Assessing Model Fit in Latent Class Analysis when Asymptotics do not hold

Geert H. van Kollenburg * Joris Mulder*
Jeroen K. Vermunt *

Abstract

The application of latent class (LC) analysis involves evaluating the LC model using goodness-of-fit statistics. To assess the misfit of a specified model, say with the Pearson chi-squared statistic, a p-value can be obtained using an asymptotic reference distribution. However, asymptotic p-values are not valid when the sample size is not large and/or the analysed contingency table is sparse. Another problem is that for various other conceivable global and local fit measures, asymptotic distributions are not readily available. An alternative way to obtain the p-value for the statistic of interest is by constructing its empirical reference distribution using resampling techniques such as the parametric bootstrap or the posterior predictive check (PPC). In the current paper, we show how to apply the parametric bootstrap and two versions of the PPC to obtain empirical p-values for a number of commonly used global and local fit statistics within the context of LC analysis. The main difference between the PPC using test statistics and the parametric bootstrap is that the former takes into account parameter uncertainty. The PPC using discrepancies has the advantage that it is computationally much less intensive than the other two resampling methods.

In a Monte Carlo study we evaluated Type I error rates and power of these resampling methods when used for global and local goodness-of-fit testing in LC analysis. Results show that both the bootstrap

*Department of Methodology and Statistics, Tilburg University, the Netherlands.

and the PPC using test statistics are generally good alternatives to asymptotic p-values and can also be used when (asymptotic) distributions are not known. Nominal Type I error rates were not met when sample size was small and the contingency table has many cells. Overall the PPC using test statistics was somewhat more conservative than the parametric bootstrap. We have also replicated previous research suggesting that the Pearson X^2 statistic should in many cases be preferred over the likelihood-ratio G^2 statistic. Power to reject a model for which the number of LCs was 1 less than in the population was very high, unless sample size was small. When the contingency tables are very sparse, the $TBVR$ statistic, which is based on bivariate relationships, still had very high power, signifying its usefulness in assessing model fit.

Key words: Goodness-of-Fit, Posterior Predictive Check, Parametric Bootstrap, Latent Class Analysis

1 Introduction

The use of latent class (LC) models is becoming more and more widespread in a broad range of fields, such as in biomedical sciences (Rindskopf, 2002), psychiatry (Roedelof, Bongers, & van Nieuwenhuizen, 2013), abnormal psychology (Crow et al., 2012) developmental psychology (Laudy et al., 2005), gambling studies (Dufour, Brunelle, & Roy, 2013) and marketing (Okazaki, Campo, Andreu, & Romero, 2014). This makes the availability of reliable methods to assess the goodness-of-fit of LC models increasingly important (Lanza, Flaherty, & Collins, 2004).

The global or overall goodness-of-fit of a LC model is typically assessed using the Pearson or the likelihood-ratio chi-squared statistic (Goodman, 1974). For local fit assessment, which involves checking whether the specified LC model describes specific aspects of the data well, various types of statistics have been proposed, such as residual log-odds-ratios and Pearson statistics computed in two-way tables (Hagenaars, 1988; Magidson & Vermunt, 2004). A convenient way to determine the extent of global or local misfit is to obtain p-values for the goodness-of-fit statistics of interest. Typically, we would get the p-values from the asymptotic distributions of the statistics, but these are not always readily available. Moreover, even when these are available, asymptotic p-values are not useful when the analysed contingency table is too sparse because the sample size is small or the number of cells in the table is large (Haberman, 1988; Langeheine, Pannekoek, & Van de Pol, 1996; Maydeu-Olivares & Joe, 2006; Reiser & Lin, 1999).

P-values can also be obtained using resampling techniques, such as the parametric bootstrap (Efron & Tibshirani, 1993) or the posterior predictive check (PPC) (Meng, 1994; Rubin, 1984). The major benefit of resampling techniques over asymptotics is that we do not need any distributional assumptions regarding the statistics. These methods generate replicated data sets based on the parameter estimates for the specified model, and for each data set they calculate the required statistics. P-values for the statistics are then obtained from their resulting empirical distributions. The main difference between the model-based PPC and the parametric bootstrap is that the former takes into account parameter uncertainty. Another variant of the PPC called the parameter-based PPC has the advantage that it is computationally much less intensive than the other two resampling methods.

While bootstrap methods have been proposed in the context of LC analysis as a way to deal with sparseness when assessing global fit (Langeheine et al., 1996; Von Davier, 1997), they have not been used so far to obtain p-values for statistics for which the distributions are unknown, such as the local fit measures proposed by Magidson and Vermunt (2004). In contrast, PPCs have been used to assess LC model fit using a range of global and local fit measures (Berkhof, Van Mechelen, & Gelman, 2003; Hoijtink, 1998; Ligtvoet & Vermunt, 2012; Meulders, De Boeck, Kuppens, & Van Mechelen, 2002; Rubin & Stern, 1994), but the performance of this approach has not been investigated in a systematic manner.

The purpose of this paper is to discuss and investigate bootstrap and PPC methods in a more integrated manner. This allows expanding the bootstrapping approach to obtain p-values not only in case of sparseness, but also with measures for which the asymptotic distribution is unknown. This allows answering the question as to whether the PPC can be an improvement over the bootstrap when the latter works less well (Von Davier, 1997). More specifically, does taking parameter uncertainty into account yield more reliable p-values when tables are extremely sparse?

The remainder of this paper is organised as follows. Section 2 reviews the LC model and describes a number of commonly used statistics to asses global and local LC model fit. In Section 3 we discuss the various methods to obtain p-values in more detail. Section 4 presents a simulation experiment in which the performance of the investigated methods to obtain p-values is compared. In Section 5 we present an empirical example and finally in Section 6 we discuss the main findings and issues in need of further research.

2 Latent Class Analysis

2.1 The Model

Suppose we have N observations on J categorical items with R_j categories for item number j ($j = 1, \ldots, J$). There are then $S = \prod_{j=1}^{J} R_j$ possible response patterns, which can be denoted as $\mathbf{y}_s = (y_{s1}, \ldots, y_{sJ}), s = 1, \ldots, S$. Letting n_s denote the observed frequency for pattern $\mathbf{y}_s$, the observed data can be summarised as pattern frequencies in $\mathbf{n} = (n_1, \ldots, n_S)$.

The LC model assumes that the N observations can be partitioned into C latent classes, which form the categories of the discrete latent variable ξ (Goodman, 1974). The LCs differ from one another with respect to the conditional response probabilities to the items. Moreover, within each LC the responses to the observed variables are assumed to be independent of one another (i.e., the local independence assumption).

Let ρ_c be the class size (proportion) of LC c and let π_{rjc} be the conditional response probability that a respondent gives response r to item j, given that he or she belongs to LC c. The probability of observing response pattern s is then a mixture of multinomial distributions with weights equal to the class proportion ρ_c. It is given by:

$$P(\mathbf{y}_s) = \sum_{c=1}^{C} \rho_c \prod_{j=1}^{J} \prod_{r=1}^{R_j} \pi_{rjc}^{y^*_{sjr}}, \tag{1}$$

where y^*_{sjr} is 1 if $y_{sj} = r$ and 0 otherwise.

Several methods exist to estimate the LC model parameters $\boldsymbol{\psi} = (\boldsymbol{\rho}, \boldsymbol{\pi})$. One might be interested in obtaining point estimates, interval estimates or posterior probability distributions for the unknown parameters. To obtain their maximum likelihood estimates we typically use the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). We may obtain estimates of their posterior distribution by means of an MCMC algorithm (Tanner & Wong, 1987).

2.2 Goodness-of-Fit Measures

An important part of the model selection procedure in LC modeling involves checking whether a model is in agreement with the data. The discrepancies between observed data and expectations under the model can be assessed using goodness-of-fit (GoF) statistics (Agresti, 2002). We will discuss statistics for the assessment of global and local fit.

Global fit statistics aggregate the disagreement between the observed frequencies n_s and the expected frequencies under the model $e_s = N \cdot P(\mathbf{y}_s)$ into a single value. Well-known chi-squared statistics are the Pearson X^2,

$$X^2(\mathbf{n}) = \sum_{s=1}^{S} \frac{(n_s - e_s)^2}{e_s}, \tag{2}$$

and the likelihood ratio statistic G^2,

$$G^2(\mathbf{n}) = 2 \sum_{s=1}^{S} n_s \ln(n_s/e_s). \tag{3}$$

These two chi-squared statistics belong to the more general family of power divergence statistics which take the form

$$PD(\mathbf{n}) = \frac{2}{\lambda(\lambda+1)} \sum_{s=1}^{S} n_s \left\{ \left(\frac{n_s}{e_s} \right)^{\lambda} - 1 \right\}. \tag{4}$$

The X^2 and G^2 statistics are obtained by setting $\lambda = 1$ and letting λ approach 0, respectively. These two statistics have been shown to be inappropriate when contingency tables are sparse; that is, when a portion of the expected frequencies is small. In such cases, the X^2 statistic tends to become very large, yielding a p-value of 0, while the G^2 statistic tends to be small, yielding a p-value of 1. It has been argued that a good trade-off is found by setting λ equal to 2/3, through which we obtain the Cressie-Read (CR) statistic (Cressie & Read, 1984).

Another global fit measure indicative of how much the observed and estimated cell frequencies differ is the Dissimilarity Index (DI):

$$DI(\mathbf{n}) = \frac{\sum_{s=1}^{S} |n_s - e_s|}{2N}. \tag{5}$$

The DI indicates which proportion of the sample should be moved to another cell to obtain a perfect fit (Vermunt & Magidson, 2013). Though this statistic is appealing due to the information it provides, its asymptotic distribution is unknown. Therefore, to obtain a p-value for this statistics, we need to resort to resampling techniques.

In LC modeling, local fit is typically assessed by computing statistics for lower-order marginals of the analysed J-way contingency table. A popular and very useful measure is the bivariate residual (BVR) statistic, which can be used to determine violations of the local independence assumption (Magidson & Vermunt, 2004; Vermunt & Magidson, 2013). The BVR quantifies the residual association between pairs of items using a Pearson-like chi-squared statistics. To show how the BVR is calculated, let the subscript

r indicate a given response to item j and subscript $r^{'}$ a response to item $j^{'}$. Then $n_{rr'}$ indicates an observed frequency in the two-way cross-tabulation of variables j and $j^{'}$. The expected frequency for this pattern, $e_{rr'}$, can be calculated from the LC model parameters as follows:

$$e_{rr'} = N \sum_{c=1}^{C} \rho_c \pi_{rjc} \pi_{r'j'c}.$$

The BVR for the item pair j-$j^{'}$ is then:

$$BVR_{jj'}(\mathbf{n}) = \sum_{r=1}^{R_j} \sum_{r'=1}^{R_{j'}} \frac{(n_{rr'} - e_{rr'})^2}{e_{rr'}}. \qquad (6)$$

Similar Pearson-like local fit measures may be computed for higher-order tables, for example, for cross-tabulations of three instead of two variables. An important advantage of the BVR statistic compared to global fit measures is that it is much less sensitive to sparseness (Maydeu-Olivares & Joe, 2006). A disadvantage is, however, that its asymptotic distribution is not known, implying that asymptotic p-values are not available.

Based on the BVR, we can derive a global fit measure that may be used as an alternative to the standard GoF chi-squared statistics. This total BVR ($TBVR$) statistic is obtained by summing the BVR statistics across all item pairs, that is,

$$TBVR(\mathbf{n}) = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} BVR_{jj'}(\mathbf{n}). \qquad (7)$$

The main advantage of the $TBVR$ is that is much less affected by sparseness than other global fit measures. However, as for the BVRs themselves, also for the $TBVR$ the asymptotic distribution is unknown. And although knowledge on lower-order fit is very useful, we cannot rule out higher-order misfit, due to multivariate interactions, based on lower-order statistics (Reiser & Lin, 1999).

3 Determining P-values for GoF Measures

3.1 Asymptotic P-values

To test whether a model deviates from the data, most often a p-value is calculated based on an asymptotic reference distribution. If a C-class model is true, the power-divergence statistics asymptotically (as N goes to infinity) follow a chi-squared (χ^2_{df}) distribution with degrees of freedom (df) equal to

$$df = \prod_{j=1}^{J} R_j - C(1 + \sum_{j=1}^{J}(R_j - 1)) \tag{8}$$

(Haberman, 1979; Magidson & Vermunt, 2004). The p-value is then equal to the tail-area probability that a value from the χ^2_{df} distribution is equal to or greater than the computed statistic. If the p-value is less than some a priori set threshold (e.g., .05), the researcher concludes that there is significant misfit between the model and the data (Fisher, 1925).

An important issue related to the use of asymptotic reference distributions is that it is not accurate when the corresponding frequency table is sparse. This occurs when the sample size is not large enough for the contingency table at hand. For example, 10 dichotomous items create a table with $2^{10} = 1024$ cells, which would be considered sparse even with 1000 observations). Sparse tables result in untrustworthy asymptotic p-values (see e.g., Collins, Fidler, Wugalter, & Long, 1993; Langeheine et al., 1996; Magidson & Vermunt, 2004; Von Davier, 1997).

For statistics such as the DI, BVR, and $TBVR$, asymptotic distributions are not known. In some cases rules of thumb are used, but these may not always be accurate. For instance, one rule of thumb says that for dichotomous items BVR values greater than 3.84 indicate significant misfit (3.84 being the 95th percentile of the χ^2_1 distribution). Others take BVR values greater than 1 to indicate misfit. It appears that each cut-off has its downsides and can result in too conservative or too liberal conclusions, depending on the situation (Oberski, van Kollenburg, & Vermunt, 2013). Resampling techniques are therefore required to obtain p-values.

3.2 Parametric Bootstrap

To overcome the problems associated with asymptotic p-values it is possible to obtain empirical reference distributions through resampling methods like the parametric bootstrap (Efron & Tibshirani, 1993), which is used in LC analysis regularly (see e.g., Formann, 2003; Jansen & van der Maas, 1997; Lin, McCulloch, Turnbull, Slate, & Clark, 2000). The parametric bootstrap simulates the probability of finding a value for a statistic T, greater than or equal to the observed value of the statistic $T(\mathbf{n})$, conditional on the ML estimates for the C-class model being the population parameters. The parametric bootstrap p-value for a statistic T is obtained as follows:

Step 1: Find the ML estimates $\hat{\boldsymbol{\psi}}$ for the C-class model (for instance using EM) and calculate the observed fit-statistic $T(\mathbf{n}^{\text{obs}})$. For example, one could use the Pearson X^2 statistic, in which case $T(\mathbf{n}^{\text{obs}}) = X^2(\mathbf{n}^{\text{obs}})$.

Step 2: Calculate the estimated pattern probabilities $\widehat{P}(\mathbf{y}_s)$ from the ML estimates $\hat{\boldsymbol{\psi}}$. Draw B random replicated samples, $\mathbf{n}^{\text{rep}}$, of size N from a multinomial distribution with parameters $\widehat{P}(\mathbf{y}_s)$:

$$\mathbf{n}^{\text{rep},(b)} \sim \text{Multin}(N, \widehat{P}(\mathbf{y}_1), \ldots, \widehat{P}(\mathbf{y}_S)), b = 1, \ldots, B \tag{9}$$

Step 3: Determine the empirical reference distribution of the statistic $T(\mathbf{n})$. That is, find the ML estimates for each data set $\mathbf{n}^{\text{rep},(b)}$ and calculate $T(\mathbf{n}^{\text{rep},(b)})$. For instance, calculate $X^2(\mathbf{n}^{\text{rep},(b)})$ (and/or other statistics of interest).

Step 4: Estimate the bootstrap p-value by the proportion of $T(\mathbf{n}^{\text{rep},(b)})$ which are greater than, or equal to $T(\mathbf{n})$ (which was calculated in Step 1):

$$\hat{p}_{\text{boot}} = B^{-1} \sum_{b=1}^{B} I(T(\mathbf{n}^{\text{rep},(b)}) \geq T(\mathbf{n}^{\text{obs}})), \tag{10}$$

where the indicator function I equals 1 if $T(\mathbf{n}^{\text{rep},(b)}) \geq T(\mathbf{n}^{\text{obs}})$ and 0 otherwise. If $\hat{p}_{\text{boot}}$ is less than a predefined value (for instance, .05) we conclude that the model does not fit the data properly.

Langeheine et al. (1996) showed that the parametric bootstrap method works well with global chi-squared statistics for small well-filled contingency tables. However, Von Davier (1997) showed that in sparse contingency tables with many cells, different conclusions about LC model fit might be obtained depending on which statistic is used. Bootstrap p-values for the G^2 statistics were shown to lead to conservative results, while p-values for the Pearson's X^2 and CR statistics did not fail systematically. Thus, although we can obtain empirical distributions for any statistic, sparseness can still have an effect on how reliable the resulting p-values are, depending on the statistic that is used.

The bootstrap has not been used so far to obtain p-values for GoF measures for which asymptotic p-values are not available, such as the DI, BVR, and $TBVR$ statistics. Whether the bootstrap is suitable for use with these measures has yet to be determined.

3.3 PPC using Test Statistics

In the parametric bootstrap, each of the B replicated data sets is generated using the same ML estimates as if it were population parameter values, implying that the uncertainty about these estimates is not taken into account.

Within the Bayesian framework, parameter uncertainty is incorporated in the posterior distribution. The PPC can be seen as the Bayesian counterpart of the parametric bootstrap which makes use of this posterior distribution.

Two versions of the PPC exist. A PPC using test statistics and a PPC using discrepancies. The PPC using test statistics which was used in LC analysis by Rubin and Stern (1994), is very similar to the parametric bootstrap. It generates a large number of replicated data sets, reestimates the LC model for each data set, and calculates the statistics of interest. The only difference is that the PPC using test statistics uses parameter draws from their posterior distributions as population values to sample the replicated data sets, rather than fixing the parameters to their ML estimates. The PPC using discrepancies (Gelman, Meng, & Stern, 1996) does not require reestimating the LC model for each replicated data set. Instead, it compares both the observed and replicated data directly to the parameter values sampled from their posterior distribution. The PPC using discrepancies will be discussed in detail in the next subsection.

In LC analysis, the PPC using test statistics to obtain a p-value for a statistic $T(\mathbf{n})$ (which is based on ML estimates) proceeds as follows:

Step 1: Find the ML estimates $\hat{\boldsymbol{\psi}}$ for the C-class model (for instance using EM) and calculate the observed fit-statistic $T(\mathbf{n}^{\text{obs}})$. For example, one could use the Pearson X^2 statistic, in which case $T(\mathbf{n}^{\text{obs}}) = X^2(\mathbf{n}^{\text{obs}})$.

Step 2: Obtain K draws $\boldsymbol{\psi}^{(k)}$ from the posterior distribution for the C-class model:

$$\boldsymbol{\psi}^{(k)} \sim p(\boldsymbol{\psi}|\mathbf{n}^{\text{obs}}), k = 1, \ldots, K. \tag{11}$$

This can be done using an MCMC algorithm (Rubin & Stern, 1994).

Step 3: Calculate the estimated pattern probabilities $\widehat{P}(\mathbf{y}_s)^{(k)}$ from $\boldsymbol{\psi}^{(k)}$. Draw K random samples of size N from a multinomial distribution with parameters $\widehat{P}(\mathbf{y}_s)^{(k)}$

$$\mathbf{n}^{\text{rep},(k)} \sim \text{Multin}(N, \widehat{P}(\mathbf{y}_1)^{(k)}, \ldots, \widehat{P}(\mathbf{y}_S)^{(k)}) \tag{12}$$

Step 4: Obtain the ML estimates (e.g., using the EM algorithm) for each data set $\mathbf{n}^{\text{rep},(k)}$ and calculate $T(\mathbf{n}^{\text{rep},(k)})$ to determine the empirical reference distribution of the statistic T. For instance, calculate $T(\mathbf{n}^{\text{rep},(k)}) = X^2(\mathbf{n}^{\text{rep},(k)})$ (and/or other statistics of interest).

Step 5: Estimate the posterior predictive p-value for a test statistic by the proportion of $T(\mathbf{n}^{\text{rep},(k)})$ which are greater than, or equal to $T(\mathbf{n})$:

$$\hat{p}_{test} = K^{-1} \sum_{k=1}^{K} I(T(\mathbf{n}^{\text{rep},(k)}) \geq T(\mathbf{n}^{\text{obs}})). \tag{13}$$

If $\hat{p}_{test}$ is less than a predefined value (for instance, .05) we conclude that the model does not fit the data properly.

PPCs are generally used to check whether specific aspects of the observed data are correctly picked up by the model (Gelman, Carlin, Stern, & Rubin, 2004). The BVR statistic is a good example of this, as it indicates one specific aspect of the model, rather than GoF at the aggregate level. However, whether the X^2, G^2 or the BVR are suitable for use as test statistics in the PPC has yet to be determined.

An issue with the PPC using test statistics, which also holds for the parametric bootstrap, is that ML estimates have to be obtained for each of the replicated data sets. This makes both procedures rather time consuming, because the model has to be estimated for each replicated data set.

3.4 PPC using discrepancies

The added value of the PPC using discrepancies (Gelman et al., 1996) over the PPC using test statistics and parametric bootstrap is that it not only incorporates uncertainty about the model parameters, but it also eliminates the need for model estimation for each replicated data set because we can define discrepancies $D(\mathbf{n}^{\text{obs}}, \boldsymbol{\psi})$ which not only depend the data $\mathbf{n}$ but also on the model parameters phi. This makes the PPC using discrepancies computationally much faster than the other resampling methods.

Using the index k for a specific draw for the parameters obtained through the data augmentation algorithm, the PPC using discrepancies proceeds as follows:

Step 1: Obtain K draws $\boldsymbol{\psi}^{(k)}$ from the posterior distribution for the C-class model:

$$\boldsymbol{\psi}^{(k)} \sim p(\boldsymbol{\psi}|\mathbf{n}^{\mathrm{obs}}), k = 1, \ldots, K.$$

Step 2: Calculate the estimated pattern probabilities $\widehat{P}(\mathbf{y}_s)^{(k)}$ from $\boldsymbol{\psi}^{(k)}$. Draw K random samples of size N from a multinomial distribution with parameters $\widehat{P}(\mathbf{y}_s)^{(k)}$

$$\mathbf{n}^{\mathrm{rep},(k)} \sim \mathrm{Multin}(N, \widehat{P}(\mathbf{y}_1)^{(k)}, \ldots, \widehat{P}(\mathbf{y}_S)^{(k)})$$

Step 3: Calculate, for each data set $\mathbf{n}^{\mathrm{rep},(k)}$ the realised discrepancies $D(\mathbf{n}^{\mathrm{obs}}, \boldsymbol{\psi}^{(k)})$ and replicated discrepancies $D(\mathbf{n}^{\mathrm{rep},(k)}, \boldsymbol{\psi}^{(k)})$. For instance, when using the Pearson X^2:

$$D(\mathbf{n}^{\mathrm{obs}}, \boldsymbol{\psi}^{(k)}) = X^2(\mathbf{n}^{\mathrm{obs}}, \boldsymbol{\psi}^{(k)}) = \sum_{s=1}^{S} \frac{(n_s - e_s^{(k)})^2}{e_s^{(k)}} \tag{14}$$

and

$$D(\mathbf{n}^{\mathrm{rep},(k)}, \boldsymbol{\psi}^{(k)}) = X^2(\mathbf{n}^{\mathrm{rep},(k)}, \boldsymbol{\psi}^{(k)}) = \sum_{s=1}^{S} \frac{(n_s^{(k)} - e_s^{(k)})^2}{e_s^{(k)}}, \tag{15}$$

where the expected frequencies $e_s^{(k)} = NP(\mathbf{y}_s|\boldsymbol{\psi}^{(k)})$ (see Equation 1). The $n_s^{(k)}$ are the pattern frequencies in the replicated data set $\mathbf{n}^{\mathrm{rep},(k)}$.

Step 4: Estimate the posterior predictive p-value for a discrepancy by the proportion of replications for which $D(\mathbf{n}^{\mathrm{rep},(k)}, \boldsymbol{\psi}^{(k)})$ is greater than or equal to $D(\mathbf{n}^{\text{obs}}, \boldsymbol{\psi}^{(k)})$:

$$\hat{p}_{disc} = K^{-1} \sum_{k=1}^{K} I(D(\mathbf{n}^{\text{rep},(k)}, \boldsymbol{\psi}^{(k)}) \geq D(\mathbf{n}^{\text{obs}}, \boldsymbol{\psi}^{(k)})), \tag{16}$$

(where the indicator function I equals 1 if $D(\mathbf{n}^{\text{rep},(k)}, \boldsymbol{\psi}^{(k)}) \geq D(\mathbf{n}^{\text{obs}}, \boldsymbol{\psi}^{(k)})$ and 0 otherwise). If $\hat{p}_{disc}$ is close to 0 or 1, depending on what discrepancy is used, we conclude that the model does not fit the data properly (Gelman et al., 1996, 2004).

Note that Steps 1 and 2 for PPC using test statistics are exactly the same as Steps 2 and 3 for PPC using discrepancies. But rather than comparing replicated statistics to a single observed value based on the ML estimates, the PPC using discrepancies compares K *pairs* of discrepancies; that is, K *realised* discrepancies, $D(\mathbf{n}^{\text{obs}}, \boldsymbol{\psi}^{(k)})$, with K *predictive* discrepancies, $D(\mathbf{n}^{\text{rep},(k)}, \boldsymbol{\psi}^{(k)})$.

It is important to note that p_{disc}-values are different from the other p-values in the sense that their distribution under the null-hypothesis is generally non-uniform (Meng, 1994) Rather its distribution tends to be peaked around .5 (Robins, van der Vaart, & Ventura, 2000). Because of this, the PPC using discrepancies will usually provide more conservative results and have lower power to reject a false model (Gelman, 2013).

4 Simulation Study

The quality of bootstrap and PPC p-values for global and local GoF testing in LC analysis was investigated using two Monte Carlo studies. The first study evaluated the Type I error rates. The second study investigated the power of the different methods and statistics. In both studies, p-values were then obtained by either comparing the statistics to

1. a χ^2 distribution with given df,
2. the empirical distribution from the parametric bootstrap,
3. the empirical distribution from the PPC using test statistics, or
4. the empirical distributions from the PPC using discrepancies.

We used the software package R 2.15.1 (R Core Team, 2012) to generate data sets, to perform the PPC using discrepancies, and to collect the results. For ML estimation, asymptotic p-value calculation, and the parametric bootstrap, we used LatentGOLD 5.0 (Vermunt & Magidson, 2013). The MCMC algorithm for the Bayesian LC analysis was implemented in a routine written in C. We used a burn-in of 1000 iterations and subsequently intervals of 10 iterations of the data augmentation algorithm between draws of $\boldsymbol{\psi}^{(k)}$.[1]

[1] Inspection of the parameter estimates indicated that a burn-in of 1000 iterations was sufficient for our models, providing estimates comparable to the population parameters.

4.1 Study 1. Type I errors

4.1.1 Design

To check Type I error rates of the different p-values, we fully crossed the following design factors:

- Sample size $N = 100$, 1000, or 5000.
- Number of LCs $C = 2$, or 3.
- Number of dichotomous items $J = 6$, or 10.
- Conditional response probabilities $\pi_{1j1} = \pi_{2j2} = .7$, .8, or .9, for all j, and $\pi_{rj3} = \pi_{rj1}$, for $j = 1, \ldots, J/2$ and $\pi_{rj3} = \pi_{rj2}$ for $j = J/2{+}1, \ldots, J$

Table 1 provides the population parameters for each LC when $\pi_{1j1} = .8$. Additionally, we analysed conditions with J=6 trichotomous items ($R_j = 3$ for all j) and sample sizes $N = 100$, 1000, or 5000. The population parameters are shown also in Table 1. In all conditions we generated 2000 data sets. Each data set was analysed using a LC model in which the number of classes was equal to the number of classes in the population model (i.e., the null-hypothesis was true). The parametric bootstrap was performed with $B = 500$ replications conditional on $\hat{\boldsymbol{\psi}}$. The PPC using test statistics and PPC using discrepancies were performed based on $K = 500$ replications/draws.

We chose the simulation conditions such that the parameter values influence the level of sparseness but are also practically relevant. The chosen sample sizes of 100, 1000 and 5000 correspond typically to small, medium

Table 1: Example of Population Parameters for the LCs, c, for conditions with $J = 6$ items.

	$R_j = 2$					$R_j = 3$			
	$c = 1$	$c = 2$	$c = 3$	$c = 4$		$c = 1$	$c = 2$	$c = 3$	$c = 4$
π_c	.25	.25	.25	.25	π_c	.25	.25	.25	.25
π_{11c}	.8	.2	.8	.2	π_{11c}	.7	.1	.7	.1
π_{12c}	.8	.2	.8	.2	π_{21c}	.2	.2	.2	.2
π_{13c}	.8	.2	.8	.2	π_{12c}	.7	.1	.7	.1
π_{14c}	.8	.2	.2	.8	π_{22c}	.2	.2	.2	.2
π_{15c}	.8	.2	.2	.8	π_{13c}	.7	.1	.7	.1
π_{16c}	.8	.2	.2	.8	π_{23c}	.2	.2	.2	.2
					π_{14c}	.7	.1	.1	.7
					π_{24c}	.2	.2	.2	.2
					π_{15c}	.7	.1	.1	.7
					π_{25c}	.2	.2	.2	.2
					π_{16c}	.7	.1	.1	.7
					π_{26c}	.2	.2	.2	.2

Note. For all conditions, including those with fewer than $C = 4$ LCs, the parameters for the class sizes always equal $1/c$.

and large data sets, respectively. The sample size influences the degree of sparseness in the contingency table: The fewer respondents, the sparser the contingency table becomes.

The number of items (6 or 10) and number of response categories affects the degree of sparseness. The number of possible patterns (i.e., cells in the table) was either $2^6 = 64$, $3^6 = 729$ or $2^{10} = 1024$. Note that in the $J = 10$ item conditions, sparseness may be a problem even with a sample size of 5000.

Conditional response probabilities of .7, .8, and .9, respectively, indicate a weak, medium and strong associations of the items with the LCs. Note that these probabilities also influence the degree of sparseness besides sample size and the number of items. When the conditional response probabilities of a particular response to an item gets closer to 1, the number of patterns decreases, leading to an increase in sparseness.

Increasing the number of classes, on the other hand, decreases the sparseness of the contingency table, since the response preferences of each class lead to different response patterns. However, because this decrease in sparseness comes with an increased model complexity, it will be interesting to see any trade-off between model complexity and sparseness in determining the fit of a LC model.

Under the null-hypothesis, p-values should be uniformly distributed (Sackrowitz & Samuel-Cahn, 1999). This also means that (approximately) 5% of the p-values should have values less than .05. We will therefore investigate the performance of the methods by checking whether the proportion of the simulation data sets yielding a p-value less than .05 is close to .05.

4.1.2 Results

Results from study 1 on Type I error rates can be found in Tables 2 through 5 for the dichotomous conditions and in Table 6 for the trichotomous conditions. The tables for the dichotomous conditions are arranged such that the least sparse condition is located top-left, meaning that by going downward or to the right, sparseness increases. For each combination of condition, fit-statistic and type of p-value, we provide the proportion of simulations in which the obtained p-value was less than .05. Due to expected fluctuations in 2000 replications per condition, we expect 99% of the p-values to lie within the "expected interval" $.05 \pm 2.58\sqrt{.05(1-.05)/2000}$ (i.e., between 0.037 and 0.063). In the traditional context of null-hypothesis testing this interval would signify close-to-nominal Type I error rates. Proportions outside the interval may indicate problems with a given method, statistic, or combination of both and these proportions are underlined in the table. Note that for the BVR statistic, the asymptotic p-values are based on a χ^2_1 distribution for the dichotomous and χ^2_4 for the trichotomous conditions, even though it has been shown to be incorrect. We include them to assess the practical implications of this common usage. No asymptotic p-values are provided for the $TBVR$ and DI.

The standard GoF chi-squared statistics

Tables 2 and 3 provide the simulation results for the standard chi-squared GoF statistics for the dichotomous two-class and three-class conditions, respectively.

As expected, the asymptotic p-values only provided close to nominal Type I error rates for the situations where sparseness was not an issue. For $J = 6$ items and $N = 5000$ or $N = 1000$ observations, the asymptotic p-values may be useful, except when using the G^2. Asymptotic p-values for G^2 only reached close-to-nominal Type I error rates when there were 5000 observations.

The bootstrap and PPC using test statistics did considerably better than the asymptotic p-values and performed comparably well, where serious problems only occurred in the most sparse condition of $J = 10$ and $N = 100$. The differences between parametric bootstrap and PPC using test statistics are generally small and mostly involve the G^2 statistic. For the G^2, the PPC using test statistics provides more conservative results than the parametric bootstrap in the $J = 10$ conditions for $N = 1000$ and $N = 100$.

Looking at the PPC using discrepancies, we see that the proportions of p-values less than .05 lie in the expected interval only in the $\pi_{1j1} = .7$, $J = 6$, and $N = 5000$ or 1000 conditions. For the G^2 this also holds for the $\pi_{1j1} = .8$ and $\pi_{1j1} = .9$ conditions when $N = 5000$ and for the X^2 when $\pi_{1j1} = .9$ and $N = 5000$. For all other conditions the proportion of p-values less than .05 was (much) less than .05, confirming the non-uniformity of the p_{disc}-value.

For the trichotomous conditions, the results found in Table 6 make it clear that the parametric bootstrap provides close-to-nominal Type I-error rates in nearly all conditions and for all global fit statistics. Only in the $N = 100$ conditions were the Type I error rates outside of the expected interval. The PPC using test statistics was overall a bit more conservative, even in the least sparse case. The PPC using discrepancies was much too in practically all conditions. Again it is shown that asymptotic p-values are very unreliable, unless when used for the X^2 statistic in the $N = 5000$ conditions.

The results for the two- and three-class model are similar, albeit that the PPCs tend to get more conservative when model complexity increases. This effect is especially noticeable in the trichotomous conditions.

Statistics without a known asymptotic distribution

Tables 4 and 5 provide the simulation results for the BVR, $TBVR$, and DI for the dichotomous two-class and three-class conditions, respectively. These are all measures for which asymptotic p-values are not available.

First of all, it can be observed that using the χ^2_1 distribution as the asymptotic reference distribution for the BVR is inadequate. The highest Type I error rate was .0110, but generally these were much smaller still.

The parametric bootstrap generally works very well for the BVR, $TBVR$, and DI, with most proportions inside or very close to the expected interval, although it seems to work less well for the DI in the most extreme sparseness condition. The PPC using test statistics had more proportions outside the expected interval, which generally resulted in somewhat more conservative conclusions. Overall, resampling techniques seem to work well when there is no reference distribution available.

For the PPC using discrepancies, only in 1 condition, for the DI, a proportion of p-values was found inside the expected interval. It can be seen that here, too, the p_{disc}-values are not uniformly distributed.

For the trichotomous conditions, the results found in Table 6 again show that the parametric bootstrap works very well when applied to the BVR, $TBVR$ and DI. It was only too conservative in the sparsest case of $N = 100$ and $C = 2$ in combination with the DI. The PPCs were too conservative, except for the PPC using the local fit measures as fit statistics in non-sparse conditions with two LCs. The BVR clearly did not follow a χ^2_4 distribution as the Type I error rate for the p_{asymp} was 0 in all conditions.

The results for the two- and three-class model are similar, but again the PPCs become more conservative when model complexity increases.

Table 2: Type I Error Rates (the Proportion of P-Values which were Less Than $\alpha = .05$) for the Global Fit Statistics based on 2000 MC Simulation Replications for the Conditions with 2 LCs.

		J=6						J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	.7	.056	.056	.058	.035	G^2	.7	.623	.053	.055	.054
		.8	.048	.044	.045	.026		.8	.591	.046	.047	.042
		.9	.063	.044	.046	.027		.9	.000	.054	.045	.043
	X^2	.7	.055	.057	.057	.031	X^2	.7	.057	.059	.057	.052
		.8	.043	.044	.043	.027		.8	.062	.049	.049	.027
		.9	.047	.048	.045	.023		.9	.252	.065	.065	.037
	CR	.7	.057	.057	.059	.031	CR	.7	.055	.058	.056	.054
		.8	.043	.044	.045	.027		.8	.012	.050	.049	.031
		.9	.044	.048	.048	.024		.9	.000	.064	.064	.035
N=1000	G^2	.7	.052	.046	.042	.025	G^2	.7	.509	.051	.055	.048
		.8	.082	.059	.057	.032		.8	.000	.048	.033	.030
		.9	.098	.054	.056	.028		.9	.000	.038	.017	.017
	X^2	.7	.041	.044	.043	.028	X^2	.7	.126	.056	.054	.043
		.8	.055	.058	.057	.027		.8	.185	.058	.057	.016
		.9	.064	.051	.051	.021		.9	.357	.038	.038	.023
	CR	.7	.041	.042	.044	.026	CR	.7	.005	.055	.059	.045
		.8	.051	.056	.053	.026		.8	.000	.059	.057	.015
		.9	.039	.053	.053	.021		.9	.000	.040	.040	.019
N=100	G^2	.7	.167	.072	.082	.033	G^2	.7	1.000	.096	.044	.022
		.8	.035	.069	.059	.021		.8	1.000	.026	.002	.002
		.9	.000	.073	.025	.011		.9	.905	.051	.000	.000
	X^2	.7	.041	.051	.054	.029	X^2	.7	1.000	.033	.021	.037
		.8	.042	.041	.044	.010		.8	1.000	.016	.016	.003
		.9	.140	.031	.048	.001		.9	1.000	.061	.076	.008
	CR	.7	.032	.060	.065	.032	CR	.7	1.000	.118	.118	.034
		.8	.017	.050	.052	.015		.8	1.000	.054	.045	.001
		.9	.011	.054	.055	.003		.9	.999	.074	.074	.005

Table 3: Type I Error Rates (the Proportion of P-Values which were Less Than $\alpha = .05$) for the Global Fit Statistics based on 2000 MC Simulation Replications for the Conditions with 3 LCs.

		J=6						J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	.7	.056	.058	.057	.023	G^2	.7	.702	.053	.052	.049
		.8	.044	.042	.044	.016		.8	.490	.052	.054	.051
		.9	.056	.043	.043	.017		.9	.000	.055	.050	.044
	X^2	.7	.055	.054	.057	.022	X^2	.7	.054	.059	.053	.048
		.8	.041	.044	.045	.014		.8	.060	.048	.048	.028
		.9	.040	.045	.041	.012		.9	.175	.057	.052	.029
	CR	.7	.053	.055	.057	.021	CR	.7	.051	.057	.056	.052
		.8	.044	.046	.046	.015		.8	.015	.049	.052	.035
		.9	.040	.045	.042	.012		.9	.001	.057	.057	.028
N=1000	G^2	.7	.068	.059	.055	.021	G^2	.7	.128	.067	.064	.055
		.8	.067	.045	.044	.015		.8	.000	.057	.035	.026
		.9	.097	.055	.056	.020		.9	.000	.047	.024	.019
	X^2	.7	.054	.060	.057	.019	X^2	.7	.133	.057	.055	.044
		.8	.043	.044	.043	.014		.8	.188	.058	.052	.012
		.9	.059	.052	.054	.014		.9	.319	.041	.042	.021
	CR	.7	.053	.061	.056	.020	CR	.7	.001	.065	.068	.048
		.8	.042	.045	.044	.014		.8	.000	.059	.057	.012
		.9	.048	.054	.055	.017		.9	.000	.048	.043	.015
N=100	G^2	.7	.097	.055	.056	.020	G^2	.7	1.000	.252	.028	.008
		.8	.029	.053	.063	.011		.8	1.000	.113	.001	.000
		.9	.003	.072	.034	.012		.9	1.000	.089	.000	.000
	X^2	.7	.059	.052	.054	.014	X^2	.7	1.000	.019	.024	.013
		.8	.029	.031	.061	.005		.8	1.000	.031	.052	.000
		.9	.083	.046	.078	.002		.9	1.000	.055	.076	.002
	CR	.7	.048	.054	.055	.017	CR	.7	1.000	.134	.118	.010
		.8	.008	.040	.073	.007		.8	1.000	.105	.087	.000
		.9	.015	.060	.073	.006		.9	1.000	.078	.071	.000

Table 4: Type I Error Rates (the Proportion of P-Values which were Less Than $\alpha = .05$) for the BVR, the total BVR and the DI based on 2000 MC Simulation Replications for the Conditions with 2 LCs.

		J=6						J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	BVR	.7	.006	.065	.059	.000	BVR	.7	.011	.059	.057	.001
		.8	.002	.052	.050	.001		.8	.002	.060	.059	.001
		.9	.000	.045	.043	.001		.9	.000	.048	.049	.000
	$TBVR$	.7	NA	.051	.052	.000	$TBVR$	.7	NA	.051	.049	.001
		.8	NA	.053	.052	.001		.8	NA	.043	.043	.001
		.9	NA	.047	.048	.001		.9	NA	.049	.047	.000
	DI	.7	NA	.057	.055	.015	DI	.7	NA	.047	.044	.038
		.8	NA	.049	.050	.006		.8	NA	.049	.042	.028
		.9	NA	.052	.053	.001		.9	NA	.047	.039	.003
N=1000	BVR	.7	.002	.043	.039	.001	BVR	.7	.010	.049	.046	.001
		.8	.001	.061	.059	.000		.8	.003	.048	.044	.000
		.9	.000	.061	.061	.001		.9	.000	.048	.046	.000
	$TBVR$	.7	NA	.046	.042	.001	$TBVR$	.7	NA	.055	.058	.000
		.8	NA	.051	.045	.000		.8	NA	.047	.042	.000
		.9	NA	.051	.044	.001		.9	NA	.045	.034	.000
	DI	.7	NA	.053	.052	.016	DI	.7	NA	.058	.057	.041
		.8	NA	.053	.052	.006		.8	NA	.031	.009	.016
		.9	NA	.060	.051	.001		.9	NA	.046	.021	.004
N=100	BVR	.7	.007	.038	.037	.001	BVR	.7	.010	.042	.040	.001
		.8	.001	.052	.039	.000		.8	.004	.053	.044	.000
		.9	.000	.047	.036	.001		.9	.001	.059	.045	.000
	$TBVR$	.7	NA	.033	.033	.001	$TBVR$	.7	NA	.047	.031	.000
		.8	NA	.042	.021	.001		.8	NA	.053	.025	.000
		.9	NA	.048	.018	.001		.9	NA	.044	.007	.000
	DI	.7	NA	.071	.072	.018	DI	.7	NA	.045	.010	.015
		.8	NA	.062	.033	.003		.8	NA	.017	.000	.008
		.9	NA	.046	.007	.002		.9	NA	.027	.000	.006

Table 5: Type I Error Rates (the Proportion of P-Values which were Less Than $\alpha = .05$) for the BVR, the total BVR, and the DI based on 2000 MC Simulation Replications for the Conditions with 3 LCs.

		J=6						J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	BVR	.7	.000	.043	.044	.001	BVR	.7	.005	.053	.051	.000
		.8	.000	.058	.057	.000		.8	.001	.052	.051	.000
		.9	.000	.056	.056	.000		.9	.001	.051	.053	.000
	$TBVR$	.7	NA	.052	.051	.001	$TBVR$	.7	NA	.054	.051	.000
		.8	NA	.061	.058	.000		.8	NA	.047	.045	.000
		.9	NA	.047	.046	.000		.9	NA	.058	.053	.000
	DI	.7	NA	.051	.052	.007	DI	.7	NA	.053	.049	.035
		.8	NA	.050	.050	.002		.8	NA	.049	.045	.023
		.9	NA	.043	.036	.000		.9	NA	.055	.048	.004
N=1000	BVR	.7	.000	.034	.023	.000	BVR	.7	.004	.042	.039	.000
		.8	.000	.047	.038	.000		.8	.001	.051	.050	.000
		.9	.000	.046	.037	.000		.9	.000	.052	.048	.000
	$TBVR$	.7	NA	.043	.033	.000	$TBVR$	.7	NA	.052	.048	.000
		.8	NA	.043	.033	.000		.8	NA	.054	.046	.000
		.9	NA	.045	.037	.000		.9	NA	.042	.038	.000
	DI	.7	NA	.049	.049	.007	DI	.7	NA	.058	.043	.035
		.8	NA	.037	.034	.002		.8	NA	.050	.029	.019
		.9	NA	.048	.044	.000		.9	NA	.046	.014	.006
N=100	BVR	.7	.000	.046	.037	.000	BVR	.7	.012	.044	.046	.000
		.8	.003	.016	.029	.000		.8	.003	.045	.037	.000
		.9	.001	.033	.030	.000		.9	.001	.043	.037	.000
	$TBVR$	.7	NA	.045	.037	.000	$TBVR$	.7	NA	.051	.048	.000
		.8	NA	.018	.025	.000		.8	NA	.040	.021	.000
		.9	NA	.036	.039	.000		.9	NA	.024	.012	.000
	DI	.7	NA	.048	.044	.000	DI	.7	NA	.135	.007	.008
		.8	NA	.054	.041	.003		.8	NA	.071	.000	.004
		.9	NA	.054	.005	.004		.9	NA	.051	.000	.003

Table 6: Type I Error Rates (the Proportion of P-Values which were Less Than $\alpha = .05$) based on 2000 MC Simulation Replications for the Trichotomous Conditions, where $J = 6$ and $\boldsymbol{\pi}_{r11} = \{.7, .2, .1\}$

		$C = 2$					$C = 3$			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	$\underline{.585}$	.049	$\underline{.021}$	.048	G^2	$\underline{.516}$	.057	$\underline{.009}$	.049
	X^2	.054	.044	$\underline{.023}$	$\underline{.033}$	X^2	.059	.052	$\underline{.010}$	$\underline{.029}$
	CR	$\underline{.027}$	.049	$\underline{.020}$	.038	CR	$\underline{.033}$	.055	$\underline{.011}$	.038
	BVR	$\underline{.000}$	.046	.045	$\underline{.004}$	BVR	$\underline{.000}$	.055	$\underline{.035}$	$\underline{.001}$
	$TBVR$	NA	.051	.049	$\underline{.000}$	$TBVR$	NA	.046	$\underline{.033}$	$\underline{.000}$
	DI	NA	.040	$\underline{.010}$	$\underline{.021}$	DI	NA	.059	$\underline{.006}$	$\underline{.024}$
N=1000	G^2	$\underline{.000}$	.055	$\underline{.015}$	.039	G^2	$\underline{.000}$	.056	$\underline{.005}$	$\underline{.035}$
	X^2	$\underline{.087}$	.053	$\underline{.031}$	$\underline{.024}$	X^2	$\underline{.083}$	.048	$\underline{.017}$	$\underline{.020}$
	CR	$\underline{.000}$	.058	$\underline{.029}$	$\underline{.020}$	CR	$\underline{.001}$	.059	$\underline{.011}$	$\underline{.021}$
	BVR	$\underline{.000}$	.046	.045	$\underline{.002}$	BVR	$\underline{.000}$	.043	$\underline{.022}$	$\underline{.004}$
	$TBVR$	NA	.054	.050	$\underline{.001}$	$TBVR$	NA	.051	$\underline{.026}$	$\underline{.000}$
	DI	NA	.045	$\underline{.007}$	$\underline{.022}$	DI	NA	.047	$\underline{.002}$	$\underline{.017}$
N=100	G^2	$\underline{1.000}$	.038	$\underline{.000}$	$\underline{.001}$	G^2	$\underline{1.000}$	$\underline{.075}$	$\underline{.000}$	$\underline{.001}$
	X^2	$\underline{1.000}$	$\underline{.020}$	$\underline{.008}$	$\underline{.002}$	X^2	$\underline{1.000}$	$\underline{.022}$	$\underline{.021}$	$\underline{.001}$
	CR	$\underline{1.000}$	$\underline{.070}$	$\underline{.017}$	$\underline{.005}$	CR	$\underline{1.000}$	$\underline{.079}$	$\underline{.023}$	$\underline{.000}$
	BVR	$\underline{.000}$	.040	$\underline{.032}$	$\underline{.004}$	BVR	$\underline{.001}$	.046	$\underline{.026}$	$\underline{.000}$
	$TBVR$	NA	$\underline{.036}$	$\underline{.018}$	$\underline{.000}$	$TBVR$	NA	.040	$\underline{.010}$	$\underline{.000}$
	DI	NA	$\underline{.021}$	$\underline{.000}$	$\underline{.006}$	DI	NA	.051	$\underline{.000}$	$\underline{.004}$

4.2 Study 2. Power Analysis

Design

After evaluating Type I error rates, we also investigated power of the different p-values. Power is the probability of rejecting a model when it is indeed false. To do this, we estimated a two-class model on data sets generated under a three-class population, and estimated a three-class model on data sets generated under a four-class population. Population parameters for these conditions were $\pi_{1j1} = .8$ with $J = 6$, or 10 items in the dichotomous cases, and $\pi_{1j1} = .7$ with $J = 6$ items in the trichotomous cases (cf. Table 1). For each condition 2000 data sets were generated and analysed.

Results

Results of the power analysis for the dichotomous conditions can be found in Tables 7 and 8 and for the trichotomous conditions in Table 9. Power of .8 or greater is generally regarded to be acceptable, and higher values are better. It is immediately clear that the power to detect that a model has too few LCs is very high in medium ($N = 1000$) to large ($N = 5000$) data sets, as most of the power values are 1.0. For small data sets ($N = 100$) the power was around .2, though it is noteworthy that the $TBVR$, when used in the parametric bootstrap or as statistic in the PPC, has high power even in the sparsest condition when $C = 2$. Also, the power to detect misfit using the $TBVR$ increases as the number of items increases.

In order to draw conclusions about the usefulness of the methods and statistics, we need to combine the results of Study 1 and 2. For example, when a statistic has high power but also has large Type I error rates (larger than the chosen level of significance α), the statistic will lead to too liberal results and general use is not recommended. In such cases we would have a high chance of rejecting a model, regardless of whether the model is actually true or false. For the p_{boot} and p_{test}, power was high and Type I error rates were very accurate in most conditions as well. The p_{test} is overall somewhat more conservative. The p_{disc} had very low Type I error rates but still had high power to detect the misspecification of the models in our simulation by means of the global chi-squared statistics. The p_{asymp} also showed high power, but also had very high Type I error rates when sparseness became an issue (e.g. when $J = 10$). When assessing LC model fit using any particular statistic, we advise researchers to use either the parametric bootstrap or PPC using fit statistics. When tables are sparse due to small sample sizes, researchers should resort to local fit statistics, which may be tailored to the research question at hand.

Table 7: Power (the Proportion of P-values which were Less Than $\alpha = .05$) to Indicate Model Misfit when a Model with C Classes is Estimated on Data Generated under Population with $C + 1$ LCs. Conditions with $J = 6$ Dichotomous items where $\boldsymbol{\pi}_{r11} = \{.8, .2\}$. Results are based on 2000 MC Simulations.

		C = 2					C = 3			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	1.000	1.000	1.000	.973	BVR	1.000	1.000	1.000	.966
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	1.000
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=1000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.711	.922	.914	.513	BVR	.515	.980	.974	.088
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	.0110
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=100	G^2	.456	.524	.497	.353	G^2	.156	.163	.160	.056
	X^2	.413	.409	.428	.257	X^2	.065	.093	.132	.023
	CR	.308	.488	.496	.313	CR	.045	.125	.152	.036
	BVR	.228	.322	.321	.048	BVR	.014	.110	.130	.000
	$TBVR$	NA	.717	.690	.002	$TBVR$	NA	.133	.118	.000
	DI	NA	.546	.492	.310	DI	NA	.217	.181	.039

Table 8: Power (the Proportion of P-values which were Less Than $\alpha = .05$) to Indicate Model Misfit when a Model with C Classes is Estimated on Data Generated under Population with $C + 1$ LCs. Conditions with $J = 10$ Dichotomous items where $\boldsymbol{\pi}_{r11} = \{.8, .2\}$. Results are based on 2000 MC Simulations.

		C = 2					C = 3			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.973	.993	.992	.861	BVR	.959	1.000	1.000	.733
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	1.000
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=1000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.614	.761	.759	.501	BVR	.514	.976	.954	.453
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	.817
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	.999	1.000
N=100	G^2	1.000	.485	.124	.100	G^2	1.000	.355	.010	.006
	X^2	1.000	.116	.126	.056	X^2	1.000	.012	.018	.002
	CR	1.000	.286	.258	.055	CR	1.000	.117	.069	.002
	BVR	.237	.316	.320	.073	BVR	.047	.167	.162	.003
	$TBVR$	NA	.967	.957	.002	$TBVR$	NA	.567	.463	.001
	DI	NA	.290	.092	.179	DI	NA	.262	.003	.029

Table 9: Power (the Proportion of P-values which were Less Than $\alpha = .05$) to Indicate Model Misfit when a Model with C Classes is Estimated on Data Generated under Population with $C+1$ LCs. Conditions with $J = 6$ Trichotomous items where $\boldsymbol{\pi}_{r11} = \{.7, .2, .1\}$. Results are based on 2000 MC Simulations.

		C = 2					C = 3			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.838	1.000	1.000	.870	BVR	.967	1.000	1.000	.992
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	1.000
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=1000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	.996	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	.998	.996	.964	.984
	CR	1.000	1.000	1.000	1.000	CR	.980	1.000	.991	.996
	BVR	.600	.840	.813	.509	BVR	.514	.976	.954	.453
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	.817
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	.999	1.000
N=100	G^2	1.000	.338	.013	.062	G^2	1.000	.221	.001	.006
	X^2	1.000	.071	.053	.014	X^2	1.000	.022	.017	.002
	CR	1.000	.215	.111	.018	CR	1.000	.096	.030	.002
	BVR	.140	.480	.463	.086	BVR	.001	.295	.224	.003
	$TBVR$	NA	.963	.932	.001	$TBVR$	NA	.614	.312	.000
	DI	NA	.283	.006	.147	DI	NA	.187	.000	.019

5 Empirical Data

We will illustrate the methods described in the paper with a data set taken from Landis and Koch (1977) (see also, Holmquist, McMahan, and Williams (1968)). It contains information on 118 slides which were evaluated on the absence or presence of cervical cancer by seven pathologists. So, we have a data set with 7 dichotomous item and a sample size of 118. Only 20 of the possible $2^7 = 128$ response patterns were observed, indicating that we are dealing with a rather sparse contingency table. This sparse table has been used by various authors who proposed using bootstrap p-values for global fit testing with G^2 (Agresti, 2002; Magidson & Vermunt, 2004; Vermunt & Magidson, 2005) Here, we will also look at other measures and consider both PPCs in addition to the bootstrap.

We estimated LC models with two or three LCs and assessed the GoF of these two models based on the X^2, G^2, CR, BVR, $TBVR$, and DI statistics. Results from these analyses can be found in Table 10.

Table 10: Fit Statistics and P-values for the Cervical Cancer Data. Model with 2 or 3 LCs.

	Model with 2 LCs						Model with 3 LCs				
	Value	p_{asymp}	p_{boot}	p_{test}	p_{disc}		Value	p_{asymp}	p_{boot}	p_{test}	p_{disc}
G^2	64.163	1.000	.000	.012	.020	G^2	17.713	1.000	.500	.948	.662
X^2	90.564	.800	.028	.144	.332	X^2	21.120	1.000	.296	.870	.852
CR	74.851	.980	.006	.062	.212	CR	18.589	1.000	.360	.908	.828
$TBVR$	32.281	NA	.000	.000	.656	$TBVR$	8.328	NA	.026	.042	.586
DI	.268	NA	.000	.000	.010	DI	.117	NA	.146	.598	.298
BVR_{12}	1.736	.188	.028	.022	.314	BVR_{12}	.051	.822	.332	.442	.364
BVR_{13}	.387	.534	.090	.188	.838	BVR_{13}	.092	.762	.318	.370	.804
BVR_{14}	.273	.601	.196	.360	.668	BVR_{14}	.575	.448	.120	.124	.612
BVR_{15}	.146	.702	.456	.422	.410	BVR_{15}	.162	.687	.416	.290	.430
BVR_{16}	.209	.648	.362	.438	.628	BVR_{16}	.043	.836	.670	.668	.526
BVR_{17}	.024	.878	.682	.542	.442	BVR_{17}	.152	.697	.428	.338	.386
BVR_{23}	.017	.896	.824	.852	.648	BVR_{23}	.006	.939	.794	.838	.648
BVR_{24}	.577	.447	.190	.228	.736	BVR_{24}	.599	.439	.172	.172	.704
BVR_{25}	8.443	.004	.000	.000	.204	BVR_{25}	.036	.850	.290	.364	.336
BVR_{26}	.445	.505	.302	.332	.618	BVR_{26}	.477	.490	.256	.236	.548
BVR_{27}	5.205	.023	.000	.000	.304	BVR_{27}	.029	.866	.532	.442	.354
BVR_{34}	.895	.344	.256	.240	.782	BVR_{34}	.019	.890	.774	.726	.568
BVR_{35}	1.106	.293	.042	.056	.786	BVR_{35}	.134	.715	.114	.320	.788
BVR_{36}	1.316	.251	.160	.158	.772	BVR_{36}	.021	.886	.820	.758	.562
BVR_{37}	.138	.711	.098	.278	.832	BVR_{37}	.078	.780	.436	.376	.760
BVR_{45}	.043	.836	.814	.746	.586	BVR_{45}	.701	.403	.092	.112	.546
BVR_{46}	7.228	.007	.006	.002	.950	BVR_{46}	4.521	.033	.004	.004	.826
BVR_{47}	.099	.753	.236	.418	.622	BVR_{47}	.426	.514	.210	.194	.592
BVR_{56}	.589	.443	.248	.204	.612	BVR_{56}	.070	.792	.286	.550	.522
BVR_{57}	3.331	.068	.000	.000	.328	BVR_{57}	.101	.751	.356	.282	.372
BVR_{67}	.075	.785	.286	.520	.584	BVR_{67}	.038	.846	.664	.628	.526

Asymptotic p-values are not appropriate here due to the sparseness of the contingency table. Based on the simulation results of sparse tables, we expect to see that, overall, the PPC using test statistics provides somewhat more conservative p-values than the parametric bootstrap does.

Indeed, for the two-class model the parametric bootstrap provides p-values of .028 for the X^2, .000 for the G^2 and .006 for the CR, indicating that the model is inadequate. The p_{test}-values were .144, .012 and .062, respectively, meaning only the G^2 statistic suggests lack of fit for the two-class model. The p_{disc} values were .332, .020 and .212 respectively. Here too, only the G^2 statistic indicated lack of fit.

Inspection of the bivariate residuals for the two-class model shows that some association remains between the item pairs {2,5}, {2,7}, {5,7} and {4,6}. Asymptotic p-values based on the χ^2_1 distribution indicate significant remaining associations, except perhaps for the BVR of items 5 and 7 (p_{asymp} = .058). The parametric bootstrap and PPC using test statistics both indicate that these remaining associations are significantly different from 0. The PPC using discrepancies did not provide p-values close to zero. However, the most extreme p_{disc}-values are generally found for the largest BVR. For BVR_{46} the p-value was .950, which also indicates misfit.

The parametric bootstrap and PPC using test statistics both indicate model misfit with regard to the $TBVR$ and DI, with p-values of .000. The PPC using discrepancies only indicated lack of fit for the DI and not for the $TBVR$.

For the three-class model, all methods indicate that the global fit of the model is adequate, based on the X^2, G^2 and CR and DI. As we expected from the simulation results, the p_{test}-values for these statistics were larger than those from the bootstrap.

Inspection of the bivariate residuals reveals that the association between the item pair $\{4,6\}$ is not picked up by the three-class model ($BVR_{46} = 4.521$). The parametric bootstrap and PPC using test statistics both indicate that this remaining association is significantly different from 0. The PPC using discrepancies did not provide extreme p-values here. This agrees with the simulation in which we virtually never saw p_{disc}-values for the BVR less than .05.

The parametric bootstrap and PPC using test statistics are able to pick up that there is remaining bivariate association through the $TBVR$ statistic, as both techniques provided small p-values for this statistic.

In summary, the analyses show that a three-class model has adequate overall fit, but lacks in local fit, as indicated by the p-values for BVR_{46} and for the $TBVR$. Also, the empirical data analysis was in agreement with our expectations from the simulation study that the PPC using test statistics yields somewhat more conservative p-values than the parametric bootstrap does.

6 Discussion

To assess the fit of a LC model when contingency tables are sparse or when asymptotic reference distributions are not available, resampling techniques can be used to obtain empirical reference distributions for any goodness-of-fit statistics. In the current paper we evaluated a number of statistics which are commonly used in the assessment of model fit, some of which are specific to LC models. We conducted a simulation study to investigate whether reliable p-values could be obtained with the parametric bootstrap, the PPC using test statistics, and the PPC using discrepancies.

The simulation study involved calculating different p-values when analysing sparse and non-sparse contingency tables both for fit statistics that have no known asymptotic distribution, as well as for statistics for which the asymptotic distributions do not hold in sparse situations. In agreement with previous studies we found that the use of asymptotic p-values resulted in (severely) distorted Type I error rates when contingency tables were sparse. Both the parametric bootstrap and PPC using test statistics performed much better in this regard than the asymptotic method. Von Davier (1997) showed that the likelihood-ratio G^2 is not suitable for use in the parametric bootstrap when contingency tables are sparse, as it will generally lead to too liberal conclusions. We have replicated this finding and have additionally shown that using the G^2 as a statistic in the PPC resulted in too conservative results. Because the G^2 had high Type I error rates and high power, we cannot be sure what a small p_{asymp}-value indicates. The Pearson X^2 and CR however, did generally provide close-to-nominal Type I error rates and had high power. These latter should therefore in many situations be preferred over the likelihood-ratio statistic G^2.

The DI statistic worked very well in combination with the parametric bootstrap and with the PPC using test statistics. The PPC using test statistics provides somewhat more conservative p-values. Only in the most sparse condition did the parametric bootstrap show severe problems. The DI appears therefore be a good statistic to assess global model fit, even when contingency tables are sparse. However, when sample size is small ($N = 100$), it lacks power like the other global chi-squared statistics.

Sparseness has little effect on the BVR statistics, especially for dichotomous items, as it only involves the second order marginals of the contingency tables. Therefore, it may be hypothesised that the use of asymptotics is justified. However, we have shown in line with Oberski et al. (2013) that the common distributional assumption for the BVR does not hold for LC models. Use of the χ^2_1-distribution produced too conservative results (i.e., low Type I error rates). We would like to stress that the poor results ascribed to the asymptotic p-values for the BVR statistics are due to the choice for this reference distribution. Future research should indicate which, if any, asymptotic reference distribution should be used for the BVR in LC analysis.

For the BVR, both the parametric bootstrap and PPC using test statistics resulted in close-to-nominal Type I error rates, even when the tables were very sparse. The latter method provided somewhat more conservative results than the former. The BVR statistic failed completely in combination with the PPC using discrepancies.

The parametric bootstrap yielded close-to-nominal Type I error rates when using the $TBVR$. In combination with the PPC using test statistics, the $TBVR$ resulted in somewhat below-nominal Type I error rates. The power of the $TBVR$ was very high, however, and it shows that taking all bivariate associations into account provides very good information on model fit, even when tables are very sparse. Note that the findings of the BVR and $TBVR$, the latter being the sum of all BVRs, should not be seen as independent.

Our power study suggested that all methods and statistics are useful to detect misfit when the number of LCs is misspecified. When sample sizes become very small, however, the results have shown that we should resort to the local fit measures. Especially the $TBVR$ has very high power, since it is not greatly affected by sparseness and still uses information on all item pairs to indicate whether misfit is present. Since no asymptotic distribution is known for this statistic, its use in the parametric bootstrap and as a statistic in the PPC will show to be of great value, even when data is sparse.

To illustrate our findings we analysed an empirical data set where, due to sparseness, the use of asymptotic p-values was inadequate. We obtained alternative p-values by means of the parametric bootstrap, the PPC using test statistics and PPC using discrepancies. In line with the results from the simulation study, we found that the PPC using test statistics provided somewhat more conservative results than the parametric bootstrap. Bootstrapping the global fit statistics strongly suggested that a two-class model did not fit the data adequately. However, when incorporating the uncertainty about the parameter estimates in the analysis, the PPC using test statistics did not provide very strong evidence to suggest model misfit. No disagreement was found between the parametric bootstrap and PPC using test statistics with regard to the BVR, $TBVR$ and DI statistics. In the better fitting three-class model, all methods indicated no lack of global fit. A nice result was that the parametric bootstrap and PPC using test statistics were well able to pick up violations in the local fit through the BVR and $TBVR$ statistics, even though the global fit measures indicated no problems.

Overall, the computationally less intensive PPC using discrepancies provided more conservative results than the other resampling techniques. This was, to an extent, expected from the fact that the distribution of p_{disc} under the null-hypothesis is peaked around .5. A number of methods have been proposed to adjust the p_{disc} value so that it provides uniform p-values (Bayarri & Berger, 2000; Hjort, Dahl, & Steinbakk, 2006; Robins et al., 2000). Future research should indicate whether extra computational burden of calibrating p_{disc}-values outweighs the benefits, compared to the properly working PPC using test statistics.

Given the established results, researchers should be weary of using asymptotic reference distributions when the sample sizes are not very large and/or when there are many variable, leading to a sparse contingency table. Resorting to lower-order statistics, like the BVR, and statistics which are specifically tailored to a certain application or research question, like the DI, is good practice, even if their distributions are unknown. Though developing new statistics was not our aim here, many others can be conceived of. For dichotomous data, one could use a bivariate Pearson correlation to assess local dependencies. When interest lies in a specific second-order relationship, trivariate residuals could be used. If one response pattern is of particular interest, one could use the observed frequency of that pattern as a statistic. In each of these cases, resampling methods can provide reliable p-values. Also, in the very sparse cases, using resampling techniques to assess combinations of the lower level associations, like the $TBVR$ proved to be very useful.

On a final note, this paper addressed the question of assessing model *fit* and not model *comparison*. Interpretation of, for instance, information criteria like the AIC and BIC does not change when the sample size is small or when contingency tables are sparse.

References

Agresti, A. (2002). *Categorical data analysis.* John Wiley & Sons. Hoboken, New Jersey.

Bayarri, M., & Berger, J. (2000). P values for composite null models. *Journal of the American Statistical Association*, *95*(452), 1127–1142.

Berkhof, J., Van Mechelen, I., & Gelman, A. (2003). A bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, *13*(2), 423–442.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*(3), 375–389.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 440–464.

Crow, S. J., Swanson, S. A., Peterson, C. B., Crosby, R. D., Wonderlich, S. A., & Mitchell, J. E. (2012). Latent class analysis of eating disorders: Relationship to mortality. *Journal of abnormal psychology*, *121*(1), 225.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, *39*(1), 1–38.

Dufour, M., Brunelle, N., & Roy, É. (2013). Are poker players all the same?

latent class analysis. *Journal of Gambling Studies*, 1–14.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Chapman & Hall/CRC. Boca Raton, FL.

Fisher, R. A. (1925). *Statistical methods for research workers.* Genesis Publishing Pvt Ltd.

Formann, A. K. (2003). Latent class model diagnosticsa review and some proposals. *Computational statistics & data analysis*, *41*(3), 549–559.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electronic Journal of Statistics*, *7*, 2595–2602.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis.* Chapman & Hall/CRC.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–759.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.

Haberman, S. J. (1979). *Analysis of qualitative data. volume 2: New developments.* Academic Press.

Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*, *83*(402), 555–560.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators local dependence models. *Sociological Methods & Research*, *16*(3), 379–405.

Hjort, N., Dahl, F., & Steinbakk, G. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*(475), 1157–1174.

Hoijtink, H. (1998). Constrained latent class analysis using the gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistica Sinica*, *8*, 691–711.

Holmquist, N. D., McMahan, C. A., & Williams, O. D. (1968). Variability in classification of carcinoma in situ of the uterine cervix. *Obstetrical & Gynecological Survey*, *23*(6), 580–585.

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*(3), 321–357.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, *24*(4), 492–516.

Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2004). Latent class analysis and latent transition analysis. In J. Schinka & W. Velicer (Eds.), *Handbook of psychology: Volume 2. research methods in psychology (pp.*

663-685). Hoboken, NJ: Wiley.

Laudy, O., Zoccolillo, M., Baillargeon, R. H., Boom, J., Tremblay, R. E., & Hoijtink, H. (2005). Applications of confirmatory latent class analysis in developmental psychology. *European Journal of Developmental Psychology*, *2*(1), 1–15.

Ligtvoet, R., & Vermunt, J. K. (2012). Latent class models for testing monotonicity and invariant item ordering for polytomous items. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 237–250.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H., & Clark, L. C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, *19*(10), 1303–1318.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences (pp. 175-198).* Thousand Oaks, CA: Sage Publications.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.

Meulders, M., De Boeck, P., Kuppens, P., & Van Mechelen, I. (2002). Constrained latent class analysis of three-way three-mode data. *Journal of classification*, *19*(2), 277–302.

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*(3), 267–279.

Okazaki, S., Campo, S., Andreu, L., & Romero, J. (2014). A latent class analysis of spanish travelers? mobile internet usage in travel planning and execution. *Cornell Hospitality Quarterly*, 1938965514540206.

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. *Sociological methodology*, *29*(1), 81–111.

Rindskopf, D. (2002). The use of latent class analysis in medical diagnosis. In *Annual meeting of the american statistical association* (pp. 2912–2916).

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, *95*(452), 1143–1156.

Roedelof, A., Bongers, I. L., & van Nieuwenhuizen, C. (2013). Treatment engagement in adolescents with severe psychiatric problems: a latent class analysis. *European child & adolescent psychiatry*, *22*(8), 491–500.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Rubin, D. B., & Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In A. Von Eye, C. C. Clogg, et al. (Eds.), *Latent variables analysis: Applications for developmental research (pp. 420–438).* Thousand Oaks, CA: Sage Publications Inc.

Sackrowitz, H., & Samuel-Cahn, E. (1999). P values as random variables – expected p values. *The American Statistician*, *53*(4), 326–331.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Vermunt, J. K., & Magidson, J. (2005). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In A. K. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences (pp. 41–62).* Mahwah, NJ: Psychology Press.

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

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