DETECTING MEASUREMENT NONEQUIVALENCE WITH LATENT MARKOV MODEL

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

Abstract

In longitudinal studies, and thus also when applying latent Markov (LM) models, it is important to test whether measurements are equivalence over time because otherwise it remains unknown whether an observed change is a true change or whether it caused by a change of the measurement of the construct of interest. However, typically, the testing for measurement equivalence when applying latent Markov (LM) models is neglected. In this study, we investigated two analytic strategies for testing measurement equivalence in the context of the LM model; that is, a SEM-like approach starting from a free baseline model and an IRT-like approach starting from a constrained baseline model. Using a simulation study, we determined the true and false positive rates in detecting nonequivalent items for the AIC, BIC, likelihood-ratio, score, and Wald statistics. Our simulation results indicate that regardless of the analytic strategy used, the power for detecting measurement nonequivalence in LM models is high as long as there is either a sufficiently large differential item functioning or a sufficiently strong measurement.

Key words: latent Markov model, latent transition model, measurement equivalence, baseline model

[1] Department of Methodology and Statistics, Tilburg University

P.O. Box 90153, 5000LE Tilburg, The Netherlands

[2] Department of Psychology, Dokuz Eylul University

Dokuz Eylül Üniversitesi, Edebiyat Fakültesi Psikoloji Bölümü Tınaztepe Yerleşkesi Buca İzmir, Türkiye

In social and behavioral sciences, an important topic of study is how individual change over time, which can be studied by collection data from the same individuals at multiple time points. To deal with measurement error in an appropriate manner, such longitudinal studies often use latent variable techniques such as item response theory (IRT) and structural equation modeling (SEM). Another popular tool is the Latent Markov (LM) model, also referred to as latent transition model, which is a longitudinal version of the latent class model (see, Collins & Lanza, 2010; Goodman, 1974; Vermunt & Magidson, 2002). Different from latent class models, LM models provide information on the probability of changing between latent states (classes) over time. Parallel with the development in statistical packages such as Latent GOLD (Vermunt & Magidson, 2013), Mplus (L. Muthen & Muthen, 1998-2007), and PROC LTA (Lanza & Collins, 2008), 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), eating behavior (Sotrez-Alvarez et al., 2013), abnormal psychology (Connell et al., 2008), and health psychology (Williams et al., 2015).

Most longitudinal studies assume that the instruments used measure the same constructs over time, which is also referred to as measurement equivalence. At the same time, 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).

Despite the growing body of literature on longitudinal measurement (in)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. One the other hand, it is understandable because there is no generally accepted strategy for investigating measurement equivalence. For instance, it is unclear whether one should use the SEM-like approach in which one starts with a free baseline model or the IRT-like approach in which one starts with a constrained baseline model.

Because a comparison of the SEM- and IRT-like analytic strategies in LM modeling is lacking, in this study we provide a detailed comparison of the performance of the constrained and free baseline approaches for studying measurement equivalence in LM models. Our aim is to provide an advice to applied researchers using LM models on which strategy to use in combination with what types of statistical tests.

The next section reviews the latent Markov model and the corresponding strategies for investigating measurement equivalence. Then, we describe the design of our simulation study and present its results. The final section provides a discussion, recommendations for applied researchers, and suggestions for future research.

Latent Markov Model

Let Y_{tj} be one of P observed variables or items measured at T occasions, where j = 1, 2, 3…,P and t = 1, 2,..,T , and let S_t represent the latent state occupied at time point t. The corresponding LM model for these variables for T=3 and P=3 is depicted in Figure 1, which shows clearly the two basic assumptions of the LM model; that is, the local independence and the first-order Markov assumption. The former implies that observed responses are independent of one another conditional on the latent states occupied at the T time points, and the latter implies that the state at time point t depends only on the state occupied at the previous time point t-1.

[FIGURE 1]

In fact, LM models consist of two parts: a measurement part containing the item response probabilities and a transition part containing the initial latent state probabilities and the latent transition probabilities. The item response probabilities $P(Y_{tj} \mid S_t)$ define how the items are related to the latent states, and thus determine how the latent states should be interpreted. As discussed in more detail below, though item response probabilities can be allowed to differ across time points, practitioners usually assume these to remain same over the time, which amounts to assuming measurement equivalence. The initial latent state probabilities $P(S_1)$ provide information on the prevalence of the latent states at the first measurement occasion. For the analysis of change, the parameters of main interest are the transition probabilities $P(S_t \mid S_{t\text{-}1})$, which indicate how likely it is to move across latent states between consecutive measurement occasions. More specifically, it is the probability of being in latent state S_t at time t conditional on being in latent state $S_{t\text{-}1}$ at time t-1.

Testing for Measurement Equivalence

Measurement equivalence implies that the latent structure (the number and the definition of the latent states) is the same across time points; that is, that we can validly impose across-time equality restrictions on item response probabilities $P(Y_{tj} \mid S_t)$. In contrast, measurement nonequivalence or differential item functioning (DIF) occurs when item response probabilities turn out to be different across occasions for some or for all items. If some but not all items have different response probabilities across time points, we are in the partial equivalence situation. At the extreme, we have the situation in which all measurement model parameters need to be allowed to be time specific, which corresponds to the complete nonequivalence situation. In such a fully unconstrained LM model – also referred to as the basic LM model (Bartolucci, Farcomeni, & Pennoni, 2013) – the definition of the latent states may fully

change over time, making the interpretation of the transition probabilities rather difficult, especially in models for more than a few time points.

It is important to note 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, we may still perform tests of equivalence by expanding the number of states, where some states are assumed not to occur at some of the time points. For example, in a study investigating non-suicidal self-injury (NSSI) behaviors, one may encounter four states at older ages, say experimental NSSI, mild NSSI, self-cutting, and multiple NSSI, and find out that the self-cutting behavior does not yet exist at the younger ages (Somer et al., 2015). Then, a researcher can estimate a four-state latent Markov model in which self-cutting state probability is fixed to zero at the first measurement occasion(s). In such a model, the response probabilities may still be equal across time points.

After deciding on the number of latent states needed, overall measurement equivalence can basically be tested by comparing two models: a constrained model with identical item response parameters and a free model with the item response probabilities free to vary across time points (Collins and Lanza, 2010). But, such a comparison does not provide information about possible partial equivalence with item-specific differences which are often referred to as DIF. Generally, researchers aim to find the most parsimonious model that is theoretically relevant, statistically sound, and easy to interpret. Therefore, when the constrained model is reject in favor of the unconstrained model, partially equivalent models become important, also because in a partially equivalent model, the latent states may remain comparable across time-points.

A practical problem in the search for an adequate partial equivalent model is that there is a large number of possible model comparisons that can be made. Consider a three-state LM model for three time points and six dichotomous items. Such a model contains 54

measurement parameters (3 occasions x 3 states x 6 items) that can be constrained. Moreover, DIF may occur for a single item or for several items, for as single or for several time points, and for a single or for several states. This shows that a clear strategy is needed to make the investigation of item-level (in)equivalences feasible. We will investigate two possible strategies which both involve comparing nested models. One strategy involves testing whether imposing item-level contraints in the free (fully nonequivalent) baseline model deteriorates the fit and the other whether relaxing constraints per item in the constrained (fully equivalent) baseline model improves the fit.

[FIGURE 2]

Model 2a depicted in Figure 2 represents the free baseline model where item response probabilities are free to vary over time (which is indicated by using different letters for the state-item relationships). In the free baseline strategy, one specifies a series of models with a single item restricted to be equivalent (Models 2b, 2c, and 2d). If the single-item constrained model does not deteriorate the fit worse, we can conclude that the item concerned can be assumed to be equivalent. In contrast, if the restricted model (say Model 2b) is rejected in favor of the free baseline model, the item concerned should be flagged as nonequivalent. A (statistical) advantage of using the free baseline model as the starting point is that the alternative model in the test is always a model that fits the data. However, a disadvantage is that the free baseline is not very parsimonious and that moreover its interpretation may be difficult if there are several nonequivalent items.

Alternatively, one may use the constrained model in which all items are assumed equivalent (Model 2e) as the baseline model, and subsequently free the across-time restrictions on the response probabilities per item (Models 2f, 2g, and 2h). Using this strategy, if the constrained baseline model is rejected in favor of the partially unrestricted alternative, the item concerned can be labeled as nonequivalent. A (statistical) disadvantage of this

strategy is that the alternative model may itself not fit the data, which occurs when there are multiple longitudinal DIF items.

Regardless of the baseline model strategy that is used, with a *P*-item scale, *P*+1 models should be estimated, which is almost always feasible in practice. In addition, after identifying the items which are nonequivalent, the model with freely estimated nonequivalent items and restricted equivalent items should be estimated. This model can be treated as the final partially equivalent model.

Now, let us turn to the statistics that one may use for the model comparisons to be performed. Since we are comparing nested models, the most obvious choice is the likelihood-ratio (LR) test, which can be computed by taking either minus twice the difference in log-likelihood value between the null and the alternative model or their difference in goodness-of-fit likelihood-ratio chi-squared (L^2) value. For instance, when using the free baseline approach and testing the equivalence of the first item (see Figure 2), the LR test equals:

$$LR = L^2_{2b} - L^2_{2a},$$

where *2a* represents the free baseline model and *2b* represents the model with item 1 constrained. Similarly, we can compare models with one-item free and the constrained baseline model, for example, models 2f and 2e. Since we are comparing nested models, the LR test statistic follows a central chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the alternative and the null model. With dichotomous items, 3 states, and 3 time points, the number of degrees of freedoms equals 6, yielding a critical value of 12.59 when working with a type I error equal to .05.

Alternatively, information criteria may be used for model selection, the most commonly used of which are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). If *df* is the number of degrees of freedom and *N* sample size, these can be defined as:

$$AIC = L^2 - 2 \cdot df \quad (1)$$

$$BIC = L^2 - \ln(N) \cdot df \quad (2)$$

As can be seen from Equation 2, the BIC controls for the sample size and it is usually preferred when the sample size is large. We will use these indices for comparing models with different assumptions related to measurement equivalence, where the model with the lowest value is the one that is preferred.

When using the free baseline model, one may use the Wald test as an alternative to the LR test. These two tests are asymptotically equivalent, but the Wald test has the advantage that it does not require estimating the *P* restricted models. More specifically, using the estimated parameter values and the estimated parameter variances from the free baseline model, for each item we can compute a Wald statistic testing its measurement equivalence. A disadvantage of the Wald is that it relies on the quality of the estimated variances, which may not be very good when the sample size is small.

Similarly, we can use Score or Lagrange multiplier tests as an alternative to the LR test when using the constrained baseline strategy. More specifically, after estimating the restricted baseline model, we can check for each item whether including an additional set of parameters yielding a nonequivalent model for the item concerned yields a significant improvement of fit (Vermunt & Magidson, 2016). Also the Score test is asymptotically equivalent to the LR test and, moreover, has the same advantage as the Wald test that it requires estimating only one model, in this case the constrained baseline model. A disadvantage of the Score test is that the alternative model may not hold, in which case the corresponding *p* value may be somewhat off.

In this study, we simulated data sets containing various forms of longitudinal measurement nonequivalence. Then, measurement equivalence at the item level was tested by

constraining or relaxing the equality of item response probabilities. Our research questions were:

1- Which of the two baseline strategies is more powerful when testing for longitudinal measurement equivalence in LM models?
2- Do different evaluation statistics (LR, AIC, BIC, Wald and Score) give different results?
3- Does sample size, measurement strength, number of DIF items, number of nonequivalent items, amount of DIF, and number of nonequivalent states have an effect on the power of detecting equivalences?

Methodology

Study Design

In the simulation study, we worked with dichotomous items (no/yes, disagree/agree, or fail/pass). The number of items was fixed to six, the number of measurement occasions to three, and the number of states to three. The initial state probabilities were set to be equal and the latent transition probabilities were specified to be time-homogeneous taking the following values:

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

Latent state 1 represented the non-master or disagree state by setting low probabilities for the positive response for all six items. High probabilities for the positive response were set for all items in latent state 3, characterizing it as a master state. Latent state 2 was defined to be an intermediate state with the first three items having high and the last three items having low positive response probabilities. Moreover, the nonequivalence was specified to occur between occasions two and three yielding nonequivalence at time point three.

Five factors were varied when simulating the data: the sample size, the number of nonequivalent items, the measurement strength, the amount of DIF, and the number of nonequivalent states. The two factors that were varied during the analysis were the type of baseline model and the type of statistic.

Sample size. Sample size was set to 100, 300, or 1000. We generated data sets without missing values, so sample sizes were equal for all time occasions. The choice of 300 and 1000 was consistent with the design by Collins and Wugalter (1992). We also added the condition with a sample size equal to 100 in order to check how the investigated approaches perform with small samples.

Number of nonequivalent items. The number of nonequivalent items was zero, one (item 1 or item 4), or three (items 1, 2, and 3 or item 4, 5 and 6). This corresponds to 0%, 17% and 50% contamination, respectively. Our motivation to also look at the complete equivalence condition (zero nonequivalent items) was to establish false positive rates for this situation.

Measurement strength. The association between latent states and response variables was either weak, moderate, or strong. Positive response probabilities were equal to 0.7, 0.8, or 0.9 (Collins & Wugalter, 1992; Gudicha, Schmittmann, & Vermunt, 2015; Vermunt, 2010). These conditions can also be expressed using entropy based R^2 statistics, yielding values of .51, .78, and .94, respectively.

DIF amount. The amount of DIF was set as small, medium, or large. For conditions with small DIF, the item response probability for the nonequivalent item was set .10 lower than the previous time points, for the medium DIF conditions it was .20 lower, and for the large DIF conditions it was .30 lower.

Number of Nonequivalent States. The first three items can be considered to be easy items for which both latent state 2 and 3 members had high response probabilities for the positive response. The last three items were simulated as difficult items, and only latent state 3

members had a high response probability of selecting the positive response. DIF occurred in either easy or difficult items. When DIF was in the easy item, only State 1 was affected. However, when the DIF item was a difficult item, both State 1 and State 2 members were affected.

Type of baseline model. The baseline model was either free (all item response parameters estimated freely) or constrained (all item response probabilities equal for all time occasions).

Type of model selection. LR test, AIC, BIC, and Wald test for free baseline model and the Score test for constrained baseline model were assessed.

In summary, we varied the following factors; sample size (3), number of nonequivalent items (3), measurement strength (2), DIF amount (3), and number of nonequivalent states (2). In total, we simulated 117 conditions. In the analysis, we also varied the type of baseline model (2) and the statistic used for model selection (4).

Data Analysis

Data generation and analysis were conducted using the Syntax module of the Latent GOLD 5.0 program (Vermunt & Magidson, 2013). The estimated model was always a 3-state LM model with homogenous transition probabilities. In total 14 models were estimated with each simulated data set: the free baseline model, 6 models with a single item constrained, the constrained baseline model, and 6 models with a single item unconstrained. For model selection, we used four different statistics: LR test, AIC, BIC, and either a Wald or Score test. The critical value for the chi-square statistics with six degree of freedom was 12.59. We evaluated true and false positive rates across 100 replications per simulation condition. The true positive results represent the proportion of nonequivalent items correctly identified as nonequivalent across replications. For the conditions with three nonequivalent items, the reported true positive results are the averages for the three items. The false positive results

represent the average proportion of equivalent items incorrectly flagged as being nonequivalent.

Results

False Positive Rates

As can be seen in Table 1, for the free baseline model, the false positive results for the conditions without DIF were between .03 and .11 for the AIC and .03 and .08 for the LR statistics. The false positive results of the constrained baseline model were slightly higher. The most striking result is the high false positive rates for the Wald statistics. Regardless of other factors the false positive rates were higher than .09 when using Wald tests. The BIC had .00 false positive results for all conditions, but as discussed below it also had very low power.

For the conditions with DIF, false positive results were similar to those for the conditions without DIF. As shown at the bottom of Table 1, the average false positive results ranged from .00 to .27. The highest false positive results were, similar to the conditions without DIF, found with the Wald test.

[TABLE 1]

True Positive Rates

We split the true positive results into three tables based on measurement strength, each presenting 36 conditions. Table 2, 3, and 4 show the true positive results for the strong, medium and weak measurement strength conditions, respectively. These tables are organized from large DIF to small DIF and from large sample size to small sample size.

[TABLE 2, & 3, & 4]

Sample Size

Regardless of other factors, when the sample size decreased power also decreased. This effect became more severe when DIF was small and nonequivalence was in one state. BIC was the most affected statistic by sample size, however. When the measurement strength was weak,

the impact of the sample size was more severe. In the following sections, we interpret results of the conditions with a sample size of 300 and 1000, because the effect of using a sample size of 100 was similar for most of the conditions. Note that the only exception was the case of strong measurement, large DIF, and DIF in two states, in which case, except for the BIC, the power is still higher than .90.

Measurement Strength

In the case of strong measurement (Table 2), when the DIF amount was large, both the free baseline and constrained baseline models had very high power. When the measurement strength was medium (Table 3), power decreased slightly. For instance, in the case of large DIF, large sample size, and DIF in three items and one state, the power of the AIC decreased from 1.00 to .82. As can be seen from Table 4, the conditions with weak measurement strength were the most problematic ones. In these conditions, true positive results close to one were only obtained with large DIF, nonequivalence in two states and large sample sizes regardless of the baseline model used.

DIF amount and Number of Nonequivalent States

As expected, the amount of DIF was one of the most important factors affecting power. In this study, we used three different DIF amounts with either one or two states affected, whereby fewer states affected made it harder to detect the nonequivalence. This shows that it is important to know whether nonequivalence affects just one or more states. According to the results, we see a decreasing pattern in the power depending on both the DIF amount and the type of DIF item. When DIF was on an easy item and small, it was almost never detected, especially in the weak measurement strength conditions. In the conditions with large DIF in difficult items, power was close to one for almost all statistics and both the free baseline and constrained baseline approaches. However, in the case of weak measurement, large DIF, large

sample size, and three DIF items, the power of the BIC decreased to .31 for the free baseline model and .09 for the constrained baseline model.

When nonequivalence occurred in two latent states, true positive rates were close to 1.00 especially for the large DIF and large sample size conditions. However, when the nonequivalence was in one latent state, power dropped dramatically in the weak measurement conditions even with large DIF in three items. Power was noticeably higher for the two DIF state than for the one DIF state conditions when the amount of DIF was small.

Statistics Used

Another important finding was that the BIC tended to select equivalent models and its power decreased to zero when the sample size was 300 or 100 and the amount of DIF was small. The AIC and LR statistic results were similar to one another, and also similar to the Wald and the Score test results. True positive rates of the AIC ranged between .12 to 1.00 and .10 to 1.00 for the free-baseline and constrained baseline models, respectively.

Number of DIF items

The number of DIF items was either one or three. In the conditions with strong measurement strength, power was similar for both conditions and depended mainly on the amount DIF and the type of DIF item. When the measurement strength decreased, the free baseline model approach was slightly more powerful than the constrained baseline model approach when the number of DIF items was three. Additionally, when there was only one DIF item, true positive rates were high for both approaches.

Discussion

The main objective of this study was to compare different strategies for investigating longitudinal measurement equivalence in latent Markov modeling. In the context of SEM and multi-group latent class models researchers usually prefer starting with the free baseline model, while researchers in the field of IRT modeling will typically start with the constrained

baseline model. In a recent study, Kim and Willson (2014) showed that using the constrained baseline model approach may yield false positive results in detecting nonequivalence with multigroup 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. Similarly, Kankaras, Vermunt and Moors (2011) recommended the free baseline model because it ensures that comparisons are always made with a model that fits the data. Thus, one question to be answered is which approach should be used in the LM models framework. Our simulation study suggest that the power of detecting measurement nonequivalence will be good as long as there is either a sufficiently large DIF or sufficiently strong measurement regardless of the analytic strategy used to detect nonequivalent items. Our findings also show that power improved when the strength of association between the observed variables and the latent states increased for both strategies. However, in the case of three DIF items and weak measurement strength, the power of the constrained baseline model was slightly lower than that of the free baseline model. This is reasonable because the constrained baseline model assumes that the items that are not the focus of interest – those which are constrained -- are equivalent. However, the conditions with three DIF items violate this assumption and, moreover, in our simulation corresponded with 50% contamination. Based on these results, we suggest applied researchers to use the free baseline model, especially if they think there might a large portion of nonequivalent items.

In our study, we also investigated the performance of different statistics (LR, AIC, BIC, Wald test, and Score test). One of the most striking results was the failure of the BIC in detecting nonequivalent items. Since the BIC selected equivalent models in most of the simulation conditions, we recommend not to use the BIC for investigating measurement

nonequivalence in LM models. AIC and LR tests gave similar results, but somewhat larger true positive rates for the free baseline compared to the restricted baseline approach. The results based on Wald tests and Score test showed that these statistics are also powerful especially for the researchers who would like to check equivalence without estimating several nested models. However, we strongly recommend researchers who use Wald tests to take into account its tendency to yield false positives. This implies that it may be wise to use a LR test as an extra check for items flagged as nonequivalent. Overall, the most accurate conclusions were reached when using either the AIC or LR test combined with the free baseline model.

The results also demonstrated that with a small sample the power to detect nonequivalent items may be very low. When the sample size increased the power also increased regardless of the baseline model strategy and model selection method. Future research using, for instance, the power computation methods proposed by Gudicha et al. (2015) focusing on finding the optimal sample size in measurement equivalence studies with LM model might be interesting.

In the small DIF conditions, both strategies and all statistics used had small true positive rates. Although this might suggest low power of LM models, most researchers would consider these conditions as extremely small DIF conditions. Frankly, such an item bias would not affect the transition probabilities or the measurement model parameters severely. One motivation to investigate these conditions was to see how small DIF we could detect. Further research investigating the impact of violation of measurement equivalence with different DIF amounts is needed.

In our study, both the true and the estimated models were models with homogenous latent transition probabilities. In many applications this assumption is too strong and may therefore affects the results. Further research might explore the effect of estimating the latent

transition probabilities with or without the time-homogeneity restrictions in the presence of nonequivalent items.

References

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent markov models for longitudinal data*. Boca Raton: Chapman and Hall/CRC press.

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.

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

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: 1211-1225.

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*, 1179-1259.

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*, DOI: 10.1080/10705511.2015.1014040

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*, Vol 40-2: 279-310.

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*, 44:446-456.

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, 279-300.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data. *Child Development Perspectives* , 4, 5-9.

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

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

Olivera-Aguliar, M. (2013). *Impacts of violations of longitudinal measurement invariance in latent growth models and autoregressive quasi-simplex models.* (Doctoral dissertation). repository.asu.edu/attachments/.../OliveraAguilar_asu_0010E_13164.pdf

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*, 50:1163-1171. 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* , 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.

Vermunt, J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18, 450-469.

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., & Magidson, J. (2016). *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations Inc.

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*, 10-18.

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* , 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). Available from ProQuest Dissertations and Theses database. (UMI No. 3331053).

Table 1. False Positive Results

			Free-Baseline				Constrained-Baseline			
DIF	Measurement Strength	*N*	AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
NO DIF	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
DIF	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

Table 2. True Positive results in case of strong measurement

			Free-baseline				Constrained-baseline			
DIF Items	DIF amount	*N*	AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
Item 4-5-6	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	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	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	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

Table 3. True Positive results in case of medium measurement strength

			Free-baseline				Constrained-baseline			
DIF Items	DIF amount	*N*	AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
Item 4-5-6	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	.003	.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	.006	.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	.003	.17	.30	.18	.006	.15	.14
Item 1-2-3	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	.003	.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	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	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

Table 4. True Positive results in case of weak measurement strength

DIF Items	DIF amount	*N*	Free-baseline				Constrained-baseline			
			AIC	BIC	L^2	Wald	AIC	BIC	L^2	Score
Item 4-5-6	large	1000	.99	.31	.99	.99	.95	.09	.93	.87
		300	.73	.003	.71	.86	.47	.003	.45	.41
		100	.40	.006	.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	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	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	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

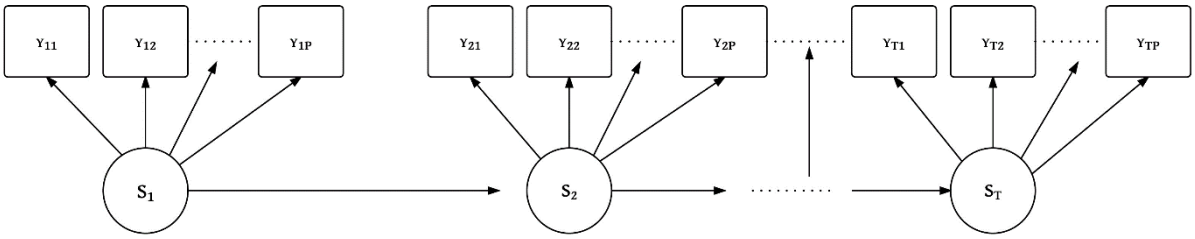

Figure 1. *First-order latent Markov model path diagram*

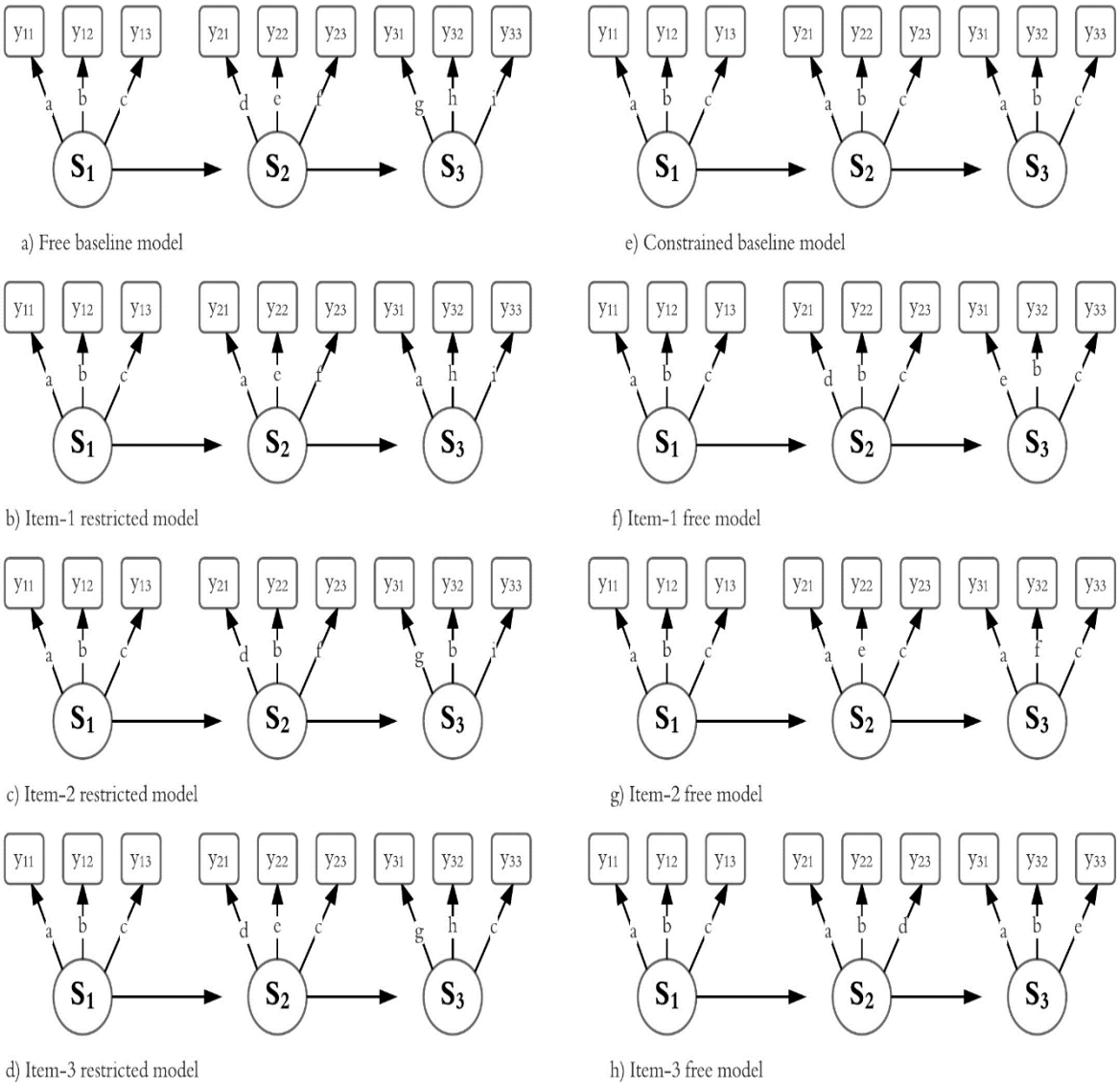

Figure 2. *Constrained and free model diagrams*

(Letters represent constrains on item response probabilities)