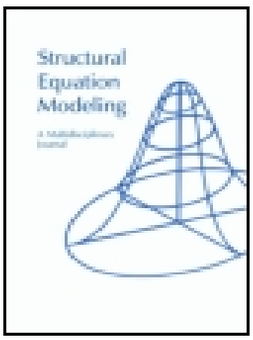

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

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

The Detection and Modeling of Direct Effects in Latent Class Analysis

Jeroen H. M. Janssen, Saskia van Laar, Mark J. de Rooij, Jouni Kuha & Zsuzsa Bakk

To cite this article: Jeroen H. M. Janssen, Saskia van Laar, Mark J. de Rooij, Jouni Kuha & Zsuzsa Bakk (2018): The Detection and Modeling of Direct Effects in Latent Class Analysis, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2018.1541745

To link to this article: https://doi.org/10.1080/10705511.2018.1541745

Published online: 26 Nov 2018.

Structural Equation Modeling: A Multidisciplinary Journal, 00: 1–11, 2018
ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2018.1541745

The Detection and Modeling of Direct Effects in Latent Class Analysis

Jeroen H. M. Janssen,[1] Saskia van Laar,[2] Mark J. de Rooij,[1] Jouni Kuha,[3] and Zsuzsa Bakk[1]

[1]*Leiden University*
[2]*University of Oslo*
[3]*London School of Economics and Political Science*

Several approaches have been proposed for latent class modeling with external variables, including one-step, two-step, and three-step estimators. However, very little is known yet about the performance of these approaches when direct effects of the external variable to the indicators of latent class membership are present. In the current article, we compare those approaches and investigate the consequences of not modeling these direct effects when present, as well as the power of residual and fit statistics to identify such effects. The results of the simulations show that not modeling direct effect can lead to severe parameter bias, especially with a weak measurement model. Both residual and fit statistics can be used to identify such effects, as long as the number and strength of these effects is low and the measurement model is sufficiently strong.

Keywords: direct effects, latent class analysis, mixture models, residual statistics

INTRODUCTION

In the early days of latent class (LC) analysis, the technique was mainly being regarded as a "qualitative data analog to factor analysis" (McCutcheon, 1987, p. 7), which could be used to identify underlying categorical latent variables to explain associations among categorical observed indicators, or, alternatively, identify different subgroups among subjects. While 30 years later this viewpoint is still the most common one, the applications have also expanded. The interest with recent developments is not only in relating the indicators to the LCs (the *measurement model*), but also relating the clustering to a set of external variables, both predictors and distal outcomes (the *structural model*).

Traditionally, the LC model is estimated using a one-step full information maximum likelihood estimation (FIML; McCutcheon, 1987; Vermunt, 2010), a statistically and computationally efficient procedure which uses all the available information in the dataset at once. This efficiency unfortunately comes with a price, namely the forced re-estimation of the entire model if even slight changes are made. Alterations in, for example, the external variables lead to reidentification of the model and thus possible changes in the definition and interpretation of the LCs (Asparouhov & Muthén, 2014; Bakk & Kuha, 2017; Bolck, Croon, & Hagenaars, 2004; Vermunt, 2010).

As an alternative to FIML, methods of three-step estimation have been developed (Asparouhov & Muthén, 2014; Vermunt, 2010). Here, estimation of the model consists of three steps, namely (1) estimating the measurement model, (2) classification of subjects to the LC, and (3) relating the classification variable to external variables. A well-known problem of this approach is that in the second step, irrespective of the assignment rule used, a classification error is introduced, which leads to biased estimates in step 3. To account for the classification error, various so-

Correspondence should be addressed to Jeroen H. M. Janssen, Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Leiden University, P.O. Box 9555, Leiden 2300 RB, The Netherlands. E-mail: j.h.m.janssen@fsw.leidenuniv.nl

called bias-adjusted three step estimators have been developed (Asparouhov & Muthén, 2014; Bakk, Tekle, & Vermunt, 2013; Bolck et al., 2004; Vermunt, 2010). These adjustments solve the problem of classification error, but there are other issues related to this model that cannot be easily solved, such as modeling the direct effect of the external variable on the indicators (see, e.g., Di Mari and Bakk (2018)).

More recently, an alternative to the one-step and three-step methods has been developed, called the two-step approach (Bakk & Kuha, 2017). This model removes the classification step of the three-step method, with the first step estimating the measurement model, and the second step estimating the structural model with the measurement parameters held fixed at their estimates from step one. This approach avoids the explicit classification step of three-step approaches but maintains their computational and interpretational advantages.

If there are external variables and/or covariates present in the LC model, the indicators of the LC are typically assumed to be conditionally independent of these external variables given LC membership. This assumption is in different contexts also known as the assumption of equivalence (or invariance) of measurement, or that of no differential item functioning (DIF; Masyn, 2017; Osterlind & Everson, 2009). If it is violated, the parameter estimates of the LC model can be severely biased (Asparouhov & Muthén, 2014; Masyn, 2017; Mellenbergh, 1989). This is caused by an unmodeled residual association between indicator(s) and external variables (Masyn, 2017). In this way, the external variable has a direct influence on the indicators through the LC (Moustaki, 2003). In the following, we will refer to this as direct effects (DEs). As a consequence, any systematic difference on the observed items may no longer reflect true differences in relation to the latent variable (Masyn, 2017; Millsap, 2011), leading to biased estimates of the overall association between LC membership and the external variable, but also of the measurement model. This consequence states the importance of being knowledgeable about whether there are any DEs in your data, and if there are, knowing how to deal with them.

The assumption of measurement invariance in a model can be relaxed by adding direct paths between the external variable and the concerned indicators (Kankaraš, Moors, & Vermunt, 2010; Masyn, 2017). This allows for the levels of the affected indicators to depend on the external variable. The various methods of estimation discussed above differ in their way in which they can do this. In one-step and two-step models, it is possible to add DEs, as discussed further below. With three-step, in contrast, adding DEs becomes increasingly difficult and often even impossible. When estimating the structural model, only information from the classification variable is used, with the information from the separate indicator variables being lost. Therefore, DEs to specific indicators cannot be added in the third step anymore. Although it is possible to add them in step one, one would have to know the exact structure of these paths beforehand, which often is not the case.

In this article we will focus on two different research questions. First we would like to know the consequences of not modeling such effects on the stability and bias of the parameter estimates (i.e., does it actually matter whether the direct effects are modeled?). This approach originated from the idea of Kuha and Moustaki (2015) who in a multigroup structural equation model framework conducted a sensitivity analysis of not modeling DEs and concluded that there are various conditions in which parameter estimates are rather unaffected by this.

The second part is focused on finding the DEs, should they be there. This part of the article focuses on answering the second research question: how can we identify the DEs that are present in our model? We investigate the performance (i.e., power) of different statistics to see how well they do in identifying DEs. Testing for DEs is common practice in general structural equation modeling (SEM), and in multi-group latent variable models in particular. The article by Kim, Cao, Wang, and Nguyen (2017) gives an overview of models used to test for DEs in an inferential way that can be used when the number of groups is large. The authors mainly focus on model fitting to test for DE, 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 DE. An overview is given by 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., Nagelkerke, Oberski, & Vermunt, 2017; Oberski, van Kollenburg, & Vermunt, 2013; Oberski, Vermunt, & Moors, 2015).

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 DEs in the model. Another possibility 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 statistic is commonly used in econometrics, where it is called the Lagrange multiplier (e.g., Breusch & Pagan, 1980). In the field of SEM, the EPC is related to the modification index (MI). For a description of the MI and differences with the EPC, see Whittaker (2012).

Although residual statistics are used to investigate DEs in FIML by Oberski et al. (2013), the performance of these

statistics is not yet tested in the newly developed two-step estimator; this is one of the main goals of the current article.

This article aims to add to the existing literature on how to deal with the presence of DEs in stepwise LC models. The two-step estimator developed by Bakk and Kuha (2017) in theory overcomes many of the problems present with one-step and three-step estimators, especially in the presence of such DEs. That is why it is interesting to know how this estimator compares to existing methods in dealing with these effects.

The remainder of this article is structured as follows. First, a short description of the general LC model and the various approaches to estimate this model is given. Then, two different simulation studies are performed to answer the research questions. We end with a generalized discussion and conclusion in which we try to give some recommendations on when to use what models and when to model DEs.

THE LATENT CLASS MODELS

In this article we only give a brief overview of the various LC models. For an extended description, we refer to for example McCutcheon (1987), Vermunt (2010), and Bakk and Kuha (2017).

Suppose we have a LC model with 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, $i \in \{1, ..., N\}$. Then define a latent variable X consisting of T different classes, such that $t \in \{1, ..., T\}$. An LC model for $P(\mathbf{Y}_i)$ can be defined as (e.g., McCutcheon, 1987):

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

The indicators are typically assumed to be independent given LC membership, which entails a restriction on the last term in this last equation:

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

with $I(Y_{ik} = r)$ an indicator variable being 1 if subject i has response r on indicator k, and 0 otherwise, and $\{\pi_{ktr}\}$ being the $(K-1)KT$ unique probabilities to be estimated (Bakk, Oberski, & Vermunt, 2014).

This basic LC model can be extended to include an observed covariate vector $\mathbf{Z}_i$. We then get a model for $P(\mathbf{Y}_i | \mathbf{Z}_i)$ (Vermunt, 2010):

$$P(\mathbf{Y}_i | \mathbf{Z}_i) = \sum_{t=1}^{T} P(X = t | \mathbf{Z}_i) P(\mathbf{Y}_i | X = t). \tag{3}$$

In case of a covariate model, the assumption of invariance of measurement or lack of DEs entails the independence of the indicators $\mathbf{Y}_i$ and the covariate vector $\mathbf{Z}_i$ given the LC X. This assumption is the main focus of the article. Whereas in the current article a single binary covariate will be used, this can be easily extended to include multiple covariates, either binary, categorical or continuous.

Methods of estimation

The one-step FIML approach estimates the LC model defined in Equation (3) by maximizing the log-likelihood function $\mathcal{L}_1$ for $P(\mathbf{Y}_i | \mathbf{Z}_i)$ (Vermunt, 2010):

$$\begin{aligned} \mathcal{L}_1 &= \sum_{i=1}^{N} \log P(\mathbf{Y}_i | \mathbf{Z}_i) \\ &= \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} P(X = t | \mathbf{Z}_i) \prod_{k=1}^{K} \prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)} \right], \end{aligned} \tag{4}$$

where the conditional LC probabilities $P(X = t | \mathbf{Z}_i)$ are parameterized using the multinomial logit

$$P(X = t | \mathbf{Z}_i) = \frac{\exp(\alpha_t + \sum_{q=1}^{Q} \beta_{qt} Z_{iq})}{\sum_{s=1}^{T} \exp(\alpha_t + \sum_{q=1}^{Q} \beta_{qt} Z_{iq})}. \tag{5}$$

with Z_{iq} one of Q covariates.

As compared to the one-step method, the stepwise approaches begin with a more limited LC model excluding the external variables. The first step in the two-step method therefore estimates the measurement probabilities π_{ktr} and the marginal LC probabilities $\xi_t = P(X = t)$ by means of maximizing a log-likelihood function $\mathcal{L}_{2(1)}$ for $P(\mathbf{Y}_i)$ (Bakk & Kuha, 2017):

$$\begin{aligned} \mathcal{L}_{2(1)} &= \sum_{i=1}^{N} \log P(\mathbf{Y}_i) \\ &= \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} \xi_t \prod_{k=1}^{K} \prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)} \right], \end{aligned} \tag{6}$$

The parameter estimates $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\pi}}$ are collected in a parameter vector $\hat{\boldsymbol{\theta}}_1$. Note that in a covariate-only model the LC probability vector $\hat{\boldsymbol{\xi}}$ is discarded in the estimation of the second step. The second step then consists of fixing the measurement parameters to the sample estimate from the first step ($\hat{\boldsymbol{\theta}}_1$) and relating this estimate to external variables. This second step is defined as

$$P(\mathbf{Y}_i|X=t,\mathbf{Z}_i) = \underbrace{P(X=t|\mathbf{Z}_i)}_{\text{free}}\underbrace{P(\mathbf{Y}_i|X=t)}_{\text{fixed}}, \tag{7}$$

and is estimated using a second-step log-likelihood function for $\boldsymbol{\theta}_2$:

$$\mathcal{L}_{2(2)}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1) = \sum_{i=1}^{N}\log\sum_{t=1}^{T}\underbrace{P(X=t|\mathbf{Z}_i)}_{\text{free}}\underbrace{P(\mathbf{Y}_i|X=t)}_{\text{fixed}}. \tag{8}$$

Using a similar logit model as defined in Equation (5), this yields the $T-1$ parameter estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, which are the step two estimates collected in the parameter vector $\hat{\boldsymbol{\theta}}_2$.

In this second step, standard errors of the step two estimates do not take into account the uncertainty of the step one estimates, because they are fixed. A way of estimating these standard errors to take into account uncertainty from both steps is discussed in Bakk and Kuha (2017).

In the three-step models, contrary to the one-step and two-step approaches, a classification step is involved. The first step consists of maximizing the log-likelihood function of Equation (6). In the subsequent second step, subjects are assigned to one of the classes, using Bayes theorem for obtaining the posterior class membership probabilities:

$$P(X=t|\mathbf{Y}_i) = \frac{P(X=t)P(\mathbf{Y}_i|X=t)}{P(\mathbf{Y}_i)}. \tag{9}$$

Assignment to one of the classes subsequently can be done by different assignment rules. The easiest of them is called modal assignment, which assigns a subject to the class for which the posterior class probability is highest. We denote the posterior class membership by W. The problem with the three-step approach is that in the second step, a classification error is made which can be defined as the probability of belonging to class $X=t$ while assigned to class $W=s\neq t$. See Vermunt (2010) for the mathematical details. The third step then relates W to the external variables; that is, it uses (Bakk et al., 2014; Bolck et al., 2004; Vermunt, 2010),

$$\mathcal{L}_{3(3)} = \sum_{i=1}^{N}\sum_{s=1}^{T}P(W=s|\mathbf{Y}_i) \log\sum_{t=1}^{T}P(X=t|\mathbf{Z}_i)P(W=s|X=t). \tag{10}$$

Several solutions have been proposed to account for this error, two of which will be described here. The first of them is called the BCH-approach, in which the third-step correction is based on the use of a weighted version of the estimated class membership, weighted by the inverse of the classification error (Bakk et al., 2013; Bolck et al., 2004). A second proposition is the ML-approach, as proposed by Vermunt (2010), which involves re-estimation of the LC model in the third step. Here, the estimated LC membership W of the second step is used as the only indicator of the latent variable with known classification error (Bakk et al., 2013; Vermunt, 2010). For the purpose of comparison, we will also use the uncorrected three-step method. This uncorrected approach uses W directly, without correcting for the known classification error, as such leading to biased estimates of the step three model. Although it is known that it will lead to biased estimates (Bolck et al., 2004; Vermunt, 2010), it is still a popular approach among researchers.

Identification and modeling of direct effects using one-step and two-step estimators

If we would like to add the DEs in our LC model, either one-step or two-step estimation can be used. Adding the DE to the one-step estimator requires the complete model to be re-estimated. However, adding those extra parameters will change the model estimations and can possibly change the definition of the LCs. In addition to that, re-estimation can take a lot of time and introduces a hurdle for researchers. Two-step estimators can also handle modeling of DEs, without the aforementioned problems. Since measurement and structural model estimation is performed separately, only those paths need to be re-estimated that have a DE. Keeping most of the measurement model fixed helps maintaining a more robust model, and decreases estimation time.

The modeling of DEs entails a relaxation of the conditional probabilities $P(\mathbf{Y}_i|X=t)$, because now $\mathbf{Y}_i$ is not only dependent on the latent variable anymore, but also on the covariate. Therefore, the LC model of Equation (3) with a single covariate Z_i now becomes

$$P(\mathbf{Y}_i|X=t,\mathbf{Z}_i) = \sum_{t=1}^{T}P(X=t|\mathbf{Z}_i)P(\mathbf{Y}_i|X=t,\mathbf{Z}_i). \tag{11}$$

While in two-step estimation this means changes in the parameters that are fixed to their first-step estimates, this only changes the estimates of the indicators on which a DE is added. The unaffected indicators keep their step one estimates.

The last term in Equation (11) is parameterized using a multinomial logit similar to Equation (5):

$$P(\mathbf{Y}_i|X=t,\mathbf{Z}_i)=\frac{\exp(\alpha_t+\beta_t X_i+\sum_{q=1}^{q}\gamma_{qt}Z_{iq})}{\sum_{s=1}^{T}\exp(\alpha_t+\beta_t X_i+\sum_{q=1}^{q}\gamma_{qt}Z_{iq})}. \quad (12)$$

While Equation (12) shows how DEs can be modeled when they are cluster-independent, in some instances the DE is present only in one (some) of the classes, not in the others. In that case a cluster-specific DE can be modeled, by adjusting Equation (12):

$$P(\mathbf{Y}_i|X=t,\mathbf{Z}_i)=\frac{\exp(\alpha_t+\beta_t X_i+\sum_{q=1}^{q}\gamma_{qt}Z_{iq}|X)}{\sum_{s=1}^{T}\exp(\alpha_t+\beta_t X_i+\sum_{q=1}^{q}\gamma_{qt}Z_{iq}|X)}. \quad (13)$$

We can only add the DE if we know which indicator(s) are affected. To know this, the DE needs to be identified. To this end, there are several options available, three of which will be discussed here.

In the following we present three approaches for identifying DEs, namely BVR, EPC and the Wald test. While the first two are residual statistics, the Wald is an inferential method.

Bivariate Residual. The BVR (Vermunt & Magidson, 2005, pp. 72–73) 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 z and y_j, both having values 0 or 1, it is defined as:

$$\text{BVR}_j=\sum_{k\in\{0,1\}}\sum_{l\in\{0,1\}}\frac{(n_{kl}-\hat{\mu}_{kl})^2}{\hat{\mu}_{kl}}, \quad (14)$$

where n_{kl} and $\hat{\mu}_{kl}$ equal the observed and expected frequencies in the 2×2 cross-table of z by y_j, respectively. As a value for the BVR for every pair of variables is given as output in standard software such as Latent GOLD (Vermunt & Magidson, 2005), this is an elegant way of locally examining whether paths should be added. While the distribution of this statistic is not defined (Nagelkerke et al., 2017), in practice it is used assuming a χ^2-distribution with $df=1$.

Expected Parameter Change. The EPC statistic (Oberski et al., 2013; Oberski & Vermunt, 2014) is a well-known residual statistic in the context of item response theory (Glas, 1999) and SEM (e.g., Oberski, 2014; Saris, Satorra, & Sörbom, 1987). 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 z and y_j, it is defined as (Oberski et al., 2013):

$$\text{EPC}_j=\frac{s_j^2}{\text{Var}(s_j)}, \quad (15)$$

with $s_j=\frac{\partial\mathcal{L}(\boldsymbol{\theta})}{\partial\psi_j}$ as a value for the 'score' in a local dependence test. In this last definition, $\mathcal{L}(\boldsymbol{\theta})$ is the log-likelihood for a model which allows for the DE between z and y_j, and ψ_j is the parameter corresponding to this effect. See Oberski et al. (2013) for a detailed definition and a discussion of the relationship between the BVR and the EPC. The EPC follows a χ^2-distribution with $df=1$ in the cluster-independent DEs, and $df=T$ in the cluster-specific DEs, with T the number of classes.

Wald Test. In addition to the two residual statistics we also use an inferential method to test for the question whether the concerned direct path coefficient value equals 0. This is done by dividing the ML coefficient estimate by its standard error (Agresti, 2002). The squared test statistic z^2 has an approximate χ^2-distribution under the null, with $df=1$ in the cluster-independent DEs, and $df=T$ for the cluster-specific effects. The Wald is the only one of the statistics used in this study that actually requires fitting the model with the DE(s).

SIMULATION STUDIES

A Monte Carlo simulation study was conducted to investigate the performance of both the different estimators (one-step, two-step and (bias-adjusted) three-step) and statistics (BVR, EPC and Wald). The population model consists of a single three-class latent variable X, six observed binary indicators $\mathbf{Y}=(Y_1,\ldots,Y_6)$ and a single binary covariate Z. Figure 1 shows a graphical representation of the population model. 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 were made in a number of parameters, while others were kept constant due to computational considerations. First of all, the probability of giving a positive response is varied ($S\in\{.80,.90\}$). These values correspond to an entropy-based pseudo-R^2 value of .65 and .90 and a 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).

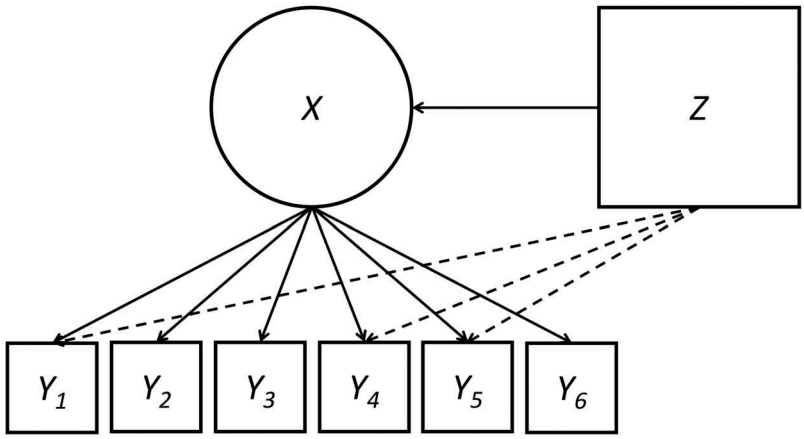

FIGURE 1 A graphical representation of the LC population model, where $Y_1, ..., Y_6$ represent the observed binary indicators, X the LC variable, and Z the observed binary covariate. The dashed lines represent the DEs of which one, two or all will be added to the model.

In the conditions with the lower entropy values, we expect the stepwise estimators to be biased, since information from the covariate is needed in order to estimate the LC variable correctly. When the entropy (i.e., separation) increases, the measurement model becomes powerful enough on its own. Second, the sample size is varied ($N \in \{500, 1000, 2000, 4000\}$). It was shown by Oberski et al. (2013) that both BVR and EPC are large-sample statistics, so it is interesting to see how they perform in relatively small sample sizes as well. The choices for the levels of S and N were slightly adapted from Bakk and Kuha (2017). These authors include a low separation condition of $S = .70$ and do not include the high sample size condition $N = 4000$. Since in the current article the model will become increasingly complex, a more stable measurement model with a higher sample size is preferred. A sample size of 500 can be regarded as a minimal sample size for LC models (Vermunt, 2010).

In terms of the DEs we varied the number ($D \in \{1, 2, 3\}$) and strength ($\gamma \in \{0.4, 0.7\}$, corresponding to a medium and strong effect). The reason for this approach is that Asparouhov and Muthén (2014) showed an increase of bias in FIML models when the number of unmodeled DEs increases. For this reason we would like to know how the statistics perform under more difficult conditions. In our simulations, the DEs are modeled on Y_1 for $D = 1$, (Y_1, Y_4) for $D = 2$, and (Y_1, Y_4, Y_5) for $D = 3$ (see Figure 1). The DEs were either cluster-specific (on class 3) or cluster-non-specific (called the *general* condition). The population values used in the simulation to asses the $Z - X$ association (see Equation (5)) were $\alpha_1 = 0.4$ and $\alpha_2 = -0.6$, $\beta_1 = -1.0$ and $\beta_2 = 1.0$. For the cluster-specific DE, the covariate paths are $\gamma_1 = \gamma_2 = 0$ and $\gamma_3 \in \{0.4, 0.7\}$, with the index referring to the class, modeled on the indicators as mentioned above. For the general DE, a value of $\gamma \in \{0.4, 0.7\}$ is used.

In order to perform the simulation studies, we used the computer software Latent GOLD version 5.1.0.17306 (Vermunt & Magidson, 2005), and RStudio version 1.1.442 (R Core Team, 2015). The script was written in RStudio, which called upon Latent GOLD. Data generation and model estimation is done by Latent GOLD. The output is stored in csv-files, which are imported in RStudio, where further analysis of the results was performed. Both BVR and EPC can be manually added to the Latent GOLD output by asking for 'bvr' and 'scoretest' in the input syntax file, respectively.

Study 1

The first simulation study focuses on the consequences on stability and correctness of parameter estimates when not modeling DEs. To this end, we compared seven methods, all described above: one-step (yes and no), two-step (yes and no), with yes/no corresponding to modeling and not modeling the DEs, respectively, and three-step with either no bias correction (*none*), BCH or ML correction. The seven methods were compared using 4 (N) $\times$ 2 (S) = 8 data conditions for the 3 (D) $\times$ 2 (γ) = 6 DE conditions, both general and cluster-specific. These 96 conditions were all replicated 500 times, to correspond with previous literature (Bakk & Kuha, 2017; Bakk et al., 2013). Comparisons were made for parameter bias (the average deviation from the true parameter value) and coverage (using corrected standard errors as proposed by Bakk and Kuha, 2017). The mean absolute bias is computed as

$$\text{Bias} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{2}(|\beta_1 - \hat{\beta}_{1j}| + |\beta_2 - \hat{\beta}_{2j}|), \tag{16}$$

with $j = 1, ..., J$ the simulations for a given condition. Coverage is defined as the "proportion of replications for

which the 95% confidence interval contains the true parameter value" (Muthén & Muthén, 2002, p. 606).

Results

First of all, in some of the simulations, one or more convergence problems occurred. This means that after the maximum number of iterations, no final solution was found. This resulted in for example an error message about non-convergence in the Latent GOLD output file or extremely large standard errors (>10). These conditions these problems occurred in were exclusively in the general DE condition, for one-step and two-step and for low sample sizes ($N \leq 1000$). Table 1 gives an overview of the percentages of these exclusions. Since the same datasets were used for Study 1 and Study 2, estimation problems in either resulted in exclusion from all analyses.

Table 2 shows for all seven estimators the mean absolute bias and coverage, averaged over the two $Z - X$-parameters $\beta_1 = -1.0$ and $\beta_2 = 1.0$, and averaged over the 8 $N \times S$ data conditions. Standard error correction has been done for the two-step models (and thus for the coverage values). In order to get an idea of the range of the bias values, Table 3 show the values for the conditions in which bias is considered to be the lowest ($N = 4000, S = .90\,(\text{high}), \gamma = 0.4$), referred to as 'best,' and the 'worst' condition ($N = 500, S = .80\,(\text{medium}), \gamma = 0.7$), where bias is expected to be highest.

With respect to bias, the first thing that can be seen from Table 2 is that the one-step and two-step methods are performing better than the three-step estimator, with a slight (but minimal) preference for the one-step method. If we first look at the cluster-specific case (the bottom half of the table) for one-step and two-step, we can see that the bias increases when not modeling the DE, but this increase seems to be rather stable over levels of γ and D. If we increase the number and strength of the DEs, bias still stays rather low, both when we model them (.01 in all cases) and when we don't (with a maximum of .05 for two-step). For this part only, there seems to be no particular preference for one-step or two-step. Coverage values also almost always reach the nominal value of .95 (range between .94 and .96).

Three-step estimators are generally performing worse than one-step and two-step in case of cluster-specific DEs, although the differences are a lot less pronounced as compared to the cases with general DE. Maximum bias values for three-step are .08 when there is no correction for classification error, with coverage values ranging from .80 to .94.

When the DEs are general as compared to cluster specific, the estimators seem to be in more trouble. In case the DEs are modeled, in both situations the parameters are rather unbiased (Bias $\leq$ .02). If DEs are not modeled, however, bias increases. For example, in the condition with two small DEs, not modeling these DEs means an increase in bias from .01 to .08 for the one-step. This effect is even stronger when we have strong DEs, where we for example see an increase in bias from .01 to .28 for three strong effects. The two-step estimator shows similar results. The expected increase with stronger DEs is more pronounced here as compared to the previous situation. Also, especially in the $\gamma = 0.7$ conditions, coverage decreases by a large amount when not modeling the DEs. With three strong DEs, coverage decreases to a

TABLE 1
Percentage of Excluded Simulations Due to Nonconvergence in the General DE Conditions

	$\gamma = 0.4$			$\gamma = 0.7$		
	D = 1	*D = 2*	*D = 3*	*D = 1*	*D = 2*	*D = 3*
1STEP	1.6	2.0	3.6	1.8	3.6	3.8
1STEPno	1.2	1.8	1.2	2.0	4.6	3.0
2STEP	0	0	0	0.8	0	2.2
2STEPno	0	0	0	0	0	0

Note. 1STEP (*2STEP*) = one-step (two-step) estimation modeling DEs. *1STEPno* (*2STEPno*) = one step (two step) not modeling DEs.

TABLE 2
The Mean Absolute Bias (Coverage) Values Over All Replications, Averaged Over the Two Slope Parameters $\beta_1 = 1$ and $\beta_2 = -1$, and Averaged Over All Eight Data Conditions

General DE

	$\gamma = 0.4$			$\gamma = 0.7$		
	D = 1	*D* = 2	*D* = 3	*D* = 1	*D* = 2	*D* = 3
1STEP	.01 (.95)	.01 (.94)	.01 (.95)	.01 (.95)	.01 (.95)	.01 (.95)
1STEPno	.03 (.94)	.08 (.91)	.15 (.82)	.06 (.92)	.15 (.82)	.28 (.64)
2STEP	.01 (.95)	.01 (.95)	.01 (.95)	.02 (.95)	.01 (.95)	.01 (.95)
2STEPno	.05 (.94)	.07 (.91)	.14 (.93)	.08 (.91)	.12 (.84)	.26 (.66)
3STEPml	.03 (.93)	.06 (.91)	.12 (.85)	.06 (.90)	.10 (.86)	.21 (.71)
3STEPbch	.03 (.86)	.06 (.84)	.12 (.76)	.06 (.83)	.10 (.62)	.21 (.62)
3STEPnone	.22 (.61)	.21 (.61)	.19 (.64)	.24 (.55)	.22 (.59)	.19 (.66)

Cluster-specific DE

	$\gamma = 0.4$			$\gamma = 0.7$		
	D = 1	*D* = 2	*D* = 3	*D* = 1	*D* = 2	*D* = 3
1STEP	.01 (.95)	.01 (.95)	.01 (.95)	.01 (.95)	.01 (.96)	.01 (.95)
1STEPno	.02 (.95)	.02 (.95)	.03 (.94)	.03 (.94)	.04 (.94)	.04 (.94)
2STEP	.01 (.95)	.01 (.95)	.01 (.95)	.01 (.95)	.01 (.96)	.01 (.95)
2STEPno	.03 (.95)	.03 (.94)	.03 (.94)	.05 (.93)	.04 (.94)	.05 (.93)
3STEPml	.02 (.94)	.02 (.94)	.02 (.94)	.03 (.92)	.04 (.93)	.04 (.93)
3STEPbch	.02 (.88)	.02 (.87)	.02 (.87)	.03 (.86)	.04 (.87)	.04 (.86)
3STEPnone	.05 (.90)	.06 (.90)	.06 (.89)	.06 (.88)	.07 (.80)	.08 (.87)

Note. 1STEP (*2STEP*) = one-step (two-step) estimation modeling DEs. *1STEPno* (*2STEPno*) = one step (two step) not modeling DEs. *3STEPml* = three-step estimation using ML correction. *3STEPbch* = three-step estimation using BCH correction. *3STEPno* = three-step estimation without correction.

TABLE 3
The Mean Absolute Bias Values, Averaged over β_1 and β_2 in the "Best" and "Worst" Condition, for All Models, Both General and Cluster Specific

General DE						
	$D = 1$		$D = 2$		$D = 3$	
	Best	Worst	Best	Worst	Best	Worst
1STEP	.00	.03	.00	.03	.00	.04
1STEPno	.02	.07	.04	.24	.06	.46
2STEP	.00	.05	.00	.04	.00	.02
2STEPno	.02	.15	.03	.17	.06	.38
3STEPml	.01	.13	.03	.14	.04	.33
3STEPbch	.01	.13	.03	.14	.04	.33
3STEPnone	.10	.39	.09	.35	.09	.31
Cluster-specific DE						
	$D = 1$		$D = 2$		$D = 3$	
	Best	Worst	Best	Worst	Best	Worst
1STEP	.00	.04	.00	.01	.01	.03
1STEPno	.01	.07	.01	.06	.01	.06
2STEP	.00	.03	.00	.06	.01	.05
2STEPno	.01	.09	.01	.11	.01	.09
3STEPml	.01	.07	.01	.10	.00	.08
3STEPbch	.01	.07	.01	.10	.00	.08
3STEPnone	.01	.15	.01	.20	.01	.23

Note. 1STEP (2STEP) = one-step (two-step) estimation modeling DEs. *1STEPno* (*2STEPno*) = one step (two step) not modeling DEs. *3STEPml* = three-step estimation using ML correction. *3STEPbch* = three-step estimation using BCH correction. *3STEPno* = three-step estimation without correction.

minimum of .64 in one step (as compared to .94 for the similar condition with a cluster-specific DE).

Three-step estimators are again showing worst performance here. The maximum bias for the general DE is .24, accompanied by a coverage level of .55 for the condition with no correction.

What was found in our results and what is summarized in Table 3 is a general increasing trend in bias when model becomes increasingly difficult. This can be seen by taking the difference between the worst and the best condition. This is caused by a weaker measurement model (lower sample size and lower separation) and larger DEs. Estimators that allow for the modeling of DEs are generally doing fine, with minimal bias even with weaker measurement models (Bias as high as .05 for one DE in two-step estimation). In case the DEs are not modeled, the estimators are doing substantially worse (up until .46 for one-step with three DEs), but no clear preference is found for any of the options, as long as for the three-step estimator the bias is corrected.

Study 2

The second simulation study focuses on the performance of inferential and residual statistics to identify DEs. As can be concluded from the first study, in some situations omitting DEs leads to biased estimators. This is especially the case when the measurement model is weak (small separation and small sample size) or when the DEs are strong (or when there are multiple). In such cases, it may be good to add DEs from the covariate to the indicator. However, in order to do so, one would obviously have to know which paths to add. In this study, the BVR, EPC, and Wald test as defined above are used.

The second study only concerns the two estimators in which the DEs can be added, that is, one-step and two-step. Again, the same 96 conditions as in Study 1 were used, with 500 replications. The performance of the methods was measured by looking at the proportion of simulations in which the correct DE was found by the concerned statistic. When there are multiple DEs, we checked for the probability of finding at least one of them and finding all of them. This approach is adapted from Nagelkerke et al. (2017). This can be regarded as a form of power analysis, since it entails the probability of finding an effect if it is present. A DE is called 'identified' when the corresponding residual association (BVR) or direct path (EPC/Wald) was significant, at the significance level of $\alpha = .05$ and critical values as defined above. For BVR/EPC, aside from being significant, the corresponding association(s) needed to be the largest (or largest two/three) of all the possible $Z - \mathbf{Y}$ effects as well.

Results

Table 4 gives an overview of the power of the various estimators and statistics in correctly identifying either one or all of the DEs, when present in the population. These results are averaged over the sample size and separation conditions.

In Table 4 we can see that the probability of finding one of the effects increases when there are actually multiple DEs (the model gets to choose, in a way) as compared to a single effect, but remains rather constant when comparing two or three effects (e.g., .62, .81, and .79 for BVR in one step for one, two, and three effects, respectively). Performances greatly increase, as can be expected, when the effect is strong (the right-hand side of the table) as compared to medium (the left-hand side). In the same line of reasoning we see that a general DE (i.e., unconditional on the LC; the top half of the table) is much easier identified as compared to a cluster-specific effect (the bottom half). If we stay with our evaluation of BVR in one-step, we see that the previously mentioned probabilities increase to .89, .95, and .93 for a strong DE ($\gamma = 0.7$), whereas in

TABLE 4
Mean (SD) Proportion of Simulations in Which the Correct Number of DEs Were Found by the Various Statistics, Averaged Over All Sample Size and Separation Conditions

	General DE											
	γ = 0.4						γ = 0.7					
	BVR		EPC		Wald		BVR		EPC		Wald	
	1STEP[a]	2STEP[a]	1STEP	2STEP	1STEP	2STEP	1STEP	2STEP	1STEP	2STEP	1STEP	2STEP
D = 1												
All[b]	.62 (.24)	.66 (.23)	.61 (.24)	.61 (.24)	.51 (.31)	.50 (.32)	.89 (.15)	.85 (.18)	.83 (.19)	.83 (.19)	.79 (.26)	.78 (.27)
D = 2												
Min.1[b]	.81 (.17)	.80 (.16)	.80 (.17)	.79 (.17)	.71 (.27)	.72 (.27)	.95 (.07)	.95 (.07)	.95 (.07)	.95 (.07)	.94 (.10)	.94 (.10)
All	.24 (.03)	.29 (.09)	.30 (.09)	.35 (.16)	.37 (.34)	.37 (.34)	.31 (.13)	.36 (.09)	.40 (.07)	.30 (.16)	.73 (.32)	.74 (.31)
D = 3												
Min.1	.79 (.14)	.81 (.12)	.75 (.14)	.80 (.12)	.79 (.22)	.78 (.22)	.93 (.10)	.92 (.09)	.88 (.13)	.91 (.09)	.96 (.07)	.97 (.07)
All	.31 (.02)	.29 (.03)	.31 (.02)	.28 (.03)	.30 (.33)	.29 (.32)	.33 (.01)	.31 (.02)	.32 (.01)	.31 (.02)	.68 (.37)	.68 (.36)
	Cluster-specific DE											
	γ = 0.4						γ = 0.7					
	BVR		EPC		Wald		BVR		EPC		Wald	
	1STEP	2STEP	1STEP	2STEP	1STEP	2STEP	1STEP	2STEP	1STEP	2STEP	1STEP	2STEP
D = 1												
All	.27 (.07)	.23 (.09)	.25 (.07)	.27 (.10)	.07 (.07)	.09 (.09)	.47 (.19)	.41 (.21)	.42 (.19)	.47 (.24)	.23 (.28)	.28 (.30)
D = 2												
Min.1	.50 (.12)	.45 (.12)	.48 (.11)	.49 (.13)	.20 (.18)	.22 (.19)	.74 (.17)	.67 (.19)	.70 (.17)	.72 (.20)	.49 (.37)	.52 (.37)
All	.24 (.11)	.24 (.11)	.24 (.11)	.18 (.02)	.02 (.03)	.02 (.04)	.28 (.10)	.28 (.10)	.31 (.12)	.22 (.11)	.20 (.28)	.23 (.30)
D = 3												
Min.1	.67 (.10)	.64 (.10)	.63 (.10)	.66 (.12)	.31 (.25)	.30 (.24)	.85 (.11)	.79 (.13)	.81 (.11)	.82 (.15)	.60 (.37)	.59 (.37)
All	.25 (.03)	.24 (.04)	.26 (.05)	.26 (.03)	.01 (.01)	.01 (.01)	.27 (.02)	.26 (.03)	.26 (.02)	.30 (.03)	.18 (.27)	.16 (.24)

[a]*1STEP* = one-step modeling. *2STEP* = two-step modeling. [b]*All* = All the correct 1, 2 or 3 DEs were found. *Min.1* = At least one of the 2 or 3 were correctly identified, regardless of which one it was.

the cluster-specific condition they decrease to a maximum of .67 for a medium DE and .85 for a strong DE.

When comparing the various statistics, we see only very small differences between BVR and EPC. Although small deviations are of course present, no structural preference can be found. What is more pronounced, however, is the difference between the two residual statistics and the Wald test statistic. Although the Wald performs similarly (perhaps even slightly worse) in finding one of the effects (around .94 for all statistics in the general DE condition), it greatly outperforms both EPC and BVR in finding all of the effects in the case of multiple strong and general DEs. In contrast, in the cluster-specific case, Wald seems to have the worst performance of all statistics, to a minimum of .01 for finding all three medium effects in the two-step estimation.

Lastly, comparing the power in the two estimators, we conclude that there were no structural patterns in favor of either one-step or two-step.

In order to show in more detail the range of results, Table 5 shows for the conditions that are considered the 'least favorable' (N = 500, S = .80 (medium), γ = 0.4) and 'most favorable' (N = 4000, S = .90 (high), γ = 0.7) the proportion of correctly identified DEs. It should be noted that these values not always correspond to the actual minima and maxima (although usually they do), but to the two conditions that are considered 'worst' and 'best,' respectively.

As compared to Table 4, these results are mostly comparable. It should be noted, however, that identifying multiple DEs turns out to be very difficult for both BVR and EPC, as can be seen from the maximum values these probabilities take (.50 for both in finding two general DEs, and .33 for finding three.) Wald only seems to do a good job (and perfect job in some cases) when we have a strong enough measurement model, indicated by the large differences between best and worst cases (.01 and 1.00 for $D = 3$ in two -step).

CONCLUSION AND DISCUSSION

The current paper focused on the questions of whether in LC models it is necessary to add DEs to our structural model when measurement invariance is violated, and if

TABLE 5
An Overview of the Worst (N = 500, S = .80 (Medium), γ = 0.4) and Best (N = 4000, S = .90 (High), γ = 0.7) Value of the Power of the Various Statistics in the Various Models

	General DE											
	BVR				EPC				Wald			
	1STEP[a]		2STEP[a]		1STEP		2STEP		1STEP		2STEP	
	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best
$D = 1$												
All[b]	.40	1.00	.31	1.00	.33	1.00	.32	1.00	.18	1.00	.16	1.00
$D = 2$												
Min.1[b]	.60	1.00	.58	1.00	.59	1.00	.56	1.00	.41	1.00	.42	1.00
All	.20	.50	.22	.50	.23	.50	.20	.16	.06	1.00	.06	1.00
$D = 3$												
Min.1	.64	1.00	.65	1.00	.57	1.00	.64	1.00	.55	1.00	.52	1.00
All	.28	.33	.24	.33	.28	.33	.25	.33	.02	1.00	.01	1.00
	Cluster-specific DE											
	BVR				EPC				Wald			
	1STEP		2STEP		1STEP		2STEP		1STEP		2STEP	
	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best
$D = 1$												
All	.20	.69	.17	.68	.14	.79	.18	.80	.18	1.00	.16	1.00
$D = 2$												
Min.1	.39	.74	.35	.72	.31	.83	.35	.84	.06	.97	.08	.97
All	.18	.10	.17	.09	.16	.08	.15	1.00	.00	.66	.00	.67
$D = 3$												
Min.1	.57	.98	.54	.96	.55	.96	.55	.98	.10	1.00	.10	1.00
All	.23	.29	.33	.29	.22	.29	.23	.32	.00	.52	.00	.55

[a]*1STEP* = one-step modeling. *2STEP* = two-step modeling. [b]*All* = All the correct 1, 2 or 3 DEs were found. *Min.1* = At least one of the 2 or 3 were correctly identified, regardless of which one it was.

so, how various statistics perform in identifying which paths should be added. To this end, two extensive simulation studies were conducted, comparing different methods under various conditions.

To answer the first question, we can conclude that allowing for measurement non-invariance is mostly needed when the measurement model is weak and/or the DEs strong. The results in Table 2 show that coverage values often reach the desired value of .95, especially when the DE is only on part of the classes. If it is not, however, coverage values tend to drop, often below its nominal values, as low as ~.60 when the effect is strong.

When the strength of the DEs are only small, the measurement model often is strong enough on its own to correctly (i.e., with little bias) estimate the structural parameters. In such cases, we can obtain relatively unbiased estimates without modeling these effects.

If we would be interested in modeling the DEs, however, because the results tend to be biased or because they are of great interest on its own, it is necessary to first know how they can be found. This was the purpose of our second simulation study, where we investigated the power of the BVR, EPC and Wald statistics to identify DEs in datasets simulated from population models in which these effects were present.

In terms of the model to be used, both models seem to do a comparable job in identifying the DEs. Given the advantages of the two-step method of estimation over the broadly used one-step approach (as discussed in Bakk & Kuha, 2017), we would advice to use the two-step method.

Answering the question what statistic to use requires an answer that is a little more elaborate, since it depends more on the situation. If one has the idea that a single direct path might be present, given that path being unconditional on the LC, it does not matter what statistic is used. Multiple general DEs, however, were most reliably detected by the Wald test statistic in our simulations. If the DEs were conditional on one of the LCs, on the other hand, they were less well detected by the Wald statistics than by the BVR and EPC statistics, with EPC showing slightly higher power levels in most cases.

Although the current article gives more insight in the modeling and identification of DEs when using LC models, there are of course some limitations associated with our findings. First of all, since there are a lot of variables involved, we have kept our LC model as simple as possible.

A single latent variable was used, with only binary indicators. This could be extended to a model with for example ordinal or even continuous indicators. The (single) covariate was also binary here, but all of the methods can also accommodate multiple covariates of different types.

Second, in terms of identifying the DEs, we use the Wald statistic as one of the options. Although this statistic is asymptotically equivalent to the likelihood ratio (LR) test, in small samples the LR tends to outperform the Wald (Agresti, 2002). In LC models, it is common practice to use the Wald test in this context, however, so we have also focused on it here. Further research has to reveal whether these performance differences between Wald and LR also apply for these kind of questions.

ORCID

Jeroen H. M. Janssen http://orcid.org/0000-0003-3641-1517
Saskia van Laar http://orcid.org/0000-0003-4077-5567
Mark J. de Rooij http://orcid.org/0000-0001-7308-6210
Jouni Kuha http://orcid.org/0000-0002-1156-8465
Zsuzsa Bakk http://orcid.org/0000-0001-9352-4812

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: Threestep approaches using MPlus. *Structural Equation Modeling*, *21*, 329–341. doi:10.1080/10705511.2014.915181

Bakk, Z., & Kuha, J. (2017). Two-step estimation of models between latent classes and external variables. *Psychometrika*. Advance online publicaiton. doi:10.1007/s11336-017-9592-7

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, *22*(4), 520–540. doi:10.1093/pan/mpu003

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. doi:10.1093/pan/mph001

Breusch, T., & Pagan, A. (1980). The Lagrange multiplier test and its applications to model specifications in econometrics. *Review of Economic Studies*, *47*, 239–253. doi:10.2307/2297111

Di Mari, R., & Bakk, Z. (2018). Mostly harmless direct effects: A comparison of different latent Markov modeling approaches. *Structural Equation Modeling*, *25*(3), 467–483. doi:10.1080/10705511.2017.1387860

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

Kankaraš, M., Moors, G., & Vermunt, J. (2010). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *The oxford handbook of innovation* (pp. 359–384). Oxford, United Kingdom: Routledge.

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*, *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/a0037802

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

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

Mellenbergh, G. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143. doi:10.1016/0883-0355(89)90002-5

Millsap, R. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, *29*, 81–117.

Muthén, L. K., & Muthén, B. O. (2002). How to use Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620. doi:10.1207/S15328007SEM0904_8

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*, *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: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*. Submitted article. Retrieved from http://daob.nl/wp-content/uploads/2014/07/lca-epc-revision2.pdf

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.

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

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. doi:10.1093/pan/mpq025

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