood function $\mathcal{L}_1$, defined as (e.g., Vermunt, 2010)

$$\mathcal{L}_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{N} \log P(\mathbf{Y}_i) = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} \xi_t \prod_{k=1}^{K}\prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)} \right]. \tag{3}$$

The sample estimate of this vector, $\hat{\boldsymbol{\theta}}_1$, can be retrieved using expectation-maximization, a quasi-Newton method or a combination of the two.

2.2 Step 2: The Structural Model

The second step of the analysis consists of two parts. First, in Step 2a, the structural model is estimated by relating the measurement model to covariates. This is done by fixing the measurement parameters to the ones estimated in the first step, $\hat{\boldsymbol{\theta}}_1$. Residual statistics then indicate whether extra paths should be added to the model. This is then done in Step 2b.

2.2.1 Step 2a: Relating the model to covariates

Let $\mathbf{Z}_i$ be the covariate vector for person i, and Z_i its value on covariate Z. Then, the second step can be written as (Bakk & Kuha, 2017)

$$P(\mathbf{Y}_i|X_i = t, Z = z_i) = \underbrace{P(X = t|Z = z_i)}_{\text{free}} \underbrace{P(\mathbf{Y} = \mathbf{y}_i|X = t)}_{\text{fixed}}. \tag{4}$$

The second term, the measurement part, is fixed to the estimates found in step 1. The first term, the structural part, can be parametrized using a multinomial logit model as follows (Bakk & Kuha, 2017):

$$P(X = t|\mathbf{Z}_i) = \frac{\exp(\beta_{0t} + \beta_t Z_i)}{\sum_t \exp(\beta_{0t} + \beta_t Z_i)}. \tag{5}$$

Then, using this parametrization the Step 2a model can be estimated using a log-likelihood function $\mathcal{L}_2$ for the second parameter vector $\boldsymbol{\theta}_2$:

$$\mathcal{L}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1) = \sum_{n=1}^{N} \log \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i) P(\mathbf{Y} = \mathbf{y}_i|X = t). \tag{6}$$

As opposed to the two-step way of estimating measurement and structural model, the FIML method estimates both parts of this model at the same time.

2.2.2 Step 2b: Evaluating residual statistics to find possible direct effects

As stated above, the last step of the model consists of investigating the residual statistics in order to see whether direct $Z - \mathbf{Y}$ paths should be added. It is however possible - if not of explicit interest - to omit this last step and stop the model building process at Step 2a.

The current study will make use of the *bivariate residual* (BVR) and the *expected parameter change* (EPC), both of which will be discussed below. The BVR (Vermunt & Magidson, 2005, pp. 72-3) for a pair of observed variables can be defined as the Pearson residual in the bivariate cross-table (Oberski et al., 2013, p. 2). For two given variables y_i and y_j, both having values 0 or 1, it is defined as:

$$BVR_{ij} = \sum_{k\in\{0,1\}} \sum_{l\in\{0,1\}} \frac{(n_{kl} - \hat{\mu}_{kl})^2}{\hat{\mu}_{kl}} \quad \forall_{i\neq j}, \tag{7}$$

where n_{kl} and $\hat{\mu}_{kl}$ equal the observed and expected frequencies in the 2×2 cross-table, respectively. As a value for the BVR for every pair of variables is given as output in basic software such as Latent GOLD (Vermunt & Magidson, 2005), this is an elegant way of locally examining whether paths should be added.

The EPC (Oberski & Vermunt, 2014; Oberski et al., 2013), based on the "classical score-test" by Rao (1948), is a well-known residual statistic in the context of item response theory (Glas, 1999) and structural equation modeling (e.g., Saris, Satorra, & Sörbom, 1987; Oberski, 2014). Recently it was described by Oberski et al. (2013) to use in binary LC models as well. The EPC is a score statistic, meaning that it estimates the strength of a given effect, should it be freed in an alternative

model. For two given variables y_i and y_j it is defined as (Oberski et al., 2013):

$$EPC_{ij} = \frac{s_{ij}^2}{\text{var}(s_{ij})} \quad \forall_{i \neq j}, \tag{8}$$

with $s_{ij} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \psi_{ij}}$ as a value for the 'score' in a local dependence test. In this last definition, ψ_{ij} is an element of $\boldsymbol{\psi}$, the two-way interaction matrix between y_i and y_j. See Oberski et al. (2013) for a detailed definition and a discussion of the relationship between the BVR and the EPC.

Although this is not done in the current study, the paths as indicated by the BVR and EPC can be added to the model, and the new model can be evaluated to see whether the fit improved. This is done by freeing the concerned paths, with a model similar to Equation (4):

$$P(\mathbf{Y}_i|X_i = t, Z = Z_t) = P(X = t|Z = z_i)P(\mathbf{Y} = \mathbf{y}_i|X = t, Z_i). \tag{9}$$

The variables that indicated a direct effect are freely estimated in the last term, whereas the ones that did not are kept at the values from Step 1.

3 Simulation Study

In order to investigate the performance of the EPC and BVR in detecting direct effects, a Monte Carlo simulation is designed and conducted. The statistics are tested in two different models: the full-information maximum likelihood (FIML) one-step model, and the limited-information maximum likelihood (LIML) two-step model.

The model that is simulated from consists of a three-class latent class variable X, six different observed binary indicators $\mathbf{Y} = (Y_1 ... Y_6)$ and one observed covariate Z_1, values ranging from 1 to 5. Figure 1 graphically shows how this model looks. The classes are modeled in such way that the first class is likely to have a positive response on all six indicators, the second class is likely to respond positive on the first three variables and negative on the last three, while class three has a high probability of responding negative to all six indicators. This approach is in line with the set-up used by for example Bakk et al. (2013) and Vermunt (2010).

Variations will be made in a number of parameters, while others will be kept constant due to computational considerations. First of all, the probability of giving a positive response is varied ($S \in \{.70, .80, .90\}$). These values correspond to an entropy based pseudo-R^2 value of .36, .65 and .90 and a low, middle and high separation between classes, respectively. This variation has an effect on the quality of the classification in three-step approaches, as shown by Vermunt (2010). Although three-step models are not discussed here, we expect a similar pattern in two-step analyses. In the

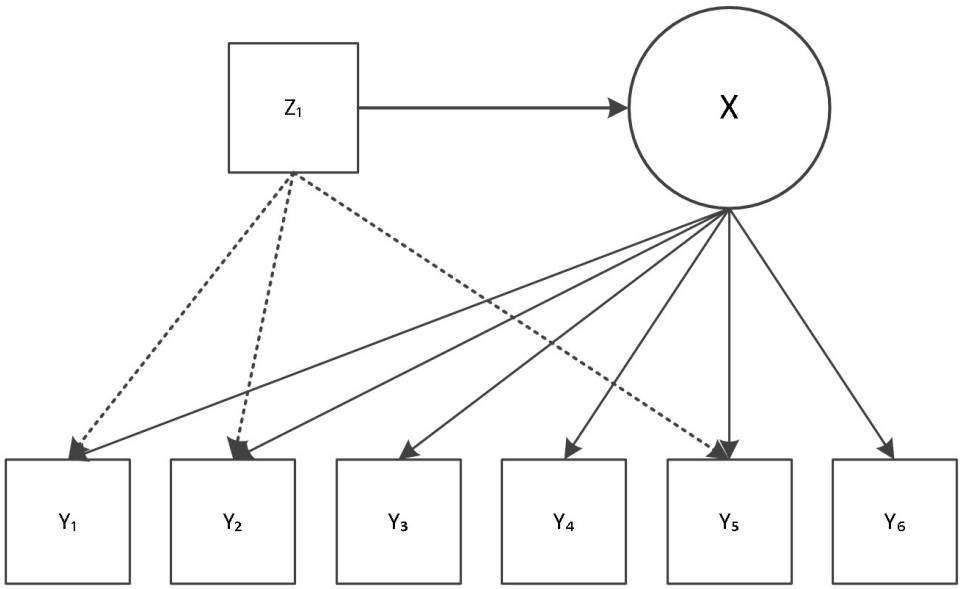

Figure 1. A graphical representation of the LC model that is simulated from, where $Y_1 ... Y_6$ represent the observed binary indicators, X the latent class variable, and Z_1 the observed categorical covariate. The dashed lines represent the direct paths that will be added to the population model according to the condition.

conditions with the lowest entropy values, we expect the stepwise procedures to be biased, since information from the covariate is needed in order to estimate the latent class variable correctly. When the entropy (i.e., separation) increases, the measurement model becomes powerful enough on its own. Second of all, the sample size is varied ($N \in \{500, 1000, 2000, 4000\}$), since it is shown by Oberski et al. (2013) that both BVR and EPC are large-sample statistics. Therefore it is interesting to see how they perform in relatively small sample sizes as well. Lastly, the number of direct effects is varied ($D \in \{0, 1, 2, 3\}$). The reason for this is that Asparouhov and Muthén (2014) showed an increase of bias in FIML models when the number of - unmodeled - direct effects increased. For this reason we would like to know how the statistics perform under more difficult (i.e., more biased) conditions. In our simulations, the direct effects are modeled on Y_1 for $D = 1$, Y_1 and Y_2 for $D = 2$, and Y_1, Y_2 and Y_5 for $D = 3$.

The population values that will be used in this simulation to asses the $Z - X$ association will be $\alpha_1 = 4.73$ and $\alpha_2 = -3.699$, $\beta_1 = -2.0$ and $\beta_2 = 1.0$. The direct effect will be class-specific, with a medium effect of $\zeta_1 = -0.5$, $\zeta_2 = 0$ and $\zeta_3 = 0$, with the index referring to the class, added to the indicators as mentioned above.

Now, the first split in research questions can be made, i.e., conditions where $D = 0$ and conditions where $D \neq 0$. In the first set of conditions, the focus of interest is in the residual statistics' type I error probability, meaning the probability of them finding direct effects when in fact there are none

(RQ1). In the second set, the focus is on the statistics' power [1], i.e., their ability to pick out direct effects. RQ2 will be the focus of this part of the study. Both RQ1 and RQ2 will be answered using both FIML and LIML Step 1 and Step 2a.

In all of the simulation conditions, we test whether the right direct effects are identified by investigating the BVR and EPC vectors for the $Z_1 - \mathbf{Y}$ associations - thus both containing six values per simulation - and see whether the largest one, two or three (depending on the condition) values are indeed on the associations we put the direct effect on. For all of the simulations where the direct effects are correctly classified, we will subsequently check whether the values are indeed significant, by comparing EPC and BVR to a χ^2-distribution with df = 3 and df = 1, respectively. Together with that, the mean bias (mean deviation from the population values) and root mean squared error (RMSE) will be reported for every condition. A similar procedure will be used to answer RQ1.

In order to perform the simulation study, we used the computer software Latent GOLD version 5.1.0.17046 (Vermunt & Magidson, 2005), and RStudio version 0.99.903 (R Core Team, 2015). The simulation consists of 4 (D) $\times$ 3 (S) $\times$ 4 (N) = 48 conditions, for all of which 500 replications were used.

4 Results

4.1 Preliminaries

Before we proceed to the actual answers to the research questions, a few small remarks about the results have to be made. First of all, a quick note about the simulations. There were several cases, especially in the low sample size and low separation conditions, in which no final solution was found (i.e., no convergence after maximum number of iterations), referred to as boundary solutions. This resulted in either an error message in Latent GOLD, not giving all six $Z_1 - \mathbf{Y}$ values for EPC, or extremely large standard errors (i.e., larger than 3.0). These simulations were excluded from further analyses, since boundary issues - though very interesting on their own - are not within the scope of this project. Table 1 gives an overview of how many replications were excluded. These numbers are not surprisingly high, and in accordance with for example Bakk et al. (2013).

Second of all, we looked at the mean bias and the root mean squared error (RMSE) of the two $Z - X$ slope parameters modeled as β_1 = -2.0 and β_2 = 1.0. The mean bias and RMSE are defined as

[1] With *power* we mean the actual probability of rejecting H_0 if it is false, with H_0 being no direct effects in the model, and H_1 being 1, 2 or 3 direct effects present, depending on the level of D. Although this is not considered power in terms of being a formal statistical test, it will still be referred to as 'power' for convenience.

Table 1

Number of Excluded Simulations Due to Non-Convergence Per Level of N, S and D, Summed Over FIML and LIML.

Condition		D			
N	S	0	1	2	3
500	Low	107	74	85	220
	Mid	2	1	1	11
	High	1	0	1	0
1000	Low	29	7	7	47
	Mid	0	0	1	0
	High	0	0	0	0
2000	Low	2	0	0	1
	Mid	0	0	0	0
	High	0	0	0	0
4000	Low	0	0	0	0
	Mid	0	0	0	0
	High	0	1	0	1

$$M_{Bias} = \frac{1}{J} \sum_{j \in \{1 \ldots J\}} (\beta_{ij} - \hat{\beta}_{ij}), \qquad RMSE = \sqrt{\frac{1}{J} \sum_{j \in \{1 \ldots J\}} (\beta_{ij} - \hat{\beta}_{ij})^2}$$

for the J simulations in every condition. Table 2 gives an overview of these bias values, averaged over the two parameters.

If we look at the misspecified models in Table 2 (i.e., models where $D \neq 0$), it can be seen that there is quite a large amount of bias, especially in the worst conditions. Bias tends to decrease with sample size and separation, leading to smaller amounts (and thus less modeling issues) in the best conditions. No clear difference is visible between FIML and LIML. These values clearly indicate that something can be modeled differently (i.e., that there is room for improvement), justifying and explaining the modeling of direct effects.

4.2 Research question 1 (type I error probability)

First of all, the results of the simulation conditions where $D = 0$ will be discussed. Table 3 shows the proportion of simulations where any of the six BVR or EPC values for the $\mathbf{Y} - Z_1$ paths were significant, i.e., significant under a χ^2-distribution with $df = 3$ or $df = 1$ for EPC and BVR, respectively. This table helps us in investigating the type I error probability of the residual statistics under certain conditions.

A couple of trends become visible from this table. First of all, for LIML there is a monotoni-

Table 2

The Mean Bias (M_{Bias}) and Root Mean Squared Error ($RMSE$), Averaged Over β_1 and β_2, for All Levels of N, S and D, for Both FIML and LIML.

Condition		$D = 0$				$D = 1$				$D = 2$				$D = 3$			
		FIML		LIML		FIML		LIML		FIML		LIML		FIML		LIML	
N	S	M_{Bias}	RMSE	M_{Bias}	RMSE	M_{Bias}	RMSE	M_{Bias}	RMSE	M_{Bias}	RMSE	M_{Bias}	RMSE	M_{Bias}	RMSE	M_{Bias}	RMSE
500	Low	0.47	1.94	-0.47	0.93	0.39	1.64	-0.30	0.84	0.24	1.50	-0.11	0.78	0.06	3.01	-0.18	0.87
	Mid	0.03	0.29	-0.08	0.32	0.08	0.30	-0.03	0.29	0.09	0.31	0.02	0.29	0.05	0.37	-0.04	0.32
	High	0.03	0.21	0.01	0.21	0.02	0.21	0.00	0.20	0.02	0.20	0.01	0.19	-0.01	0.21	-0.04	0.20
1000	Low	0.11	0.79	-0.35	0.71	0.12	0.53	-0.17	0.57	0.18	0.39	0.00	0.53	0.14	0.95	-0.02	0.63
	Mid	0.02	0.20	-0.03	0.21	0.06	0.21	0.01	0.19	0.08	0.23	0.04	0.21	0.01	0.25	-0.04	0.23
	High	0.02	0.14	0.01	0.14	0.02	0.14	0.01	0.14	0.01	0.14	0.00	0.14	-0.02	0.15	-0.03	0.14
2000	Low	0.02	0.24	-0.22	0.45	0.10	0.24	-0.04	0.34	0.15	0.27	0.07	0.32	0.13	0.33	0.05	0.39
	Mid	0.00	0.13	-0.02	0.14	0.05	0.15	0.02	0.14	0.07	0.17	0.04	0.15	0.01	0.19	-0.02	0.17
	High	0.00	0.10	0.00	0.10	0.01	0.10	0.01	0.10	0.01	0.10	0.01	0.10	-0.02	0.11	-0.02	0.11
4000	Low	0.01	0.16	-0.10	0.25	0.10	0.18	0.03	0.19	0.13	0.22	0.11	0.22	0.15	0.27	0.12	0.25
	Mid	0.01	0.09	0.00	0.09	0.04	0.11	0.02	0.11	0.06	0.13	0.04	0.12	0.02	0.16	0.00	0.14
	High	0.00	0.07	0.00	0.07	0.01	0.07	0.00	0.07	0.01	0.07	0.01	0.07	-0.02	0.08	-0.02	0.08

Table 3

Proportion of simulations in which any of the direct paths where significant according to BVR and EPC, for $D = 0$, for all levels of N and S

Condition		FIML		LIML	
N	S	EPC	BVR	EPC	BVR
500	Low	.51	.00	.30	.87
	Mid	.42	.00	.26	.31
	High	.36	.00	.02	.01
1000	Low	.43	.00	.57	.90
	Mid	.32	.00	.18	.25
	High	.30	.00	.01	.01
2000	Low	.37	.00	.79	.85
	Mid	.28	.00	.10	.24
	High	.29	.00	.00	.01
4000	Low	.28	.00	.75	.80
	Mid	.28	.00	.10	.24
	High	.26	.00	.00	.01

cally decreasing type I error probability if the separation increases. For EPC in FIML, a similar yet somewhat less pronounced trend can be found. The sample size N, however, has the opposite effect. Whereas a decreasing trend can be found for BVR in FIML in all the separation conditions, sample size seems to be less important in explaining the statistics' performance in LIML.

In BVR for FIML, a slightly underfitting effect is visible, as almost no (i.e., rounded down to .00) significant simulations were found in any of the conditions, as opposed to the nominal level of 5% that would be expected.

4.3 Research question 2 (power)

The second part of the current research focuses on situations where $D \neq 0$, i.e., where there were direct paths present in the population model. We check the residual statistics in their performance of picking out these direct effects. This part of the research question thus focuses on their power.

Figure 2 shows the results of these simulation conditions. The power is plotted against the sample size, separately for all levels of D (except for $D = 0$) and S. These values can also be found in Table A1, together with the significance levels for all of these values.

Figure 2 graphically shows us different aspects that are present in the current simulation results. First of all, we see that in almost all of the conditions, FIML is performing better than LIML. This effect is more visible in the lower separation conditions as compared to the high separation

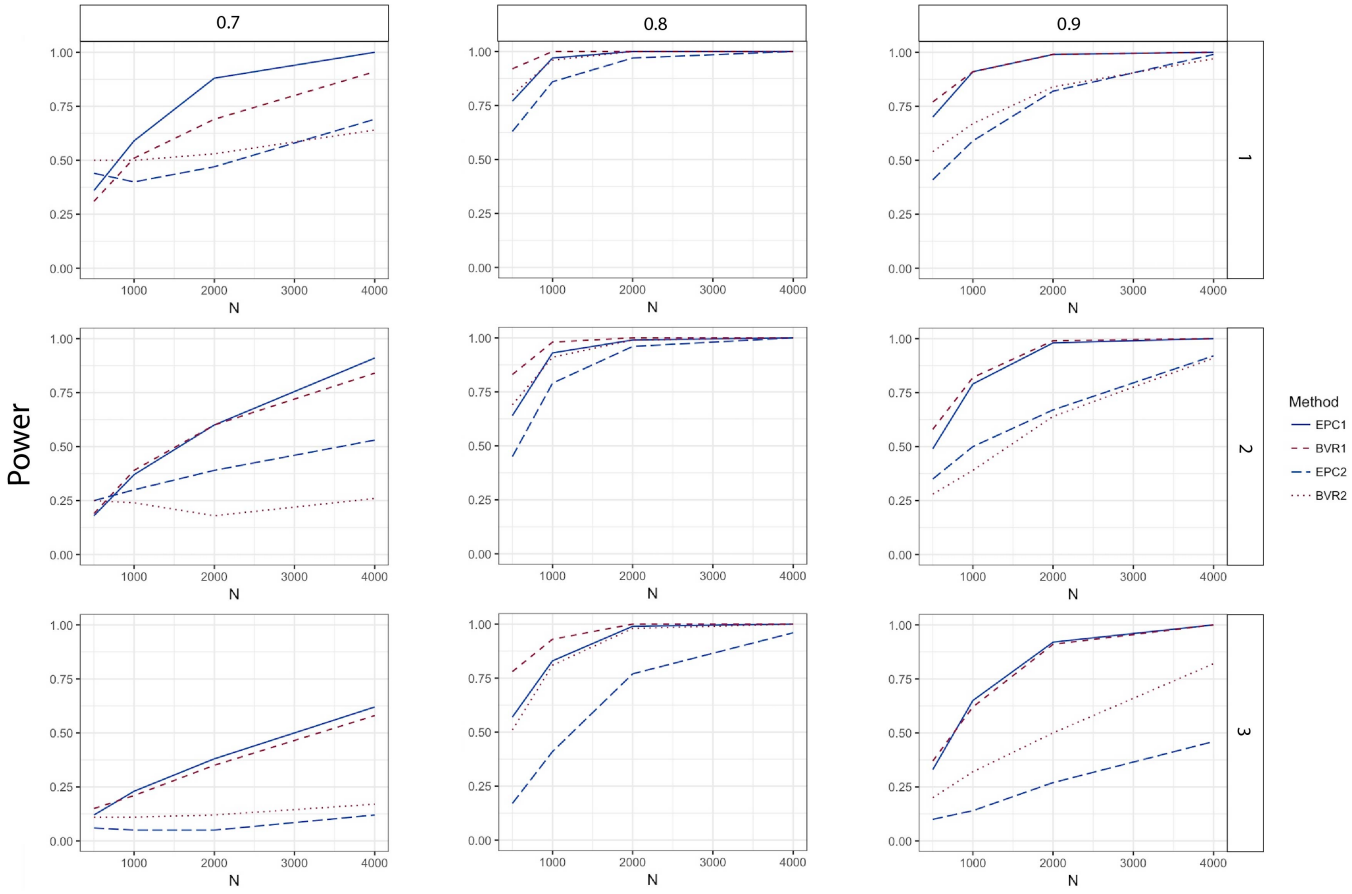

Figure 2. The power plotted as a function of the sample size, in a separate panel for every separation level (horizontally) and number of direct effects (vertically). The different lines represent the results for BVR and EPC, both for FIML (1) and LIML (2).

condition.

Second of all, performance seems to increase, for both FIML and LIML, as sample size and separation are increasing. Whereas FIML and LIML are somewhat diverging in the low separation condition - with FIML increasing more as opposed to LIML - the two methods show a more converging trend in performance over sample size as separation increases. The best conditions are those with high separation and high sample size.

The number of direct effects is shown to have a decreasing effect on power levels. Though less pronounced than separation and sample size effects, performance seems to decrease if D increases. This especially becomes clear in worse conditions. This downward trend is somewhat more visible for LIML as compared to FIML.

To see in a more inferential way what the important factors are for BVR and EPC performance, two binary logistic regression analyses were conducted. The number of direct effects D = $\{1, 2, 3\}$, separation levels S = $\{0.7, 0.8, 0.9\}$, method M = $\{$FIML = 0, LIML = 1$\}$ and sample size N = $\{500, 1000, 2000, 4000\}$ were used as predictors in the regression of the correct identified direct effects Y = $\{$yes = 1, no = 0$\}$, where correctly identified means that all of the one, two or three direct effect were identified. Significance levels were not included in these analyses. A separate regression analysis was conducted for BVR and EPC, with all main effects and interactions added at the same time.

In this analysis, sample size, separation and number of direct effects were treated as continuous variables, while in fact they are ordinal. This is done because treating them as categorical would make interpretation much less convenient due to the large number of levels. Furthermore, a more or less linear trend over the levels of these variables is expected due to the choice of these levels.

A total of 35,084 simulations were included in the logistic regression. Table 4 shows the parameter estimates $\hat{B}$, standard errors and odds ratios $\exp(\hat{B})$ for all direct effects and all two- and three-way interactions.

The model of interest, containing all effects, had a significantly better fit as compared to the constant-only model, for both BVR ($\chi^2(15)$ = 16,059.63, $p < .001$) and EPC ($\chi^2(15)$ = 16,245.15, $p < .001$), with Hosmer and Lemeshow's pseudo R^2 values of .37 (BVR) and .36 (EPC).

Due to the large number of interactions involved in these analyses, interpretation of single, specific effects becomes increasingly difficult. In terms of main effects, for BVR a significant increase in odds of finding the correct direct effects is found when the number of direct effects becomes smaller, and separation becomes larger - keeping everything else constant. The same holds for EPC, with an additional significant main effect of sample size. Both BVR and EPC also show a significant increase in odds if the method changes from FIML to LIML.

Table 4

Parameter Estimates, Standard Errors and Significance Levels of the Logistic Regression Analysis.

	BVR		EPC	
Effect	$\hat{B}$ (SE)	$\exp(\hat{B})$	$\hat{B}$ (SE)	$\exp(\hat{B})$
Constant	-2.13 (0.40)†	0.12	-2.05 (0.38)†	0.13
D	-0.60 (0.18)†	0.55	-0.42 (0.17)*	0.66
M	2.94 (0.51)†	18.83	1.84 (0.50)†	6.27
N	0.18 (0.20)	1.20	1.00 (0.20)†	2.73
S	0.91 (0.25)†	2.49	0.46 (0.21)*	1.59
D*M	-0.91 (0.25)†	0.40	-0.43 (0.25)	0.65
D*N	0.02 (0.08)	1.02	-0.28 (0.08)†	0.76
D*S	0.02 (0.10)	1.02	-0.02 (0.09)	0.98
M*S	-1.07 (0.30)†	0.34	-0.57 (0.26)*	0.57
M*N	-0.55 (0.24)*	0.58	-1.00 (0.23)†	0.37
N*S	0.87 (0.16)†	2.39	0.57 (0.14)†	1.78
D*M*N	-0.03 (0.11)	0.98	0.17 (0.10)	1.18
D*M*S	0.27 (0.13)*	1.30	0.05 (0.12)	1.06
D*N*S	-0.10 (0.06)	0.91	0.04 (0.06)	1.04
M*N*S	-0.31 (0.18)	0.74	0.06 (0.16)	1.06

Note. $^{*}p < .05$; $^{\diamond}p < .01$; $^{\dagger}p < .001$.

As for BVR, the effect that D has on the odds of identifying the correct effects, differs between the two methods, with this trend being less pronounced in LIML. The same holds for both sample size (though no significant main effect) and separation. Almost similar results are found for EPC, as can be seen from the significant two-way interactions in Table 4.

Table A1 shows us the significance levels for the correctly identified simulations in the different conditions. Few clear trends can be found in these results. Significance levels overall seem to increase with sample size and separation (although some exceptions are found), as well as with the number of effects. In the D = 3 condition, only few simulations were still significant (especially in the worse conditions). No formal test has been conducted for significance levels.

5 Discussion

5.1 Discussion of the results

The current research focused on two different aspects of the same process, namely the performance of residual statistics in identifying direct effects in the newly developed two-step latent class models. To this end, the performance was compared to a more classical one-step approach. Two different parts of the performance were investigated, i.e., the power and the type I error probability.

Both of these aspects will be discussed below.

With regard to the α-level, usually a nominal level of .05 or 5% is chosen, meaning that in 5% of the cases the null hypothesis is allowed to be rejected while true. With the null hypothesis in the current research being that there are no direct effects present, a type I error is made if any of those paths are significant. We also expect that to be the case in 5% of the cases.

If we look at the results that were found, we see that those nominal values are nowhere really found. In FIML, the EPC error rates tend to be too high even in the best conditions, whereas the BVR values tend to be too low. This latter trend could be caused by the fact that the $\chi^2(1)$ distribution we use for significance of BVR is not entirely true, as the asymptotic distribution of the BVR is unknown (Oberski et al., 2013). The increased α-levels in almost all other conditions will increase the false positive rate of the model, and can therefore lead to overfitting effects.

To the best of our knowledge, only the article by Oberski et al. (2013) investigated type I error in LC models, though in a slightly different way. In a less complex model with categorical distal outcomes, a single residual association between two indicators was modeled and/or investigated, where a type I error was made if that specific association was significant though not modeled. In the current research, all $\mathbf{Y} - Z_1$ paths were investigated. Although there are some differences between the two studies, somewhat similar results were found, meaning an underfitting effect of BVR. EPC error rates were at nominal levels in the article by Oberski et al. (2013), while here they are somewhat higher. One of the possible explanations for these differences could be the increased difficulty of the current models.

The second part of our simulation studies focused on the power of the statistics, meaning the probability to reject the null hypothesis when it is untrue. With the null hypotheses being the absence of either one, two or three effects (depending on the level of D we are in), the power indicates the proportion of simulations in which the (correct) direct effects were identified. Normally, power levels of .80 are desired.

The results of these simulations first indicate that there is a general difference between FIML and LIML, being the core of our second research question. The effect that FIML generally performs better than LIML, might be explained by the structure of those two models. In FIML, the complete model is estimated at the same time, meaning that in estimating the structural part all information of the measurement model can still be used (though both processes happening simultaneously), possibly leading to more stable results. In LIML, however, the parameters in the measurement model are fixed when estimating the structural part, leading to a loss of information.

Results of the logistic regression analysis show a significant increase in odds when changing from FIML to LIML, which is in contrast with our expectations and the results of the simulations. A

possible explanation for this could be the large number of significant interation effects this variable has. An analysis containing only the four main effects shows a significant negative effect of method.

Furthermore, next to a general effect of methods, there are also some significant interactions with the used methods, namely those of sample size and separation. The difference between FIML and LIML seems to become larger if samples and separation increase. A higher separation between indicators means that more information can be taken out of the measurement model. The previous reasoning thus can also be used to explain this significant interaction. We know that a larger value means more information can be gathered from the measurement model. In datasets with lower separation levels, information from the covariate is needed in order to correctly estimate the model, since the measurement model on its own is not enough. In such weaker models, the covariate is the strongest indicator in the model. Leaving it out in estimating the measurement part, will lead to a lower entropy value and thus to less stable results in stepwise procedures like LIML (Vermunt, 2010).

A significant interaction of method and sample size was also found. As compared to separation levels, sample size is another way of influencing the strength of the measurement model. As mentioned before, the small sample and low separation conditions are a bigger problem for stepwise procedures as compared to full information procedures, since measurement information is needed for a stable model estimation.

Overall, the results of the logistic regression analysis largely correspond with the findings of the simulation results in Table 3, Table A1 and Figure 2. There is a visible increase in performance when separation and sample size increase (though not always significant), and this increase differs between methods. There also seems to be a small yet significant decrease in performance when the number of direct effects increases, possibly caused by the definition of a correctly identified simulation and the increase in model complexity.

The last part of the study focused on the significance levels. A simulation was called significant if all of the modeled $Z_1 - \mathbf{Y}$ paths were. This is a possibly cause for the decreased levels if D increases. Increased model complexity as well as pure chance levels can be underlying this trend. Furthermore, although somewhat less pronounced, there are situations visible where the correct effect are found, and yet the statistics' values are not significant. This could potentially lead to problems if significance is not investigated.

5.2 Implications of current research

In the current research, we have tried to contribute to the testing of the newly developed models by Bakk and Kuha (2017). We checked the power and type I error probabilities of BVR and EPC in FIML and LIML models. Some conclusions might be drawn from the results. First of all, there are some situations, especially when separation and thus entropy levels or sample sizes are low, where stepwise models (or maybe even LCA in general) might not be the best idea, since this might lead to problems in estimating the models and/or significance of the estimations. Furthermore, it is shown by the current research that the investigated situations - both the worst and better conditions - sometimes lead to unexpected results. EPC and BVR, in the form they are used in at the moment, are not always the best idea when using LIML models. This implicates that further research should look at alternatives for these statistics, to investigate in what conditions the model does work.

5.3 Suggestions for further research

Further research should focus on a few aspects. Since larger samples yield better results, it may be interesting to search for the point (i.e., sample size) where performance switches from worse to better. Furthermore, the current research uses a continuous covariate, whereas real-life research often works with categorical covariates (e.g., what is the effect of gender (male/female) on the statements of the *Stemwijzer*?). This may influence the performance due to the computational differences in building the model.

Another option may be to look beyond the residual statistics. If the problem of the current models is in the standard errors, residual statistics may not be such a good way to go. Other options are for example the likelihood-ratio test, which can compare the likelihood ratio value of two models (i.e., with and without a certain effect), or, the Wald test, testing whether a certain effect is zero or not (Agresti, 2002).

Concluding all the aforementioned remarks and results, it is safe to say that the current project is a good start of research in the development and improvement of newly developed LC models. Further research, building on this project or taking a completely different point of view, should try to reach further understanding of what is going on, and try to improve the models as good as possible. For now, we can say that the used methods are possibly not the optimal techniques for the current dataset, that stepwise LC models may not even be suitable in very bad conditions, and that we are very curious to see how this will develop.

References

Agresti, A. (2002). *Categorical data analysis. Second edition*. Hoboken, NJ: John Wiley and Sons, Inc.

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

Bakk, Z., & Kuha, J. (2017). Two-step estimation of models between latent classes and external variables. *Working paper*.

Bakk, Z., Tekle, F. B., & 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. doi: 10.1177/0081175012470644

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3-27. Retrieved from `http://www.jstor.org/stable/25791751`

Breusch, T., & Pagan, A. (1980). The Lagrange Multiplier test and its applications to model specifications in econometrics. *Review of Economic Studies*, *47*. Retrieved from `http://www.jstor.org/stable/2297111`

Dufour, M., Brunelle, N., & Roy, É. (2013). Are poker players all the same? Latent class analysis. *Journal of Gambling Studies*, *31*, 441-454. doi: 10.1007/s10899-013-9429-y

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

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Newbury Park, CA: Sage.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge University Press.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 524-544. doi: 10.1080/10705511.2017.1304822

Kuha, J., & Moustaki, I. (2015). Non-equivalence of measurement in latent modeling of multigroup data: A sensitivity analysis. *Psychological methods*, *20*, 1-47. doi: 10.1037/met0000031

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. doi: 10.1080/10705511.2016.1254049

McCutcheon, A. L. (1987). *Latent class analysis* (No. 64). Newbury Park, CA: Sage.

Monga, N., Rehm, J., Fischer, B., Brissette, S., Bruneau, J., El-Guebaly, N., ... Bahl, S. (2007). Using latent class analysis (LCA) to analyze patterns of drug use in a population of illegal opioid users. *Drug and Alcohol Dependence*, *88*, 1-8. doi: https://doi.org/10.1016/j.drugalcdep.2006.08.029

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2017). Power and type I error of local fit statistics in multilevel latent class analysis. *Structural equation modeling: a multidisciplinary journal*, *24*, 216-229. doi: 10.1080/10705511.2016.1250639

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

Oberski, D. L., & Vermunt, J. K. (2014). *The expected parameter change (EPC) for local dependence assessment in binary data latent class models.* Accepted for publication in Psychometrika.

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

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Thousand Oaks, CA: Sage.

Pickles, A., Bolton, P., Macdonald, H., Bailey, A., Le Coutour, A., Sim, C. H., & Rutte, M. (1995). Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: A twin and family history study of autism. *American Journal of Human Genetics*, *57*, 717-726.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*, 159–203. Retrieved from `http://www.jstor.org/stable/2983775`

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

Van der Schoot, R., Ligtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-492. doi:

10.1080/17405629.2012.686740

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469. Retrieved from http://www.jstor.org/stable/25792024

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for latent gold 4.0: Basic and advanced.* Belmont, MA: Statistical Innovations.

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*, 26-44. doi: 10.1080/00220973.2010.531299

Appendix A. Supplementary Tables

A1. Power and Significance for $D \neq 0$

A2. Mean and SD values for included simulations

Table A1

The proportion simulations for which the correct direct effects were identified, in all levels of N and S for different levels of D, both FIML and LIML, both EPC and BVR. In parenthesis, for every proportion correctly identified, the proportion significant simulations is given.

Condition		$D = 1$				$D = 2$				$D = 3$			
		FIML		LIML		FIML		LIML		FIML		LIML	
N	S	BVR (sig)	EPC (sig)	BVR (sig)	EPC (sig)	BVR (sig)	EPC (sig)	BVR (sig)	EPC (sig)	BVR (sig)	EPC (sig)	BVR (sig)	EPC (sig)
500	Low	.31 (.00)	.36 (.81)	.50 (.94)	.44 (.37)	.19 (.00)	.18 (.07)	.25 (.19)	.25 (.03)	.15 (.00)	.12 (.03)	.11 (.04)	.06 (.00)
	Mid	.77 (.04)	.70 (.80)	.54 (.60)	.41 (.18)	.58 (.00)	.49 (.27)	.28 (.06)	.35 (.01)	.37 (.00)	.33 (.11)	.20 (.01)	.10 (.00)
	High	.92 (.27)	.77 (.84)	.80 (.40)	.63 (.00)	.83 (.04)	.64 (.45)	.69 (.07)	.45 (.00)	.78 (.00)	.57 (.24)	.51 (.00)	.17 (.00)
1000	Low	.51 (.00)	.59 (.82)	.50 (.91)	.40 (.73)	.39 (.00)	.37 (.23)	.24 (.18)	.30 (.16)	.21 (.00)	.23 (.09)	.11 (.06)	.05 (.00)
	Mid	.91 (.15)	.91 (.96)	.67 (.72)	.59 (.14)	.82 (.00)	.79 (.65)	.39 (.10)	.50 (.02)	.62 (.00)	.65 (.43)	.32 (.06)	.14 (.00)
	High	1.00 (.57)	.97 (.97)	.96 (.65)	.86 (.00)	.98 (.29)	.93 (.87)	.91 (.32)	.79 (.00)	.93 (.10)	.83 (.72)	.81 (.16)	.41 (.00)
2000	Low	.69 (.00)	.88 (.95)	.53 (.94)	.47 (.82)	.60 (.00)	.60 (.48)	.18 (.11)	.39 (.26)	.35 (.00)	.38 (.25)	.12 (.05)	.05 (.00)
	Mid	.99 (.52)	.99 (1.00)	.84 (.86)	.82 (.35)	.99 (.17)	.98 (.97)	.64 (.32)	.67 (.07)	.91 (.00)	.92 (.89)	.50 (.16)	.27 (.00)
	High	1.00 (.96)	1.00 (1.00)	1.00 (.94)	.97 (.01)	1.00 (.86)	.99 (.99)	.99 (.81)	.96 (.00)	1.00 (.67)	.99 (.99)	.98 (.70)	.77 (.00)
4000	Low	.91 (.01)	1.00 (1.00)	.64 (.95)	.69 (.89)	.84 (.00)	.91 (.88)	.26 (.17)	.53 (.40)	.58 (.00)	.62 (.58)	.17 (.07)	.12 (.04)
	Mid	1.00 (.96)	1.00 (1.00)	.97 (.96)	.99 (.77)	1.00 (.81)	1.00 (1.00)	.91 (.75)	.92 (.31)	1.00 (.30)	1.00 (1.00)	.82 (.54)	.46 (.06)
	High	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (.45)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (.12)	1.00 (.99)	1.00 (1.00)	1.00 (.98)	.96(.02)

Note. Significance means significant under a χ^2-distribution with $df = 1$ (BVR) or $df = 3$ (EPC) and $\alpha = .05$.

Table A2

Mean and standard deviation values for all included simulations, for all levels of N, S and D, for both FIML and LIML.

Condition		$D = 0$				$D = 1$			
		FIML		LIML		FIML		LIML	
N	S	BVR	EPC	BVR	EPC	BVR	EPC	BVR	EPC
500	Low	0.12 (0.20)	6.51 (53.21)	2.90 (4.15)	3.33 (27.16)	0.67 (0.50)	55.14 (197.08)	17.95 (9.98)	14.63 (57.03)
	Mid	0.15 (0.23)	3.36 (11.16)	0.96 (1.64)	2.95 (34.53)	1.50 (1.00)	14.43 (9.07)	5.93 (4.42)	7.87 (22.72)
	High	0.15 (0.24)	2.97 (3.97)	0.31 (0.51)	0.64 (2.10)	2.88 (1.72)	15.23 (22.44)	3.84 (2.51)	2.03 (0.86)
1000	Low	0.11 (0.17)	3.59 (17.28)	3.47 (5.34)	7.60 (83.62)	0.69 (0.44)	92.23 (866.26)	21.93 (16.86)	33.76 (145.36)
	Mid	0.14 (0.22)	2.76 (3.23)	0.85 (1.44)	2.59 (31.59)	2.47 (1.42)	20.21 (9.41)	6.87 (4.53)	5.44 (3.55)
	High	0.15 (0.23)	2.69 (2.93)	0.29 (0.48)	0.60 (0.76)	4.68 (2.39)	20.51 (8.31)	5.62 (3.19)	2.69 (1.17)
2000	Low	0.10 (0.16)	2.87 (3.99)	3.23 (5.21)	23.41 (493.49)	0.92 (0.58)	63.54 (618.98)	17.68 (15.01)	23.28 (39.77)
	Mid	0.14 (0.23)	2.64 (2.70)	0.80 (1.32)	1.61 (2.12)	4.19 (1.93)	33.99 (12.54)	9.78 (5.89)	7.37 (3.47)
	High	0.15 (0.24)	2.59 (2.69)	0.29 (0.49)	0.57 (0.62)	8.93 (3.31)	36.92 (11.15)	9.98 (4.57)	4.20 (1.39)
4000	Low	0.10 (0.15)	2.63 (2.82)	2.68 (4.36)	7.75 (67.08)	1.45 (0.78)	42.41 (25.68)	18.54 (12.99)	21.18 (18.66)
	Mid	0.15 (0.23)	2.68 (2.58)	0.82 (1.33)	1.61 (1.98)	8.51 (3.05)	66.87 (18.50)	16.56 (9.29)	11.45 (4.31)
	High	0.14 (0.24)	2.54 (2.52)	0.28 (0.49)	0.55 (0.60)	17.69 (5.04)	71.33 (15.72)	19.35 (6.72)	7.63 (1.89)

Condition		$D = 2$				$D = 3$			
		FIML		LIML		FIML		LIML	
N	S	BVR	EPC	BVR	EPC	BVR	EPC	BVR	EPC
500	Low	0.33 (0.33)	27.71 (110.66)	7.01 (6.23)	9.37 (25.11)	0.36 (0.35)	25.46 (114.85)	6.59 (5.87)	9.00 (22.77)
	Mid	1.05 (0.87)	13.32 (18.47)	3.70 (3.35)	3.37 (2.88)	1.05 (0.88)	13.11 (18.02)	3.71 (3.32)	3.36 (2.90)
	High	2.20 (1.65)	12.34 (7.09)	2.87 (2.23)	1.71 (0.88)	2.20 (1.65)	12.34 (7.09)	2.87 (2.23)	1.71 (0.88)
1000	Low	0.49 (0.42)	57.65 (553.92)	9.92 (9.42)	28.82 (104.72)	0.49 (0.42)	58.90 (561.67)	9.81 (9.43)	30.18 (108.90)
	Mid	1.57 (1.07)	16.60 (9.05)	4.72 (3.58)	4.42 (4.41)	1.57 (1.07)	16.60 (9.05)	4.72 (3.58)	4.42 (4.41)
	High	3.91 (2.21)	19.16 (8.09)	4.81 (2.92)	2.32 (1.08)	3.91 (2.21)	19.16 (8.09)	4.81 (2.92)	2.32 (1.08)
2000	Low	0.74 (0.49)	29.17 (179.88)	8.97 (10.46)	8.62 (9.29)	0.74 (0.49)	29.17 (179.88)	8.97 (10.46)	8.62 (9.29)
	Mid	2.60 (1.42)	26.75 (11.42)	6.39 (4.35)	4.79 (2.42)	2.60 (1.42)	26.75 (11.42)	6.39 (4.35)	4.79 (2.42)
	High	7.32 (3.12)	34.25 (11.02)	8.55 (4.16)	3.67 (1.38)	7.32 (3.12)	34.25 (11.02)	8.55 (4.16)	3.67 (1.38)
4000	Low	1.11 (0.75)	31.29 (64.18)	8.85 (8.19)	10.31 (6.30)	1.11 (0.75)	31.29 (64.18)	8.85 (8.19)	10.31 (6.30)
	Mid	4.87 (2.03)	49.04 (16.15)	10.73 (6.33)	7.75 (3.15)	4.87 (2.03)	49.04 (16.15)	10.73 (6.33)	7.75 (3.15)
	High	14.30 (4.45)	65.44 (15.28)	16.44 (6.06)	6.64 (1.94)	14.30 (4.45)	65.47 (15.26)	16.45 (6.05)	6.64 (1.94)

Note. Mean values with SD in parenthesis.