DETECTING MEASUREMENT NONEQUIVALENCE WITH LATENT MARKOV MODELS

Erwin Nagelkerke[1][2], Duygu Güngör[3], & Jeroen K. Vermunt[1]

Abstract

In longitudinal studies it is important to test whether measurements are equivalent over time, because it needs to be known if observed changes are true change or caused by a change in the measurement of the construct of interest. However, in the application of latent Markov (LM) models, measurement nonequivalence is typically neglected and not tested for. In this paper two analytic strategies for such tests are investigated in the context of LM models: An approach that is common in structural equation modeling (SEM) by starting from a free baseline model and progressively restricting parameters, and an approach that is common in item response theory (IRT) modeling by starting from a fully constrained baseline model and progressively freeing parameters. Using a simulation study, we determine the true and false positive rates in detecting nonequivalent items for different model fit statistics. The results indicate that, regardless of the analytic strategy, the power to detect measurement nonequivalence in LM models is high as long as there is a sufficiently large measurement variance, or a sufficiently strong measurement. Out of the different model fit statistics considered the AIC and likelihood-ratio tests are most promising, whereas the BIC lacks power to detect nonequivalence.

Funding

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK)

[1] Tilburg University, Tilburg, The Netherlands
[2] Corresponding author: E.Nagelkerke@tilburguniversity.edu
[3] Dokuz Eylül University, İzmir, Turkey

Arguably the most central topic in the social and behavioral sciences is individual change over time, which can be studied by data collected from the same individual at multiple time points. To adequately deal with measurement error, longitudinal studies often use latent variable techniques such as item response theory (IRT) and structural equation modeling (SEM). A popular tool in cases where the data is categorical is the latent Markov (LM) model, also referred to as the latent transition, hidden Markov, or regime switching model[4] that can be considered to be a longitudinal version of the latent class (LC) model (see Collins & Lanza, 2010; Goodman, 1974; Vermunt & Magidson, 2002). Similar to the LC model, the LM model classifies subjects based on their responses, but the LM model simultaneously estimates the probabilities of transitioning from one class to another between measurement occasions. Parallel with the development in statistical packages such as Latent GOLD (Vermunt & Magidson, 2013), Mplus (Muthen & Muthen, 1998-2007), PROC LTA (Lanza & Collins, 2008), and the LMest R-package (Bartolucci & Pandolfi, 2018) there is a growing body of research using LM models for the analysis of change in a wide range of applied areas. Examples include studies on substance misuse (Lanza & Bray, 2010), delinquency (Bright, et al., 2017), eating behavior (Sotrez-Alvarez, Herring, & Siega-Riz, 2013), abnormal psychology (Connell et al., 2008), quality of life (Bartolucci, Lupparelli, & Montanari, 2009), work psychology (Bujacz, Bernhard-Oettel, Rigotti, Magnussen Hanson, & Lindfors, 2018), and health psychology (Williams et al., 2015).

Most longitudinal studies assume that the instruments lead to identical measurement of the same constructs over time. That is, the meaning of the instrument does not change and the scores that result from it indicate an identical presence or level of the measured concept. This is referred to as measurement equivalence or invariance. The importance of testing for measurement equivalence in longitudinal data has been stressed (Millsap, 2010; Millsap & Cham, 2012), and methods have been proposed to investigate longitudinal measurement equivalence using latent variable models such as IRT (Meade, Lautenschlager, & Hecht, 2005; Millsap, 2010) and latent growth models (Olivera-Aguliar, 2013; Widaman, Ferrer, & Conger, 2010; Wirth, 2008). Yet, despite the growing body of literature on

[4] These terms are frequently used interchangeably, despite all these models having a particular definition. For an overview see Bartolucci, Farcomeni, & Pennoni (2014).

longitudinal measurement (non)equivalence, researchers using LM models still tend to fully neglect this problem. On the one hand, this is somewhat surprising because it seems to be rather straightforward to investigate measurement equivalence in LM models by comparing nested models using statistics such as AIC, BIC, likelihood-ratio, Wald, and score statistics. On the other hand, it is understandable because there is no generally accepted strategy for investigating measurement equivalence.

In order to develop a more comprehensive testing framework we will here focus on the comparison of the SEM- and IRT-like analytic strategies in LM modeling. Our aim is to provide an advice to applied researchers using LM models on which strategy is more effective: the approach more common to SEM of starting with a free baseline model and progressively testing more constrained models, or the approach more common to IRT of starting with a constrained baseline model and testing progressively freed parameters. Moreover, different model fit statistics and information criteria will be compared to see whether the strategies align with certain statistics to further aid the testing for measurement equivalence in applied research.

In the following section the LM model and the two approaches to constraining and freeing parameters will be introduced. Next the simulation design and its results will be presented. The final section provides a discussion, recommendations for researchers, and suggestions for further research.

The Latent Markov Model

The LM model is mostly applied to measure a latent phenomenon based on categorical data, and model the change in that latent phenomenon over time. To achieve this, the model classifies respondents into latent states based on their responses, and estimates the probabilities of moving between states from one measurement occasion to the next.

Let Y_{tj} be one of J observed variables or items measured at T occasions, where $j = 1, 2, \dots, J$ and $t = 1, 2, \dots, T$, and let S_t be a latent variable of which the value s_t represents the latent state occupied at time point t. The response of an individual to item j at time t can then be denoted as y_{itj}, with i being one out of N individuals. In the case of categorical data, denoting the response category as r_{tj} out of R_j responses allows some technical efficiency, as the possible responses are finite. Then, finally, the responses of an individual to all items at one occasion can be denoted $\boldsymbol{y}_{it}$ and $\boldsymbol{r}_{it}$, and further concatenating those vectors for all measurement occasions gives all responses by one individual as $\boldsymbol{y}_i$ and $\boldsymbol{r}_i$.

As an illustration the LM model is depicted in Figure 1, which shows the two central assumptions of the LM model: The local independence assumption and (first order) Markov assumption. The former implies that the observed responses are independent of one another, conditional on the latent states occupied at the T time points. That is, no covariance exists between the observed variables conditional on the latent variable. The latter implies that the state at time point t only depends on the state occupied at the previous time point $t-1$, and only indirect relations exist between the latent variables that are not adjacent. It thus holds that $P(S_3 = s_3|S_2 = s_2, S_1 = s_1) = P(S_3 = s_3|S_2 = s_2)$ (Vermunt, Langeheine, & Böckenholt, 1999). These are signified by the product terms in the equation for the LM model:

[FIGURE ONE ABOUT HERE]

$$P(\boldsymbol{y}_i = \boldsymbol{r}_i) = \sum_{s_1=1}^{S} \dots \sum_{s_T=1}^{S} P(S_1 = s_1) \left[\prod_{t=2}^{T} P(S_t = s_t|S_{t-1} = s_{t-1})\right] \left[\prod_{t=1}^{T}\prod_{j=1}^{J} P\left(y_{itj} = r_{tj}\middle|S_t = s_t\right)\right] \quad (1)$$

LM models can be thought of as a series of LC models that make up the measurement part of the model. The longitudinal structural part of the model comprises the initial state and transition probabilities. In terms of Equation 1, the first element $P(S_1 = s_1)$ are the initial state probabilities, or proportions, indicating the prevalence of each state at the first measurement occasion, here denoted as $t = 1$. The transition probabilities $P(S_t = s_t|S_{t-1} = s_{t-1})$ make the Markov assumption apparent, where the current state s_t depends on state membership at the previous measurement s_{t-1}. The last elements form the measurement of the states, where the probability of a certain response is conditional on the current state occupied $P\left(y_{itj} = r_{tj} \middle| S_t = s_t\right)$.

Although identified when there are three or more consecutive observations, assuming more manifest than latent elements (Kasahara & Shimotsu, 2008), this model is hardly ever applied in practice. Generally, the assumption of measurement invariance is applied to fix the state definitions, and the transition probabilities are (partially) fixed over time. Not doing so largely negates one of the advantages of the model, namely its parsimony. That is, when transition probabilities vary over time, each pair of adjacent measurement occasions has its own set of probabilities and thus interpretations in terms of the transitions from and to states that also change definition at each occasion.

Testing for Measurement Equivalence

Measurement equivalence implies that the latent structure, that is the number of latent states and their definition, is the same across all time points and leads to identical scores. If true equivalence holds it is valid to impose across-time equality constraints on item response probabilities $P(Y_{tj} \mid S_t)$. In contrast, measurement nonequivalence occurs when estimated item response probabilities turn out to be different across occasions for one or more of the items. If some, but not all items have different response probabilities at different measurement occasions this is referred to as partial equivalence, and the extreme situation where all items are time specific corresponds to complete nonequivalence. The latter situation requires a fully unconstrained LM model, also referred to as the basic LM model (Bartolucci, Farcomeni, & Pennoni, 2013), in which the definition of latent states may fully change over time. This makes the interpretation of state membership and state transitions extremely challenging, as respondents are moving in and out of states that themselves change in substantive meaning.

Important to note is that a necessary requirement for measurement equivalence in LM models is that the number of latent states is equal across time points. However, when such a model does not hold, equivalence can still be tested by increasing the number of states, where some states are assumed to not occur at some of the measurement occasions. An example of this is a study investigating non-suicidal self-injury (NSSI) behaviors. At older ages four states may be encountered, namely experimental, mild, and multiple NSSI and self-cutting. The latter state, self-cutting behavior, generally does not yet exist at younger ages (Somer, et al., 2015). Such situations can be modeled with measurement equivalence by estimating a four state LM model in which the self-cutting state membership probability is fixed to zero for the first measurement occasions. This would still allow response probabilities to be equivalent across time points.

After deciding on the number of latent states needed, overall measurement equivalence can be tested by comparing two models. One constrained model where all item response probabilities are identical for each occasion, and one free model where the item response probabilities are allowed to be different for all the different measurement occasions (Collins & Lanza, 2010). When the fully constrained model does not fit significantly worse than the fully free model it gives credence to the assumption that item response parameters are equivalent across occasions. However, this comparison alone does not provide information on partial equivalence where some, or only one of the items is nonequivalent over time. Generally, researchers aim to find the most parsimonious and best interpretable model that is still theoretically relevant and statistically sound. Therefore, when the constrained model is rejected in favor of the unconstrained model, the partially equivalent models in between these two extremes become important, because it may be the case that a large number of parameters can still be considered equivalent over time. In LM models this is of extra importance, because the item response probabilities determine the definition of a latent state. This implies that in situations where a large number of parameters is still equivalent, class definitions may remain broadly comparable over time while allowing for partial equivalence. In turn this allows state membership and transitions between states to still be interpreted, without violations related to equivalence assumptions.

A practical problem in the search for an adequate partially equivalent model is that there is a large number of possible model comparisons that can be made. For example, for a three-state LM model, with three measurement occasions and six dichotomous items there are 54 measurement parameters that can be constrained. Moreover, nonequivalence may occur for each combination of items, time points, and states. In practice this inhibits the theoretical option of estimating and comparing all possible models as even for a very limited study there are thousands of alternative model specifications.

To make investigation of item-level (non)equivalence feasible, we will next consider two strategies that both involve testing nested models to one another. One by progressively moving from the fully unconstrained model to constraining more parameters, and one in the other direction.

[FIGURE 2 ABOUT HERE]

Model A in Figure 2 represents the free baseline model where item response probabilities are free to vary over time (indicated by using different subscripts for the state-item relationship). In the free strategy models are specified with a single item restricted to be equivalent over time. If that constraint does not deteriorate fit, the item concerned can be assumed to be equivalent over time. In contrast, if the restricted model (e.g. Model 2B) is rejected in favor of the unconstrained model, the item concerned should be flagged as being nonequivalent. A statistical advantage of using the unconstrained model as a baseline is that the alternative model in the test is always a model that fits the data. However, this advantage comes at the expense of parsimony and interpretability of the initial baseline model.

Alternatively, the model constrained to full equivalence may be used as a starting point (Model 2E). Subsequently across-time restrictions on the item response probabilities can be removed (Models 2F, G, and H). Now, when the constrained model is rejected for a partially unrestricted alternative the item concerned should be flagged as nonequivalent. Vice versa, accepting the constrained model would signify measurement equivalence for the tested item. Statistically the disadvantage of this approach is that the alternative model does not by definition fit the data. That is, if items other than the item being tested are nonequivalent the tested model may indicate that freeing up parameters is not warranted, but this does not mean that other, untested parameters do not need to be freed.

Regardless of the choice between these two starting points, with a J item scale, $J + 1$ models should be estimated, namely the one-by-one restriction of all the items, and the comparison between the fully constrained and fully unconstrained models. In practice this is feasible in the vast majority of situations, and allows identifying nonequivalent items. After identifying these items, the model with freely estimated nonequivalent items and restricted equivalent items can be estimated, which can be treated as the selected model that is at least close to the most parsimonious partially equivalent model.

Now, in addition to the strategy, the statistics that may be used for the model comparisons also require consideration. As nested models are compared, one of the most obvious choices in this respect is the likelihood-ratio (LR) test. The LR test uses either the minus two log-likelihood ($-2LL$) or likelihood-ratio chi-squared (L^2) difference between the null and alternative model. For example, using the unconstrained model as a starting point and comparing models A and B from Figure 2 the LR test value would equal:

$$LR = L_b^2 - L_a^2, \tag{2}$$

where L_a^2 represents the likelihood ratio chi-squared value for the unconstrained model, and L_b^2 represents the model with one item constrained over time. Similarly, models with one free item and the constrained baseline model can be compared. Because these models are nested the LR test follows a known central chi-square distribution when the alternative model fits to the data with degrees of freedom equal to the difference in the number of parameters between the two models.

Information criteria form one of the alternatives for model selection, the most commonly used of which are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Using the general notation of N for sample size and df for degrees of freedom these criteria are defined as:

$$AIC = L^2 - 2df \tag{3}$$

$$BIC = L^2 - \ln{(N)} \cdot df \tag{4}$$

As can be seen from Equation 4F the BIC penalizes the model for the number of free parameters weighted by sample size, whereas the AIC only penalizes free parameters with a constant. As a result the BIC is often preferred when sample sizes are large. When comparing models the lowest value indicates the preferred model for both criteria.

Other alternatives for model testing are specific to the baseline model used. The Wald test can be used as an alternative to the LR test when starting with the unconstrained model. Asymptotically the LR and Wald test are equivalent, but the Wald test has the advantage that it does not require estimating the J restricted models. More specifically, it can be obtained for each item using the estimated parameter values and the estimated parameter variances from the unconstrained model. This does imply that the quality of the test statistic is dependent on the variance estimates, which may be biased or subject to large sampling error when the sample size is small (Agresti, 1990).

When using the constrained model as the baseline model the Score or Lagrange multiplier test can be used as an alternative to the LR test. The Score test, similar to the Wald test, is asymptotically equivalent to the LR test, with the same advantage of not having to estimate the J unrestricted models by using the estimated parameter variance. More specifically, it tests whether a parameter of interest is equal to a particular value. Here that allows a test of whether the inclusion of an additional set of parameters that would relax the equivalence assumption leads to a significant improvement of the model. The major disadvantage of the Score test is that it is powerful when the estimated parameter values are close to their true value, but suffers from increasing bias the further the alternative model is from the truth. That is, the test statistics and p-values become biased when the less restricted model is still a bad fitting model.

In the next section the strategy of equivalence testing and the different fit statistics will be compared using a simulation study containing various forms of longitudinal measurement nonequivalence.

Methodology

In the simulation study measurement equivalence is tested by relaxing or constraining item response probabilities as described. By varying several characteristics of the generated population data the goal is to answer the following research questions:

1- Which of the two baseline strategies is more powerful when testing for longitudinal measurement equivalence in LM models?

2- Do the different evaluation statistics (LR, AIC, BIC, Wald and Score) give different results?

3- Do sample size, measurement strength, the number of nonequivalent items, the number of equivalent items, the degree of nonequivalence, and the number of nonequivalent states have an effect on the power of detecting equivalence?

Study design

In the simulation study, a number of characteristics is kept fixed in order for the results to remain interpretable and to be able to focus the effect of the number of (non)equivalent items over design factors that may confound. All conditions use six dichotomous items (pass/fail, or agree/disagree), measured at three occasions, and classified into three states. The initial state probabilities were set to be equal, so all states at occasion one are the same size, and the transition probabilities are specified to be time-homogeneous with the following values:

$$\begin{bmatrix} .75 & .20 & .05 \\ .05 & .80 & .15 \\ .01 & .04 & .95 \end{bmatrix}.$$

The first latent state represents a non-mastery, or disagree, state where the probabilities of giving a positive response are low. The opposite is true for latent state three, where the positive responses are set to a high probability. The second state is an intermediate state with positive responses being likely for the first three items and negative responses get a high probability for the other three items. Nonequivalence was specified to occur between the second and third measurement occasion, yielding response probabilities at the third occasion that are different from occasions one and two, where occasion one and two are equivalent to one another.

Note that throughout the simulation study, because of the above, it is assumed that the correct number of latent states or too many states are specified. Determining the number of states is an extensive topic of study and is beyond the scope of this paper. However, when more than the true number of states is specified in a situation without nonequivalent items it would theoretically result in empty states, and in practice result in overfitting. In situations with nonequivalent items, part of the nonequivalence may be modeled through the transition and response probabilities if there are too many states, which would make testing for it through model comparisons unreliable. Yet, this is a largely theoretical problem, because the variance would still be accounted for in the model parameters.

Five factors are varied when simulating the data: the sample size, the number of nonequivalent items, the strength of measurement, the degree of nonequivalence and the number of nonequivalent states. In the analyses of the generated data the approach is varied, that is the baseline model used as a starting point, and the different model fit statistics are compared.

The sample size is set to 100, 300 or 1000, consistent with the design of Collins and Wugalter (1992), but adding a small sample scenario. Sample size is kept equal over measurement occasions.

The number of nonequivalent items is set to zero, one, or three, corresponding to 0%, 17% and 50% of items. The first condition gives an indication of Type I error, or false positive rates. The items that are set to be nonequivalent are also varied, setting nonequivalence once for both the first and fourth item, and both items one through three and four through six. In combination with the specification of the intermediate second class this causes nonequivalence to be alternated between both the states with high positive and with high negative response probabilities.

The strength of measurement was varied between weak, moderate and strong by setting the dominant response probabilities (disagree for the negative state, agree for the positive state, and the respective items in the intermediate state) equal to either 0.7, 0.8, or 0.9. The response probabilities of the opposing response then become the complements 0.3, 0.2 or 0.1 given the dichotomous nature of the items. These probabilities can also be expressed in an explained variance statistic (R^2) based on the entropy or class separation, yielding values of .51, .78, and .91 respectively, which are often used as weak to strong state entropy values (e.g. Collins & Wugalter, 1992; Gudicha, Schmittmann, & Vermunt, 2015).

The degree of nonequivalence was set to be small, medium or large by raising the negative response probabilities of the third occasion to be different from occasion one and two. These differences are .10, .20, and .30 respectively. I.e. a low nonequivalence in strongly separated classes would lead to positive response probabilities of 0.1 – 0.9 – 0.9 in states 1, 2, and 3 respectively, except for occasion 3 where it would be 0.2 – 0.9 – 0.9. Medium would lead to 0.3 – 0.9 - 0.9 in state 3, etc.

The number of states with nonequivalence changes as a result of the design. Given the state definitions, the first three items can be considered easy items, having a high positive response probability in the second and third state, the latter three items are difficult with only high positive response probabilities in the third state. Nonequivalence occurs in either one or all easy, or one or all difficult items. That is, for the one or three easy items the nonequivalence would occur only in state one, for the difficult items it would occur in states one and two due to the items having low positive response probabilities in those states.

The strategy to detect nonequivalence is varied between the two directions of constraining or freeing parameters as discussed. That is, either starting from the free baseline model, or the fully constrained baseline model.

The generally applicable model fit statistics, that is the AIC, BIC, and L^2 are reported for all models. The Wald and Score tests are reported for all the items for the baseline model to which they apply, the free and constrained starting model respectively.

In summary, the varied factors and the number of conditions are: Sample size (3), the number of nonequivalent items (3), measurement strength or entropy (2), the degree of nonequivalence (3), and the number of nonequivalent states (2). In total 117 conditions are simulated for both approaches, and contain all relevant statistics.

Data Analysis

Data generation and analysis were conducted using Latent GOLD 5.0. The estimated model is always a 3-state LM model with homogeneous transition probabilities. In total 14 models are estimated for each population model, that is for each generated data set: The free baseline model, the constrained baseline model, and six models with either one item constrained or one item unconstrained compared to the baseline. For model selection four statistics are used: The LR test, AIC, BIC, and either a Wald test or a Score test for the unconstrained and constrained baseline model, respectively. Since the asymptotic distribution for the LR test is assumed to hold, the critical value for the chi-square test with six degrees of freedom equals 12.59, associated with a p-value of 0.05.

The true and false positive rates are evaluated across 100 replications per simulation condition. The true positive rate here represents the proportion of nonequivalent items that are correctly identified across the replications. For conditions with three nonequivalent items the reported rates are the averages over the three items. The false positive rate similarly indicates the average proportion, but of the items that are incorrectly flagged as nonequivalent.

Results

False Positive Rates

Table 1 presents the detection rates of nonequivalent items that are generated to be equivalent for the two baseline models. For the conditions without nonequivalence the false positive rates for the AIC are between .03 and .11, and .03 and .08 for the LR statistic, being structurally higher in the constrained baseline model compared to the unconstrained baseline model. In conditions with adequate information, either through high state separation or large N, the rates for the LR statistic are close to nominal alpha levels. The most striking result is the high false positive rate of the Wald test that incorrectly has a high detection rate for nonequivalence regardless of the condition. The BIC in contrast has no false detections, but this will be returned to as the true positive rate is also very low.

[TABLE 1 ABOUT HERE]

For the conditions with nonequivalence the false positive rates are largely similar. The AIC and LR test do well in medium to strong state entropy conditions, the BIC has low overall detection rates, and the Wald test is by far the worst.

[TABLES 2, 3 and 4 ABOUT HERE (OR TOPPING SUBSEQUENT PAGES)]

True Positive Rates

The true positive rates are presented for the different levels of measurement strength, with Tables 2, 3, and 4 presenting 36 conditions for the strong, medium, and weak entropy conditions, respectively. The tables are ordered by degree of nonequivalence and sample size.

Sample Size

Regardless of other factors, decreasing the sample size results in lower detection rates. This effect becomes more severe for weaker nonequivalence, or when nonequivalence was located in one state. As the BIC controls for sample size in model selection, it is most affected by the changes. State separation, or measurement strength acts as a confounding factor, whereby a decrease in sample size results in a sharp decrease in detection rates for conditions with weaker separation of states. The latter result is sensible in terms of the information available in the population data, where weakly separated states and low N allow for a redistribution of dependence throughout the model. That is, the nonequivalence over time can be accommodated by the model in the parameter estimates without leading to large misfit in terms of the response and transition probabilities, because these are unstable and weakly defined to begin with.

Note that the differences between conditions are relatively limited when the sample size is 100, and generally detection rates are unacceptably small. Only when there is strong nonequivalence in two well defined states are detection rates in the range of 0.90. Because of this, the results discussed for the following factors are those for the sample sizes of 300 and 1000.

Measurement Strength

In conditions where the states are well-defined, presented in Table 2, the strategy of starting from a fully constrained or unconstrained model are both viable when there is a high amount of nonequivalence. As the measurement strength of the states decreases, the rate of detection also decreases, illustrated by the drop in detection rates between Tables 2 and 3. Nonetheless, detection rates for the LR test and AIC remain relatively high except in the least favorable conditions. From the results in Table 4 it becomes apparent that detection of equivalence in weakly separated states is only reliably achieved when other factors are most favorable, thus in high N conditions with a large nonequivalence, where that nonequivalence occurs in two states rather than one.

Degree of Nonequivalence and Nonequivalent States

As can be expected, the degree of nonequivalence is the most important factor in affecting its detection rate. Additionally, the conditions are such that either one or two states are affected by nonequivalence, whereby fewer affected states make detection harder. This does imply that it is valuable to know, or have theoretical reasons as to whether one or more states are affected by nonequivalence.

When nonequivalence occurs in two latent states, true detection rates are high, especially in favorable conditions such as those with a large nonequivalence or high N. However, when only one state is affected the detection rates drop quite dramatically, even when nonequivalence is large. This again is largely due to the nature of the model and its dependence structure. Nonequivalence in two states results in far more parameters being affected, and the resulting model misfit, or decrease in model fit is substantially higher than when only one state is affected.

Related to this it also seems that easy and difficult items show different detection rates. An easy item, one with high response probabilities in two of the states, shows far lower detection rates and is practically not detected when nonequivalence is small. Conversely, detection rates are practically one for all statistics and both strategies when large enough nonequivalence occurs in a difficult item with low response probabilities in two of the states. This, however, is an artifact of the setup of the simulation study whereby difficult items show nonequivalence in states one and two at the third occasion, whereas easy items only show nonequivalence in state one at the third occasion.

Model Fit Statistics

As noted, the AIC and LR test are most promising in terms of detection rates. Because the BIC strongly favors model parsimony given its penalty term for added parameters, its detection rates are low. This is not surprising as the nonequivalence in this study affects a relatively small part of the model. Only a part of the indicator items and classes are affected in one measurement occasion. According to the BIC the decrease in model fit that this causes does not warrant additional parameters. Although not surprising, it is worrisome that the BIC fares so badly as it is the most used model selection criterion, and of the better known criteria also performs well on other issues such as detecting the number of classes (Tein, Coxe, & Cham, 2013). The Wald and Score test are on par with the AIC and LR test in most conditions, although they do show a tendency to perform slightly worse in favorable conditions (good class separation, high N, large nonequivalence, multiple states, difficult items) and slightly better in unfavorable conditions (limited class separation, low N, small nonequivalence, one state, easy items). It must also be noted that in this study, the Wald and score tests are not obtained per freed up or constrained item and they may fare better when obtained stepwise.

Number of Nonequivalent Items

In the conditions with well separated classes power was similar for both conditions and depended mainly on the degree of nonequivalence and the type of item. When the measurement strength decreases, the free baseline model approach was slightly more powerful than the constrained baseline model approach when there are three nonequivalent items. For conditions with a single nonequivalent item the true positive rates are high for both approaches.

Example Application

To illustrate the use of the approaches to detect longitudinal measurement nonequivalence a Markov model is applied to panel data on drugs use from the National Youth Survey (Vermunt, Tran, & Magidson, 2008; Elliot, Huizinga, & Menard, 1989). This data set contains data on 1725 children and teenagers that were followed for 16 years from the ages of 11 to 17 through the ages of 27 to 33 starting in 1976. A total of 13665 observations are available, but due to the data structure a subset of these is used: The first five observations are annual, and the following observations triannual which results in 9 observations per respondent. To keep the example more succinct only the observations with a three year separation are used (2, 5, 6, 7, 8 and 9) for cases without missing responses on the indicators relating to alcohol, marijuana and hard drugs use and abuse, resulting in 1063 cases with 6378 observations. The data is further simplified by dichotomizing the original indicators into whether or not the respondent has used alcohol, marijuana or hard drugs in the past year. The data are freely available online after registration[5] and are included in LatentGOLD as an example data set.

Previous analyses have indicated that a 4-state LM model fits best based on a combination of statistical and substantive arguments (Nagelkerke, 2018). The BIC selects a 7-state LM, but the states added in addition to these four have a definition that is almost identical to another state. This indicates that the fifth, sixth and seventh state mainly exist to better model the transitions over time. This can be resolved by relaxing the assumption that transition probabilities remain identical between occasions and allowing heterogeneous transitions over time. The four states, presented in Table 5, can be defined as 'No substance use', 'Alcohol Only', Alcohol and Marijuana', and 'All Substances'.

[TABLE 5 ABOUT HERE]

Using this model as the starting point the two baseline models are estimated, namely the fully constrained and fully free model with regard to longitudinal measurement variance. Subsequently all of the three separate items are either freed up for the fully constrained model, or constrained for the fully free model. This is achieved by including the measurement occasion, here the year of observation, and allowing it to affect the probability (or logit) describing the relation between the observed indicator and the latent variable. Intuitively the measurement moment can then be understood as a nominal moderator variable, allowing the effect of the latent variable on the observed indicator to differ over time.

Note that this does bring in a secondary assumption related to longitudinal measurement variance, namely that this change in the relation between the manifest and latent variable is equal for all the states as it pertains to the latent variable as a whole. To fully free up the model in terms of

[5] See https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/88

measurement variance the specification needs to be extended such that the effect of time is allowed to differ per state, i.e. moving from $P\left(Y_{tj}\ |S_t, T\right)$ to $P\left(Y_{tj}\ |\ S_t, T|\ S_t\right)$. The latter is a third specification that is used as fully free baseline and per item nonequivalence testing, resulting in the sixteen models presented in Table 7, where the more parsimonious $P\left(Y_{tj}\ |S_t, T\right)$ models are presented in Table 6.

[TABLE 6 ABOUT HERE]

Comparing the two baseline models in Table 6, only the AIC prefers the unconstrained over the constrained model. The LR test indicates that the constrained model does not fit significantly worse than the unconstrained model and the BIC indicates a better suited model given its preference for parsimony. However, the LR test, respective Wald and Score tests, and both AIC and BIC prefer a model where at least one item is allowed to have a measurement variance. This shows that checking for measurement equivalency per item is important even when the global model fit tests indicate no nonequivalence: The large amount of parameters added to the model and complex dependence structure may obscure certain assumption violations when only considering global model fit. Further note that the interpretation of the Score test here is a significant improvement in fit when adding parameters, and the Wald is interpreted as a significant reduction in fit after constraining parameters.

Both approaches in Table 6 reach the same conclusion when inspecting the per indicator item models. The free baseline approach indicates that fixing the first indicator does not significantly decrease the fit of the model, and the BIC and AIC indicate that the more parsimonious model is preferred. The results for the constrained baseline approach indicate that both items two and three should be made nonequivalent to improve fit. Further testing (not in table) of the model relaxing the invariance assumption for item two against the model relaxing items two and three indicates that the Wald (16.116, *df* = 5, *p* = .007), Score (25.971, *df* = 5, *p* < .001) and LR (15.117, *df* = 5, *p* = .010) tests all indicate a significantly better fit when relaxing the equivalence assumption for both indicators. I.e. fixing the first item, or relaxing the second and third items are the best fitting models out of these alternatives, which of course are the same model.

Theoretically, this model is also warranted. In terms of the acceptance of marijuana use the years in which the panel was fielded show relatively strong fluctuations (Keyes, et al, 2011), which may translate to other substances and affect the way people respond to self-report questions on using drugs (Richter & Johnson, 2001; Johnson, 2014). In addition to this cohort effect, there is a longstanding idea that a period effect occurs whereby there is a tendency to over report substance use during high school (Gfroerer, 1997). Although these effects are hard to disentangle, the interpretation of the questions on drug use and potential measurement error such as systematic socially desirable reporting is likely to affect measurement. For alcohol use such effects are expected to be considerably smaller as a result of it being considered a far less sensitive topic, and the stigma surrounding a drug is central to the degree of biased reporting (Hser, 1997).

[TABLE 7 ABOUT HERE]

Furthermore, given the extensive overview in (Johnson, 2014) it can be expected that for different types of drug users these changes in measurement differ. Therefore, in Table 7 the models are further extended to allow measurement to change longitudinally, and this change to differ between states. That is, an interaction between the occasion and latent state is allowed, in addition to the main effect of occasion on the response probabilities.

These additional interaction effects are not preferred by the fit statistics presented. The score test expects the parameter to be non-significant, the LR test shows no global model improvement, and both the AIC and BIC prefer the more parsimonious model. The notable exception is the Wald test indicating that the interaction effects are significant for marijuana and drug use. However, this is an artifact of testing, whereby the Wald test cannot distinguish between the main effect of measurement occasions and the added interaction effect. As a result most of the explained variance ends up being arbitrarily attributed to the interaction effect, and the true likelihood improvement is nowhere near the model improvement suggested by the Wald test.

Substantively, the final model after testing for measurement invariance, is a model with four states, heterogeneous transitions, and a relaxation of the measurement equivalence assumption for the drug and marijuana use indicators that does not differ between latent states. Adding these ten parameters improves the likelihood chi-square from 2254 to 2178. More importantly, the additions to the model allow a better substantive description of drug and marijuana use. Table 8 displays the model response probabilities substance use per year per latent state for the original and measurement variant model.

For the two items that are allowed to be measurement nonequivalent a non-linear development can be concluded from Table 8. There is a markedly stronger jump in the probability of having used marijuana and drugs in the measurement occasion (5) where all respondents have entered adolescence, aged 17-23. At later ages marijuana use tapers off after this initial jump in all classes, whereas the use of other drugs increases in all classes.

[TABLE 8 ABOUT HERE]

These findings do indicate that during adolescence there may be overreporting of marijuana use, with the higher probabilities for having used marijuana in the past year in all classes. Note that the longitudinal effects of actual higher use should already be captured by the original Markov process that allows heterogeneous transitions between classes, and this sudden increase holds for all types of users. A similar result is seen with regards to drug use, where the expected increase over time does indeed seem to hold from the increase in the response probabilities of having used drugs other than marijuana in the previous year. This is possibly due to increased acceptance over the years.

Discussion

The main objectives of this study is to compare two different strategies to investigate and detect longitudinal measurement nonequivalence in LM models, including the type of statistic that is used for model selection. The strategies align with the preferred model testing approach in SEM and multigroup latent class models or IRT, where the former prefer to start from a free baseline model, while researchers in the latter typically start with the constrained baseline model.

Our simulation study in this regard suggests that the detection rates for measurement nonequivalence will be good regardless of the analytic strategy used as long as the nonequivalence is sufficiently large or sufficient information is available to detect it in the form of a large sample, well defined states, or nonequivalence that affects larger areas of the model. The major difference between the two approaches is that in the case of multiple nonequivalent items, the detection rates of this nonequivalence is slightly higher when starting from the free baseline model, as compared to the constrained baseline. The reason for this is that, given the dependence structure of the model, the constrained baseline model starts from the assumption of full equivalence. Freeing up parameters in this situation does not guarantee that the dependence between indicator items is modeled adequately, and the model fit may not increase enough to warrant that particular parameter to be freed individually. However, the parameter may be of enough added value in combination with other parameters to warrant its estimation. The free-baseline model suffers a similar problem where deterioration of the model by constraining an item may only surface after constraining a second item, but this happens to a far smaller extent, because it makes no a-priori assumptions about the other parameters. Moreover, it would also more naturally lead to a situation in which, after constraining an item, the initial item is tested again.

Based on these results, we suggest applied researchers to use the free baseline model in a generic scenario of testing for measurement nonequivalence, especially if a substantial portion of items is expected to show nonequivalence. This is also in line with previous research in related fields. Kim and Willson (2014) showed that using the constrained baseline model approach may yield false positive results when detecting nonequivalence in multi-group second-order latent growth models. Stark, Chernyshenko and Drasgow (2006) compared the free baseline and constrained baseline strategies in a simulation study using confirmatory factor analysis and item response theory to detect measurement equivalence across groups. They showed that false positive results were higher when the constrained baseline strategy was used. A further theoretical argument for the free-baseline model is made by Kankaras, Vermunt and Moors (2011) as it ensures that comparisons are always made with a model that is known to fit the data.

A second recommendation in light of the results is that, with regard of global model fit indices, the AIC or likelihood-ratio test should be preferred when testing for longitudinal measurement nonequivalence. The widely used BIC prefers parsimony to such an extent that it fails to indicate that model fit may improve by relaxing the equivalence assumption. The Wald and score tests in this study are used as more global tests too, and are obtained from their respective baseline models. They generally perform as well as using the AIC or LR tests in the current simulation setup, but as they are tests designed for individual parameters, they will likely outperform these measures when obtained iteratively for each individual parameter constraint. Do note that this would require reversing the logic, where the score test would be applied to the free-baseline instead of the constrained-baseline model and test whether a certain constraint is warranted.

Do note that a strong word of warning is in order here, as following these recommendations to detect measurement variance does not necessarily result in obtaining the correct model. One issue not taken into account here is the inflation of Type I error when testing many items subsequently. When the number of items is large, the number of $J + 1$ tests is too, and the chances of falsely detecting measurement invariance increase. Following test outcomes blindly then results in capitalization on chance, and will lead to overfitting or accepting a wrong model. This is an effect that is seen for most modification indices and local fit measures when not controlling for inflated Type I error. Detection of nonequivalence should be well considered, and model adjustments should not be made blindly merely to improve overall model fit, but should be validated and theoretically sensible.

The other results from the simulation study are generally related to the population and are all in the direction that they would be expected. Generally, measurement nonequivalence is easier to detect with more favorable conditions of either the nonequivalence itself being more pronounced, or more information through a larger sample or better separated classes. These findings nonetheless doe lead to additional questions for future research. To further address the issue of sample size and inspect the required sample size to reliably detect nonequivalence a valuable next step would be to apply and extend the power computation methods proposed by Gudicha et al. (2015).

Furthermore, in situations where nonequivalence is limited detection rates are very low. However, in these situations the detection rate may not actually matter, and a future study may focus initially on the impact on parameter estimates and parameter recovery when the nonequivalence is ignored. It may well be the case that the model is able to incorporate limited nonequivalence without it resulting in extensive model misfit or parameter bias, negating the need for high detection rates.This would similarly allow the Wald and score tests to be applied in a per-model basis, and extend the investigation to include additional fit statistics to inspect whether they reliably detect nonequivalence, specifically when it affects smaller parts of the model.

Finally, one aspect that is ignored in this study are the transition probabilities, which are assumed to be homogeneous. In many applications this assumption is too strong, and heterogeneous transitions are required, as is the case in the application. The impact that this may have on nonequivalence detection requires further research, since it complicates the dependence structure of the data and may cause measurement nonequivalence to be harder to detect.

Despite all these additional unknowns, we do feel confident in our recommendation for the type of strategy and selection of fit statistics, and the application shows that modeling longitudinal measurement nonequivalence may indeed alter substantive outcomes as well as the types of research questions that can be answered adequately.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley & Sons.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates. *Test, 23*(3), 433-465. doi: 10.1007/s11749-014-0381-7.

Bartolucci, F., Lupparelli, M., & Montanari, G. E. (2009). Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *The Annals of Applied Statistics*, *3*(2), 611-636. doi: 10.1214/08-AOAS230.

Bartolucci, F. & Pandolfi, S. (2016), *LMest package for the R Project for Statistical Computing*, available at: https://cran.r-project.org/web/packages/LMest/index.html

Bright, C. L., Sacco, P., Kolivoski, K. M., Stapleton, L. M., Jun, HJ., & Morris-Compton, D. (2017). Gender differences in patterns of substance use and delinquency: A latent transition analysis. *Journal of Child & Adolescent Substance Abuse, 26*(2), 162-173. doi: 10.1080/1067828X.2016.1242100.

Bujacz, A., Bernhard-Oettel, C., Rigotti, T., Magnussen Hanson, L., & Lindfors, P. (2018). Psychosocial working conditions among high-skilled workers: A latent transition analysis. *Journal of Occupational Health Psychology, 23*(2), 223-236. doi: 10.1037/ocp0000087

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New Jersey: John Wiley & Sons. doi: 10.1002/9780470567333.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*(1), 131–157. doi: 10.1207/s15327906mbr2701_8.

Connell, A., Bullock, B. M., Dishion, T. J., Shaw, D., Wilson, M., & Gardner, F. (2008). Family intervention effects on co-occuring early childhood behavioral and emotional problems: A latent transition analysis approach. *Journal of Abnormal Child Psychology*, *36*(8), 1211-1225. doi: 10.1007/s10802-008-9244-6.

Elliot, D. S., Huizinga, D., & Menard, S. (1989). *Multiple problem youth: Delinquency, substance use, and mental health problems.* New York, NY: Springer-Verlag.

Gfroerer, J., Lessler, J., & Parsley, T. (1997). Studies of nonresponse and measurement error in the National Household Survey on Drug Abuse. In L. Harrison & A. Hughes (Eds.), *The validity of self-reported drug use: Improving the accuracy of survey estimates*, (pp. 273-295). Rockville, MD: National Institute on Drug Abuse.

Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach. *American Journal of Sociology, 79*(5), 1179-1259. doi: 10.1086/225676.

Gudicha, D., Schmittmann, V., & Vermunt, J. K. (2015). Power computation for likelihood ratio tests for the transition parameters in latent Markov models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 234-245. doi: 10.1080/10705511.2015.1014040.

Hser, Y. (1997). Self-reported drug use: Results of selected empirical investigations of validity. In L. Harrison & A. Hughes (Eds.), *The validity of self-reported drug use: Improving the accuracy of survey estimates*, (pp. 320-343). Rockville, MD: National Institute on Drug Abuse.

Johnson, T. P. (2014). Sources of error in substance use prevalence surveys. *International Scholarly Research Notices, 2014.* doi: 10.1155/2014/923290.

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, *40*(2), 279-310. doi: 10.1177/0049124111405301.

Kasahara, H., & Shimotsu, K. (2008). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, *77*(1), 135-175. doi: 10.3982/ECTA6763.

Keyes, K. M., Schulenberg, J. E., O’Malley, P. M., Johnston, L. D., Bachman, J. G., Li, G., & Hasin, D. (2011). The social norms of birth cohorts and adolescent marijuana use in the United States, 1976-2007. *Addiction, 106*(10), 1790-1800. doi: 10.1111/j.1360-0443.2011.03485.x.

Kim, E. S., & Willson, V. L. (2014). Testing measurement invariance across groups in longitudinal data: Multigroup second-order latent growth model. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 566-576. doi: 10.1080/10705511.2014.919821.

Lanza, S. T., & Bray, B. C. (2010). Transitions in drug use among high-risk women: An application of latent class and latent transition analysis. *Advances and applications in Statistical Sciences*, *3*(2), 203-235.

Lanza, S. T., & Collins, L. M. (2008). A new SAS procedure for latent transition analysis: Transition in dating and sexual behavior. *Developmental Psychology*, *44*(2), 446-456. doi: 10.1037/0012-1649.44.2.446.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*(3), 279-300. doi: 10.1207/s15327574ijt0503_6.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data. *Child Development Perspectives*, *4*(1), 5-9. doi: 10.1111/j.1750-8606.2009.00109.x

Millsap. R. E. & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card. (Eds.), *Handbook of Developmental Research Methods*, (pp. 109-126). New York: Guilford.

Muthen, B. & Muthen, L. (1998-2007). *Mplus user's guide fifth edition*. Los Angeles: Muthen & Muthen.

Nagelkerke, E. (2018). *Local fit in multilevel latent class and hidden Markov models* (Doctoral dissertation). Retreived from https://pure.uvt.nl/portal/files/23344636/Nagelkerke_Local_16_02_2018.pdf

Oberski, D. L. & Vermunt, J. K. (*Under review*). The expected parameter change (EPC) for local dependence assessment in binary data latent class models. Retrieved from https://arxiv.org/abs/1801.02400.

Olivera-Aguliar, M. (2013). *Impacts of violations of longitudinal measurement invariance in latent growth models and autoregressive quasi-simplex models.* (Doctoral dissertation). Retrieved from https://repository.asu.edu/items/18699.

Richter, L., & Johnson P. B. (2001). Current methods of assessing substance use: A review of strengths, problems, and developments. *Journal of Drug Issues, 31*(4), 809-832. doi: 10.1177/002204260103100401.

Somer, O., Bildik, T., Kabukçu-Başay, B., Güngör, D., Başay, Ö., & Farmer, R.F. (2015). Prevalence of non-suicidal self-injury and distinct groups of self-injurers in a community sample of adolescents. *Social Psychiatry and Psychiatric Epidemiology*, *50*(7), 1163-1171. doi: 10.1007/s00127-015-1060-z.

Sotrez-Alvarez, D., Herring, A. H., & Siega-Riz, A. (2013). Latent transition models to study women's changing of dietary patterns from pregnancy to 1 year postpartum. *American Journal of Epidemiology, 177*(8), 852-861. doi: 10.1093/aje/kws303

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6): 1292-1306. doi: 10.1037/0021-9010.91.6.1292.

Tein, J-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 640-657. doi: 10.1080/10705511.2013.824781.

Vermunt, J. K., & Langeheine, R., & Böckenholt, U. (1999). Discrete time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics, 24*(2), 179-207. doi: 10.2307/1165200.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis,* (pp. 56-85). Cambridge, UK: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2013). *LG-Syntax User's Guide: Manual for Latent GOLD 5.0 Syntax Module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp.

373-385). Burlington, MA: Elsevier.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10-18. doi: 10.1111/j.1750-8606.2009.00110.x.

Williams, J., Miller, S., Cutbush, S., Gibbs, D., Clinton-Sherrod, M., & Jones, S. (2015). A latent transition model of effects of a teen dating violence prevention initiative. *Journal of Adolescent Health* , *56*(2), S27-S32. doi: 10.1016/j.jadohealth.2014.08.019.

Wirth, R. J. (2008). *The effects of measurement non-invariance on parameter estimation in latent growth models* (Doctoral dissertation). Retrieved from https://cdr.lib.unc.edu/record/uuid:50fe21c8-052d-4995-bc09-c4c30f7beb41.

Table 1. False positive detection rates (Proportion of replications where nonequivalence is detected when equivalence holds)

			Free-baseline				Constrained-baseline			
	Measurement Strength	*N*	AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
Eq.	Strong	1000	.04	.00	.03	.15	.07	.00	.06	.06
		300	.06	.00	.05	.13	.08	.00	.07	.06
		100	.08	.00	.07	.11	.12	.00	.10	.07
	Medium	1000	.03	.00	.02	.09	.05	.00	.04	.05
		300	.05	.00	.03	.12	.07	.00	.06	.06
		100	.08	.00	.07	.15	.11	.00	.08	.13
	Weak	1000	.05	.00	.04	.13	.06	.00	.05	.07
		300	.11	.00	.08	.25	.08	.00	.07	.13
		100	.11	.00	.08	.33	.13	.00	.11	.15
Noneq. [a]	Strong		.05	.00	.04	.15	.09	.00	.07	.06
	Medium		.06	.00	.05	.16	.09	.00	.08	.09
	Weak		.11	.00	.10	.27	.14	.00	.12	.15

[a] Equivalent items in nonequivalent conditions.

Table 2. True positive detection rates for strong measurement strength conditions (Proportion of replicates where existing nonequivalence is detected).

			Free-baseline				Constrained-baseline			
Items	Degree of Noneq.	*N*	AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
Item 4-5-6[a] (States 1 & 2)	Large	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		300	1.00	1.00	1.00	1.00	1.00	.86	1.00	1.00
		100	.88	.23	.86	.86	.86	.19	.84	.86
	Medium	1000	1.00	1.00	1.00	1.00	1.00	.99	1.00	1.00
		300	.97	.35	.97	.98	.98	.26	.96	.97
		100	.59	.04	.54	.52	.62	.03	.57	.63
	Small	1000	.96	.22	.96	.98	.97	.14	.96	.96
		300	.54	.00	.51	.59	.57	.00	.53	.56
		100	.21	.00	.17	.22	.22	.00	.19	.18
Item 1-2-3 (State 1)	Large	1000	1.00	.74	1.00	1.00	1.00	.41	1.00	1.00
		300	.78	.05	.75	.89	.68	.006	.61	.77
		100	.37	.003	.32	.33	.32	.01	.28	.31
	Medium	1000	.97	.19	.97	.99	.93	.09	.93	.94
		300	.54	.02	.51	.58	.43	.003	.41	.52
		100	.26	.00	.23	.18	.27	.00	.24	.29
	Small	1000	.62	.00	.58	.67	.57	.00	.53	.59
		300	.21	.00	.16	.21	.22	.00	.18	.20
		100	.10	.00	.09	.08	.13	.00	.11	.13
Item 4 (States 1 & 2)	Large	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		300	1.00	.98	1.00	1.00	1.00	.99	1.00	1.00
		100	.93	.37	.92	.86	.95	.41	.95	.94
	Medium	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		300	.98	.44	.98	.98	.98	.51	.98	.98
		100	.60	.07	.53	.56	.63	.12	.61	.62
	Small	1000	.98	.20	.98	.98	.99	.25	.98	.98
		300	.61	.03	.57	.61	.66	.03	.64	.65
		100	.25	.02	.23	.19	.35	.02	.30	.24
Item 1 (State 1)	Large	1000	1.00	.98	1.00	1.00	1.00	1.00	1.00	1.00
		300	.91	.25	.91	.93	.97	.34	.94	.97
		100	.60	.02	.56	.43	.68	.01	.64	.70
	Medium	1000	1.00	.46	1.00	1.00	1.00	.56	1.00	1.00
		300	.66	.02	.65	.70	.75	.02	.71	.80
		100	.26	.01	.25	.20	.39	.01	.35	.46
	Small	1000	.56	.00	.54	.60	.68	.01	.67	.71
		300	.16	.00	.16	.19	.25	.00	.25	.25
		100	.09	.00	.07	.08	.14	.00	.13	.11

[a] For multiple items the average proportion is presented.

Table 3. True positive detection rates for medium measurement strength conditions (Proportion of replicates where existing nonequivalence is detected).

			Free-baseline			Constrained-baseline				
Items	Degree of Noneq.	*N*	AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
Item 4-5-6[a] (States 1 & 2)	Large	1000	1.00	.99	1.00	1.00	1.00	.96	1.00	1.00
		300	.97	.25	.96	.99	.94	.18	.94	.93
		100	.62	.04	.58	.76	.51	.00	.49	.50
	Medium	1000	1.00	.85	1.00	1.00	1.00	.64	1.00	1.00
		300	.85	.06	.84	.89	.76	.03	.73	.77
		100	.36	.01	.33	.43	.31	.01	.28	.38
	Small	1000	.83	.02	.81	.86	.76	.01	.73	.76
		300	.33	.00	.30	.41	.35	.00	.32	.35
		100	.19	.00	.17	.30	.18	.01	.15	.14
Item 1-2-3 (State 1)	Large	1000	.82	.02	.79	.96	.70	.01	.66	.69
		300	.32	.00	.29	.64	.30	.00	.27	.25
		100	.24	.00	.20	.41	.17	.00	.15	.18
	Medium	1000	.72	.00	.70	.84	.60	.00	.57	.66
		300	.24	.00	.20	.46	.23	.00	.21	.28
		100	.18	.00	.16	.35	.13	.00	.11	.15
	Small	1000	.29	.00	.27	.39	.29	.00	.26	.31
		300	.09	.00	.08	.17	.13	.00	.11	.14
		100	.13	.00	.11	.16	.13	.00	.11	.15
Item 4 (States 1 & 2)	Large	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		300	1.00	.75	1.00	1.00	1.00	.83	1.00	1.00
		100	.75	.16	.78	.81	.83	.20	.82	.84
	Medium	1000	1.00	.96	1.00	1.00	1.00	.99	1.00	1.00
		300	.89	.13	.88	.91	.92	.24	.92	.92
		100	.49	.03	.46	.54	.57	.02	.55	.61
	Small	1000	.82	.02	.82	.86	.92	.04	.89	.91
		300	.31	.00	.25	.34	.40	.00	.35	.39
		100	.14	.00	.10	.32	.23	.00	.20	.21
Item 1 (State 1)	Large	1000	1.00	.60	1.00	1.00	1.00	.83	1.00	1.00
		300	.71	.02	.70	.77	.83	.04	.83	.87
		100	.34	.00	.32	.31	.38	.01	.34	.42
	Medium	1000	.93	.06	.91	.96	1.00	.15	.99	1.00
		300	.38	.01	.35	.47	.55	.01	.51	.61
		100	.21	.00	.16	.28	.20	.00	.18	.25
	Small	1000	.31	.00	.29	.39	.45	.00	.42	.46
		300	.12	.00	.10	.16	.18	.00	.18	.22
		100	.14	.00	.12	.13	.12	.00	.10	.15

[a] For multiple items the average proportion is presented.

Table 4. True positive detection rates for weak measurement strength conditions (Proportion of replicates where existing nonequivalence is detected).

			Free-baseline			Constrained-baseline				
Items	Degree of Noneq.	*N*	AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
Item 4-5-6[a] (States 1 & 2)	Large	1000	.99	.31	.99	.99	.95	.09	.93	.87
		300	.73	.00	.71	.86	.47	.00	.45	.41
		100	.40	.01	.37	.54	.30	.01	.26	.23
	Medium	1000	.93	.03	.91	.96	.84	.01	.80	.83
		300	.47	.00	.44	.72	.38	.00	.34	.40
		100	.34	.00	.27	.46	.25	.00	.22	.21
	Small	1000	.52	.00	.49	.63	.43	.00	.40	.47
		300	.22	.00	.19	.43	.14	.00	.12	.24
		100	.28	.00	.23	.31	.24	.00	.20	.20
Item 1-2-3 (State 1)	Large	1000	.27	.00	.22	.78	.14	.00	.13	.06
		300	.29	.00	.26	.61	.15	.00	.13	.14
		100	.21	.00	.17	.35	.24	.00	.22	.16
	Medium	1000	.29	.00	.28	.64	.19	.00	.17	.17
		300	.22	.00	.19	.47	.12	.00	.10	.13
		100	.22	.00	.16	.33	.18	.00	.15	.19
	Small	1000	.14	.00	.11	.32	.11	.00	.09	.18
		300	.16	.00	.14	.36	.13	.00	.11	.18
		100	.15	.00	.12	.33	.17	.00	.17	.19
Item 4 (States 1 & 2)	Large	1000	1.00	.99	1.00	1.00	1.00	1.00	1.00	1.00
		300	.93	.11	.93	1.00	1.00	.51	1.00	.83
		100	.62	.04	.56	.74	.73	.09	.71	.49
	Medium	1000	.99	.49	.99	.99	1.00	.84	1.00	1.00
		300	.70	.01	.67	.83	.92	.08	.87	.68
		100	.38	.00	.34	.42	.52	.00	.50	.30
	Small	1000	.68	.00	.66	.74	.77	.00	.74	.76
		300	.26	.00	.21	.39	.35	.00	.30	.38
		100	.18	.00	.14	.31	.23	.00	.21	.25
Item 1 (State 3)	Large	1000	.88	.04	.87	.97	.99	.31	.99	.73
		300	.42	.00	.36	.64	.61	.01	.57	.47
		100	.23	.00	.20	.39	.41	.00	.35	.28
	Medium	1000	.68	.00	.65	.77	.85	.03	.85	.78
		300	.28	.00	.25	.45	.35	.00	.31	.40
		100	.17	.00	.13	.32	.25	.00	.25	.22
	Small	1000	.28	.00	.25	.34	.34	.00	.32	.44
		300	.13	.00	.13	.33	.10	.00	.08	.19
		100	.15	.00	.14	.25	.19	.00	.16	.20

[a] For multiple items the average proportion is presented.

Table 5. Profile of the 4-state latent Markov model classifying substance users assuming measurement invariance and linear heterogeneous transitions.

		State 1 - Alcohol	State 2 - Alcohol & Marijuana	State 3 - None	State 4 - All
Alcohol					
	No	.018	.020	.893	.005
	Yes	.982	.980	.107	.995
Marijuana					
	No	.978	.226	.989	.078
	Yes	.022	.774	.011	.922
Drugs					
	No	.976	.932	.991	.157
	Yes	.024	.068	.009	.843
Prevalence		.428	.239	.187	.146
Transition[a]					
	State 1	.845	.245	.229	.027
	State 2	.100	.545	.113	.233
	State 3	.041	.010	.637	.071
	State 4	.014	.201	.020	.669

[a] Columns are originating states at $t-1$, rows are the probabilities to transition to the current state at t. E.g. there is a probability of .637 that non users stay in their state, and a probability of .229 to start belonging to the alcohol state.

Table 6. Model fit statistics for the latent Markov models testing possible measurement variance. P-values for the likelihood ratio and Wald or Score test between parentheses.

	Free-baseline (Item constrained)					Constrained-baseline[a] (Item freed)				
	Pars.	AIC	BIC	L^2	Wald[b]	Pars.	AIC	BIC	L^2	Score[c]
Baseline	90	13771	14218	10.080 (.073)	-	75	13824	14197	-	-
Alcohol	85	13768	14191	7.475 (.188)	6.675 (.250)	80	13824	14222	10.080 (.073)	8.885 (.110)
Marijuana	85	13783	14206	22.763 (.000)	23.157 (.000)	80	13773	14171	60.965 (.000)	57.271 (.000)
Drugs	85	13774	14196	13.314 (.021)	13.177 (.022)	80	13784	14181	50.474 (.000)	32.179 (.000)

[a] For the free baseline model the respective item is constrained leaving two unconstrained items, for the constrained baseline model the respective item is freed leaving two constrained items.

[b] Wald and Score tests for the direct effect of observed time on the indicator variable taken from the baseline model.

[c] Score tests are obtained using the variance through the expected information matrix, for details see Oberski & Vermunt (*Under review*).

Table 7. Model fit statistics for the latent Markov models testing longitudinally changing measurement variance: Testing the $P(Y_{tj} \mid S_t, T \mid S_t)$ specification to the $P(Y_{tj} \mid S_t, T)$ models in Table 6. P-values for the likelihood ratio and Wald or Score test between parentheses.

	Free-baseline (Item constrained)					Constrained-baseline[a] (Item freed)				
	Pars.	AIC	BIC	L^2	Wald[b]	Pars.	AIC	BIC	L^2	Score[c]
Baseline	135	13805	14475		-	75	13824	14197		-
Alcohol	115	13785	14356	43.429 (.054)	10.655 (.780)	95	13826	14298	27.780 (.023)	15.289 (.430)
Marijuana	115	13806	14377	37.431 (.165)	48.951 (.000)	95	13781	14253	21.929 (.110)	21.726 (.110)
Drugs	115	13785	14362	43.280 (.055)	105.243 (.000)	95	13795	14268	18.528 (.236)	21.550 (.120)

[a] For the free baseline model the respective item is constrained leaving two unconstrained items, for the constrained baseline model the respective item is freed leaving two constrained items.

[b] Wald and Score tests for the direct effect of observed time on the indicator variable taken from the baseline model.

[c] Score tests are obtained using the variance through the expected information matrix, for details see Oberski & Vermunt (*Under review*).

Table 8. Response probabilities per measurement occasion in the 4-state model with heterogeneous transitions and measurement variance for marijuana and drug use indicators.

			Measurement Invariant				Measurement Variant			
	T	Ages	Alc	Alc & MJ	None	All	Alc	Alc & MJ	None	All
Alcohol	2	12-18	.983	.980	.107	.995	.982	.979	.106	.997
	5	15-21	.983	.980	.107	.995	.982	.979	.106	.997
	6	18-24	.983	.980	.107	.995	.982	.979	.106	.997
	7	21-27	.983	.980	.107	.995	.982	.979	.106	.997
	8	24-30	.983	.980	.107	.995	.982	.979	.106	.997
	9	27-33	.983	.980	.107	.995	.982	.979	.106	.997
Marijuana	2	12-18	.022	.774	.011	.922	.020	.770	.006	.930
	5	15-21	.022	.774	.011	.922	.099	.948	.033	.986
	6	18-24	.022	.774	.011	.922	.022	.784	.007	.935
	7	21-27	.022	.774	.011	.922	.021	.783	.007	.934
	8	24-30	.022	.774	.011	.922	.012	.662	.003	.885
	9	27-33	.022	.774	.011	.922	.013	.682	.004	.894
Drugs	2	12-18	.024	.068	.009	.843	.004	.039	.003	.658
	5	15-21	.024	.068	.009	.843	.011	.088	.007	.821
	6	18-24	.024	.068	.009	.843	.014	.112	.009	.857
	7	21-27	.024	.068	.009	.843	.019	.151	.012	.894
	8	24-30	.024	.068	.009	.843	.018	.145	.012	.890
	9	27-33	.024	.068	.009	.843	.031	.224	.020	.933

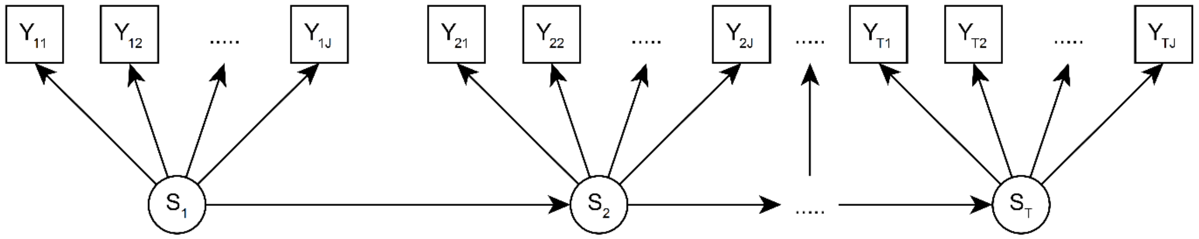

Figure 1. Latent Markov model with T measurement occasions and J items

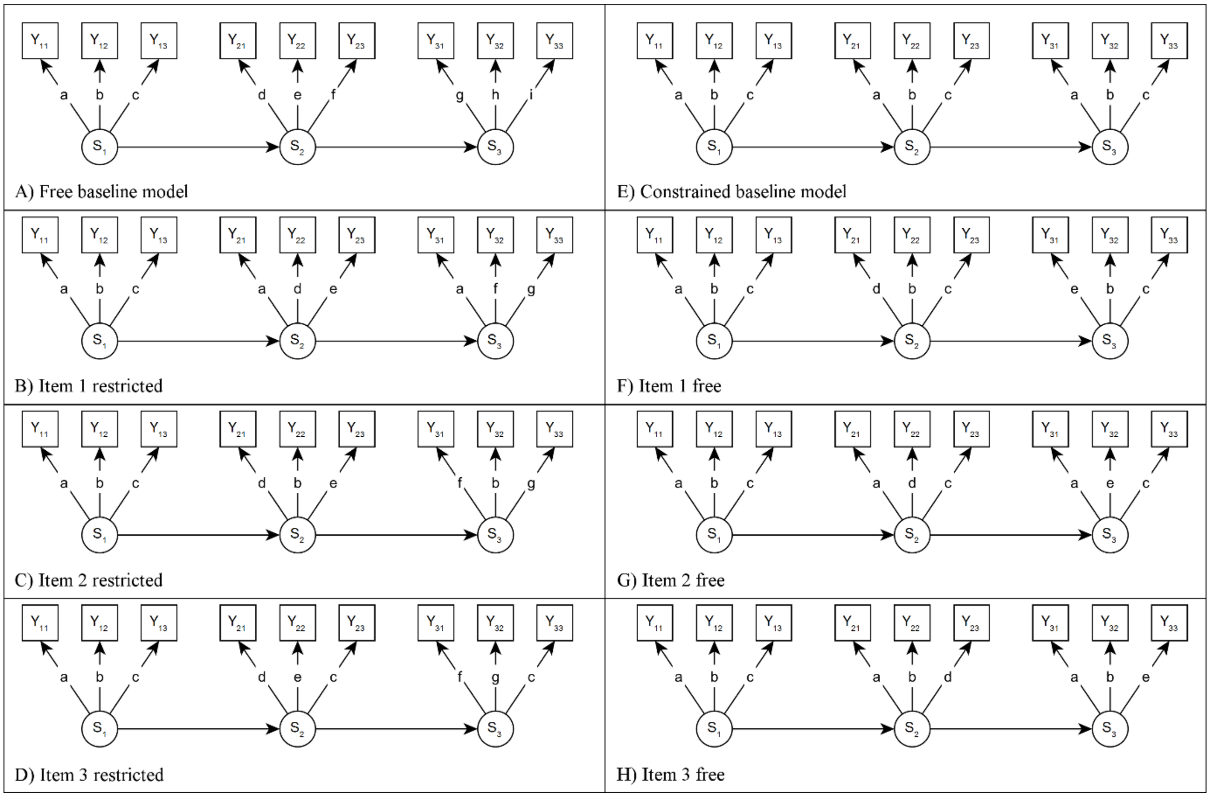

Figure 2. Latent Markov models with and without parameter constraints: Arrow labels indicate parameter estimates where two identical labels indicate a longitudinal constraint. E.g. in model B the first item is constrained to be invariant.