Assessing measurement invariance for longitudinal data through latent Markov models

Roberto Di Mari
Department of Economics and Business, University of Catania, Italy

Francesco Dotto
Department of Economics, Roma Tre University of Rome, Rome, Italy

Alessio Farcomeni
Department of Economics and Finance, University of Rome “Tor Vergata”, Rome, Italy

Antonio Punzo
Department of Economics and Business, University of Catania, Italy

Abstract

We propose a general approach to detect measurement non-invariance in latent Markov models for longitudinal data. We define different notions of differential item functioning in the context of panel data. We then present a model selection approach based on the Bayesian Information Criterion (BIC) to choose both the number of latent states and the measurement structure. We show the practical relevance by means of an extensive simulation study, and an empirical application of tolerance types toward minorities. Our results indicate that BIC is able to select the correct measurement equivalence structure more than 95% of times.

Keywords: latent Markov model with covariates, measurement invariance, differential item functioning, direct effects, model comparison

1 Introduction

Central assumptions in psychological, social and economic evaluations are that all items in a questionnaire have the same discrimination power (i.e. they are equally related to the latent trait), unidimensionality of the latent trait, and measurement equivalence - or invariance - of the scale.

In this work we focus on evaluations that are repeated over time, using the same items, for the same agents (subjects, households, etc.). The resulting longitudinal or panel data are usually treated by specifying agent-specific random effects, which capture dependence

Put address of corresponding author

and unobserved heterogeneity. Incidentally, random effects describe the latent trait, also known as the person-specific ability (De Andrade & Tavares, 2005).

Our main focus is on assessment of measurement invariance. This happens when indicators of the measurement model do not depend on covariate values, regardless of effects for the latent variable. This means that the latent trait, given the covariates, fully explains the observable behaviour of a subject. Intuitively, e.g., propensity to tolerance changes with a covariate (e.g., age); but conditionally on the covariate the meaning of an item is the same at all ages.

Violations of the measurement invariance assumption can distort the interpretation of the latent scale. In latent class models, how to detect measurement invariance and evaluating the effects of failing to ignore the presence of direct effects of the covariates on the indicators has been largely studied in recent literature (Bakk & Kuha, 2020; Vermunt & Magidson, 2020; Janssen et al., 2019).

Such violations are sometimes termed differential item functiong (DIF), since some items have differing conditional response probabilities given the covariate values. For instance, if measurement invariance is violated when comparing household portfolio prototypes across countries or regions, differences in product acquisitions might be due either to substantive differences or as differing behaviors over countries.

We describe below how a flexible model specification can be naturally used, conditionally on unidimensionality assumptions, to allow person-specific abilities to change over time, and, more importantly, to assess measurement invariance. We work with a generalization of the latent class model (Goodman, 1974) known as the latent Markov (LM) model (Collins & Wugalter, 1992; Van de Pol & De Leeuw, 1986; Van de Pol & Langeheine, 1990; Wiggins, 1973; Bartolucci et al., 2013), also referred to as hidden Markov or latent transition model. LM models describe how individuals move between latent states over time, accounting for measurement error (see, for instance, Van de Pol & De Leeuw, 1986). Examples in the social and behavioral sciences of the use of LM models include estimating the change in brand choice behavior over time (Poulsen, 1990), modeling disease progression (Jackson et al., 2003), analyzing the behavior of domestic violence batterers (Ip et al., 2010), public trust (Pennoni & Genge, 2020), measuring dynamics in households' portfolio composition (Di Mari et al., 2016; Paas et al., 2007) as well as their material deprivation (Dotto et al., 2019; Farcomeni et al., 2020).

In many applications of LM models researchers investigate possible predictors and correlates of the latent state membership. One well-known extension of the baseline LM model is obtained by including time-constant and/or time-varying covariates explaining differences in initial states and transitions across individuals and/or time points (Bartolucci & Farcomeni, 2009; Bartolucci et al., 2013, 2014; Collins & Lanza, 2010; Farcomeni, 2015; Vermunt et al., 1999). Further extensions include, for instance, dynamic latent class models (Asparouhov et al., 2017), mixed hidden Markov (Altman, 2007; Maruotti, 2011) and multilevel latent Markov models (Montanari et al., 2018; Bartolucci et al., 2011).

It is typical for LM models to assume the existence of a latent process following a first–order Markov chain with a finite number of states which affects the distribution of the response variables - measurement model. Although important, the measurement model is often not of primary interest in applied research. The researcher might rather be interested in questions like how the unobserved portfolio prototype of a household relates to the amount of hours worked per week or the country or region of residence, or whether a specific unobserved disease progression state can be predicted by the smoking status. There are two main advantages for our purposes in the use of latent transition models: first of all, finite mixtures can adapt very well to almost any underlying latent trait distribution, thus reducing bias due to misspecification of the random effects distribution. Secondly, posterior probabilities can be used to cluster subjects and evaluate their transitions over repeated questionnaire administrations.

The issue of measurement non-invariance is pervasive in latent variable models (Vriens et al., 2017, Muthén & Asparouhov, 2018, Munck et al., 2018, Kankaraš et al., 2011, Davidov et al., 2018 and Cieciuch et al., 2018). See also Vermunt (2010); Asparouhov & Muthén (2014); Kim et al. (2016); Nylund-Gibson & Masyn (2016) about bias due to ignoring the direct effects. In the latent Markov literature, Di Mari & Bakk (2018) investigate the rubustness of currently available approaches in the presence of direct effects of the covariates on the indicators, but they do not deal with locating scale (or item) level misspecification. More recent contributions include the works of Vogelsmeier et al. (2019) and Vogelsmeier et al. (2020), for tracking and diagnosing measurement non-invariance in latent Markov models for intensive longitudinal data, for continuous and ordinal responses respectively. How to detect measurement non-invariance for categorical data remains an open issue.

In this work we will define several new notions of DIF in the panel data context. In order to detect lack of measurement invariance, and select the most appropriate DIF scenario, we propose simply comparing the different options through the Bayesian Information Criterion (BIC). BIC has advantages over traditional hypothesis-testing tools (Lorah & Womack, 2019). For instance, the ability to compare non-nested models (McCoach & Black, 2008), or to make selection among several competing models (Raftery, 1995); perhaps, more in general, it allows to “show that a smaller model is better than a larger model” (Weakliem, 2004, p. 179). For these reasons, researchers tend to prefer to report BIC in order to guard against some of the weaknesses of hypothesis-testing procedures, like arbitrary selection of p-values or an increased Type I error rate when multiple models are presented for explaining variation in a data set (Lorah & Womack, 2019; Cohen, 1992). Finally and importantly, BIC is model consistent: if the true model (that is, invariance or non-invariance structure) is among the set of candidates, as the sample size grows it will be selected with probability one (Yang, 2005).

To help detecting local misspecifications, local fit measures have been proposed for latent class and latent class multilevel models (Nagelkerke et al., 2016, 2017; Magidson & Vermunt, 2004) based on Pearson residuals - available also to LM models (Vermunt & Magidson, 2016) - and a measure based on the Expected Parameter Change (EPC, Oberski, 2014; Oberski & Vermunt, 2018). Still, the issue of detecting measurement non-invariance is not yet resolved: the above local fit measures are not asymptotically chi squared and, for statistical conclusions, correct p-values are to be obtained based on resampling procedures (Oberski et al., 2013), whereas the EPC is not yet available for Markov models. Moreover, recommendations on how to select the model at scale level and power assessment of global fit measures are not available - that is, regarding possible restrictions on the kind of differential item functioning that is imposed simultaneously on all items. Although such local fit measures are gaining attention for differential item functioning detection at local (item) level, they require a prior idea on the scale-level model.

In this paper, we propose a principled and logical approach for detecting potential sources of DIF with different levels of severity, to allow applied researchers to avoid model misspecifications in longitudinal studies and get accurate estimates accounting for possible measurement non-invariance - without necessarily having to resort to overparametrized models.

To do so, we combine and extend ideas from Kankaraš et al. (2010) and Masyn (2017) to latent Markov models for longitudinal data. We start by expanding the parametrization of the measurement model of Kankaraš et al. (2010) to allow time- state-varying DIF, and all parametrization instances are nested in the overparametrized specification. We then develop a procedure to evaluate (approximate) measurement invariance at scale level based on BIC, which we assess by means of an extensive simulation study and an application on tolerance types toward minorities. Results point out that BIC allows to choose the correct model in most of the cases (more than 95% of times) and, as a consequence, precise estimates, regardless of the type of underlying DIF and intra class separation, are provided. The main take-home message is that misspecification of the measurement model causes non-negligible bias on structural model parameters: hence prior model selection at scale level is mandatory.

The paper is structured as follows. In Section 2 we introduce the modeling framework, and discuss the estimation strategy in Section 3. In Section 4 we report the results from the simulation study, and then present an empirical application on tolerance types toward minorities from the General Social Survey (Section 5). Section 6 concludes.

2 The latent Markov model with covariates

Let Y_{ith}, $h = 1, \dots, H$, denote the h-th dichotomous indicator measured for the i-th subject, $i = 1, \dots, n$, at time t, $t = 1, \dots, T$. Let additionally X_{it} denote a time-specific vector of covariates which have to be assessed for measurement invariance. Assume furthermore that there is a discrete latent variable U_{it} with support $\{1, \dots, K\}$. In the following we make assumptions of local independence (that is, that conditionally on U_{it} and possibly covariates, Y_{ith} is independent of Y_{isl} for $h, l = 1, \dots, H$ when $s \neq t$, and for all $l \neq h$ when $t = s$) and that U_{it} follows a possibly inhomogeneous first-order Markov chain (that is, conditionally on $U_{i,t-1}$ and possibly covariates, U_{it} is independent of U_{is} for $s < t - 1$).

Let $\text{logit}(x) = \log[x/(1-x)]$, and denote $\phi_{h|k} = \Pr(Y_{ith} = 1 | U_{it} = k)$. In case of measurement invariance (no DIF), we assume the data arise from the following model:

$$\begin{cases} P(Y_{it1} = y_1, \dots, Y_{itH} = y_H \mid U_{it} = k) = \prod_{h=1}^{H} \phi_{h|k}^{y_h} (1 - \phi_{h|k})^{1-y_h} \\ \log\left[\dfrac{P(U_{i1} = k \mid X_{i1})}{P(U_{i1} = 1 \mid X_{i1})}\right] = \alpha_{1k} + \beta_{1k} X_{i1} \\ \log\left[\dfrac{P(U_{it} = k \mid U_{i,t-1} = j, X_{it})}{P(U_{it} = j \mid U_{i,t-1} = j, X_{it})}\right] = \alpha_{tkj} + \beta_{tkj} X_{it}, \end{cases} \tag{1}$$

where the first equation assumes a time-homogeneous measurement model, the second equation is for $k \neq 1$ and $t = 1$, while the last equation is repeated for all $k \neq j$ and all $t > 1$. The first equation in (1) defines the measurement model, while the remaining equations define the structural model. In addition, we let $\boldsymbol{\delta}_i$ be the K initial state probabilities with generic element $\delta_{ik} = P(U_{i1} = k \mid X_{i1})$, and $\boldsymbol{Q}_{it}$ the transition probability matrix with generic element $q_{itkj} = P(U_{it} = k \mid U_{i,t-1} = j, X_{it})$.

Note that we specified a model with several parameters, which in many cases can be simplified by assuming uniformity for the latent distribution, that is, $\alpha_{tkj} = \alpha_{lkj}$ and $\beta_{tkj} = \beta_{lkj}$ for all $t > 1$ and $l > 1$. This assumption can be incorporated in the notation by dropping the subscript t from the structural model parameters. Thus, the number of free parameters to be estimated is $HK + 2(K-1) + 2[K(K-1)]$.

It is possible to relax the measurement invariance assumption by allowing for DIF. With reference to the generic h-th item, we can distinguish the following DIF scenarios. In case of an all-DIF model, the first equation of (1) is modified by letting the conditional probabilities depend on X_{it} and parameterizing

$$\text{logit}(\phi_{ht|k}) = \gamma_{hk} + \eta_{htk} X_{it} \tag{2}$$

for all $h = 1, \ldots, H$ and $t = 1, \ldots, T$. This corresponds to the nonuniform (full) DIF model in which the effect of the covariate is state- and time- dependent. In this case, the number of free parameters to be estimated is increased by HTK with respect to that of (1). Note that, by allowing a time-varying effect of the covariate on the items, we are relaxing the homogeneity assumption of the measurement model.

The other DIF scenarios can be derived from the full DIF case by imposing suitable restrictions, as shown in Figure 1. The strongest restriction is obtained by assuming the effect of the covariate on the h-th item to be state- and time- independent. The equation for the state-conditional response probabilities becomes

$$\text{logit}(\phi_{h|k}) = \gamma_{hk} + \eta_{h\cdot\cdot} X_{it}. \tag{3}$$

Measurement invariance for item h is obtained, as a special case, when $\eta_{h\cdot\cdot} = 0$. Two intermediate DIF conditions can be allowed for by specifying either time-constant or state-constant DIF, respectively

$$\text{logit}(\phi_{ht|k}) = \gamma_{hk} + \eta_{h\cdot k} X_{it} \tag{4}$$

in the first case, and

$$\text{logit}(\phi_{ht|k}) = \gamma_{hk} + \eta_{ht\cdot} X_{it} \tag{5}$$

in the second case.

3 Differential Item Functioning: a scale-level estimation strategy

In this section we describe in details the procedure to select the most plausible scale-level DIF specification. We start with ML estimation of the model parameters of the most general (full DIF) specification, from which all other nested specifications we discussed above can be derived straightforwardly (refer to Figure 1). Then we describe how a sensible initialization for computing ML estimates can be achieved, and how to classify sample units into states based on known recursions. Note that to denote the generic value assumed by U_{it}, we will use k and k_{it} interchangeably.

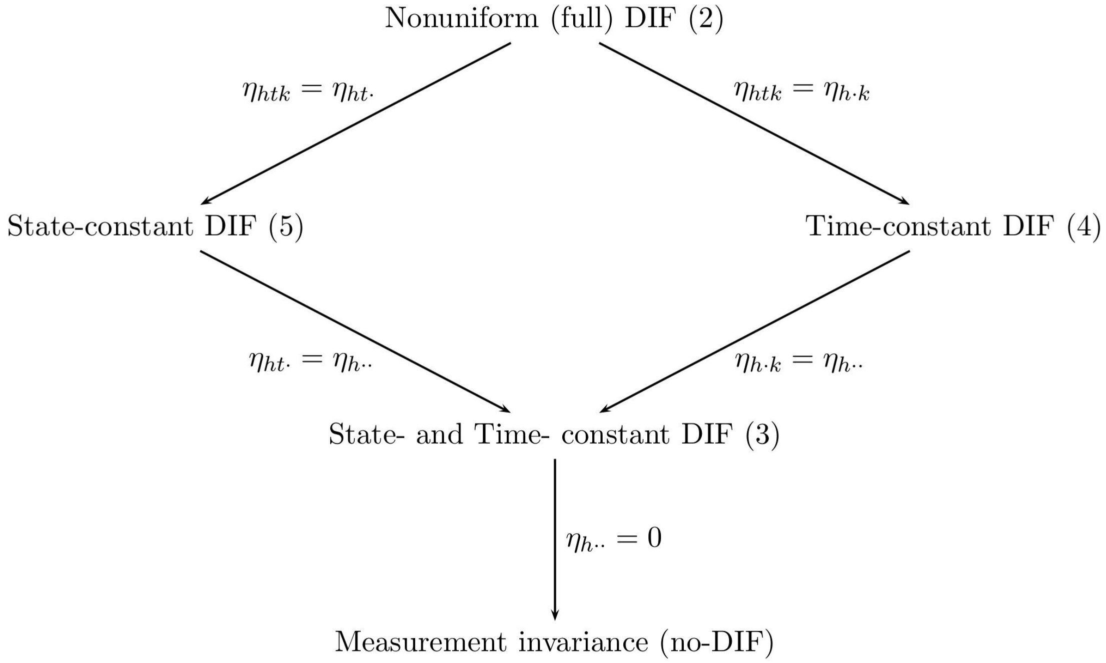
Nonuniform (full) DIF (2)
$\eta_{htk} = \eta_{ht\cdot}$
$\eta_{htk} = \eta_{h\cdot k}$
State-constant DIF (5)
Time-constant DIF (4)
$\eta_{ht\cdot} = \eta_{h\cdot\cdot}$
$\eta_{h\cdot k} = \eta_{h\cdot\cdot}$
State- and Time- constant DIF (3)
$\eta_{h\cdot\cdot} = 0$
Measurement invariance (no-DIF)

Figure 1. Hierarchy of relationships between models for the generic h-th item.

3.1 Maximum likelihood estimation

The joint distribution of the full response vector $\boldsymbol{Y}_i$ for unit i in case of full DIF can be written as

$$P(\boldsymbol{Y}_i|\boldsymbol{X}_i) = \sum_{k_{i1}=1}^{K} \dots \sum_{k_{iT}=1}^{K} P(U_{i1} = k_{i1}|X_{i1}) \prod_{t=2}^{T} P(U_{it} = k_{it}|U_{i,t-1} = k_{i,t-1}, X_{it}) \prod_{t=1}^{T} P(\boldsymbol{Y}_{it}|X_{it}, U_{it} = k_{it}), \tag{6}$$

where k_{it} denotes one of the possible latent states for unit i at time t, $i = 1, \dots, n$ and $t = 1, \dots, T$. By letting $\boldsymbol{\theta}_{\text{FULL}}$ be the vector of all model parameters for the full DIF model, parameter estimation is normally carried out by maximizing the sample log-likelihood function

$$\ell(\boldsymbol{\theta}_{\text{FULL}}) = \sum_{i=1}^{n} \log P(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) \tag{7}$$

with respect to $\boldsymbol{\theta}_{\text{FULL}}$. This is typically accomplished either by iterative algorithms like the EM-algorithm (Baum et al., 1970 and Dempster et al., 1977), or by direct maximization. Here we illustrate how to find ML estimates of $\boldsymbol{\theta}_{\text{FULL}}$ by direct maximization. To do so, we will exploit a well known recursion (Baum et al., 1970) that allows to compute the (log)-likelihood quite efficiently (see also Bartolucci et al., 2013 and Zucchini et al., 2016), which we will directly maximize to find the vector of unknown model parameters.

Let us introduce the so-called forward probabilities

$$a_{it}(k) = P(\boldsymbol{Y}_{i1}, \dots, \boldsymbol{Y}_{it}, U_{it} = k_{it} \mid X_{i1}, \dots, X_{it}), \tag{8}$$

that can be seen as the probability of observing a partial sequence ending up in state k_{it} at time t for unit i. This quantity can be computed recursively as

$$a_{i1}(k) = \delta_{ik} P(\boldsymbol{Y}_{i1} \mid U_{i1} = k, X_{i1}), \tag{9}$$

and

$$a_{it}(k) = \sum_{m=1}^{K} a_{i,t-1}(m) q_{itkm} P(\boldsymbol{Y}_{it} \mid X_{it}, U_{it} = k). \tag{10}$$

The log-likelihood can then be computed as (Zucchini et al., 2016)

$$\ell(\boldsymbol{\theta}_{\mathrm{FULL}}) = \sum_{i=1}^{n} \log \sum_{m=1}^{K} a_{iT}\left(m\right). \tag{11}$$

Note that the recursions of Equations (9), (10) and (11) can be performed under the model assumption (Sec. 2). In addition note that formulas (8)-(11) can be adapted straightforwardly to accommodate any of the constrained specifications presented above. Inference on the model parameters, under each specification, can be done using the observed information matrix estimated from the inverse of minus the (numerical) Hessian.

3.2 Initialization of the model parameters

Numerical solvers usually require starting values for the model parameters. We suggest a hierarchical initialization strategy that provides a warm start for subsequent model fitting. The strategy works as follows.

1. Fit a simple latent class model (without covariates) on the pooled data. Use the response probabilities to initialize $\phi_{h|k}$ for all $h = 1, \ldots, H$ and $k = 1, \ldots, K$.

2. Use posterior class membership probabilities from the simple latent class model to classify units with a maximum a posteriori rule and take such values as if they were realizations of an observed (not latent) Markov process and compute the initial and transition probabilities.

3. (Simple LM model) Use the log-linear transformation to map the probabilities from the previous step into intercept-only log-linear models: these are used to initialize α_{1k} and α_{kj}, for all $k, j = 1, \ldots, K$. Set β_{1k} and β_{kj} to zero, for all $k, j = 1, \ldots, K$.

4. Fit the no DIF LM model. Use the model parameter estimates to fit all other DIF configurations - with η_{htk} parameters initialized to zero.

3.3 Classification with posterior membership probabilities

To classify observations within states by means of posterior membership probabilities, we introduce the so-called backward probabilities

$$b_{it}(k) = P(\boldsymbol{Y}_{i,t+1}, \ldots, \boldsymbol{Y}_{i,T} \mid U_{it} = k_{it}, X_{i1}, \ldots, X_{i,t+1}). \tag{12}$$

In words, $b_{it}(k)$ indicates the probability, given state membership at time t, to observe $(\boldsymbol{Y}_{i,t+1}, \dots, \boldsymbol{Y}_{i,T})$. These probabilities can be computed, similarly to the forward probabilities, recursively as follows

$$b_{it}(k) = \sum_{m=1}^{K} q_{i,t+1,mk} P(\boldsymbol{Y}_{i,t+1} \mid X_{i,t+1}, U_{i,t+1} = m) b_{i,t+1}(m), \tag{13}$$

with

$$b_{iT}(k) = 1, \tag{14}$$

for $k = 1, \dots, K$.

At time t, posterior membership probabilities for unit i can be computed as

$$P(U_{it} = k \mid \boldsymbol{Y}_i, \boldsymbol{X}_i) = \frac{a_{it}(k) b_{it}(k)}{\sum_{m=1}^{K} a_{it}(m) b_{it}(m)}. \tag{15}$$

3.4 Model selection

3.4.1 Number of latent states. The first choice to be made concerns the number of states (K). Typically, to do so, information criteria like AIC or BIC are used. However, although using these criteria is common practice, their performance has not been studied enough in detail - especially in connection with LM models (Bartolucci et al., 2013). The current best practice in latent class analysis (LCA) is to select the number of states based on unconditional models - i.e. without covariates (Asparouhov & Muthén, 2014, Masyn, 2017 and Vermunt, 2010).

Based on available recommendations for LCA, to avoid selecting a number of states larger than optimal - due to, for instance, misspecifications like unmodelled local dependencies and DIF - we also suggest selecting K based on a LM model without covariates. The unconditional model can be fitted using the same initialization strategy of Section 3.2 for an increasing number of states. Then, the optimal K is selected minimizing BIC.

3.4.2 Specification of DIF structure. The next step of model selection involves choosing the DIF specification. For a given K, we propose to fit a LM model with covariates with no DIF, and DIF as specified in Figure 1 - see also Equations (2), (3), (4) and (5). The model that is then selected is the one minimizing BIC.

From the point of view of interpreting the model, the two most interesting configurations are the no DIF and the time- and state- constant DIF. By contrast, a full DIF model, i.e. with time and state varying DIF, is selected if the latent structure is not enough to describe the association between items and between items and covariate(s). In this case, the resulting overall model is hardly interpretable.

4 Simulation Study

The aim of this simulation study is twofold: first, we show the potentially dangerous consequences of ignoring the direct effect(s) of covariate(s) on the measurement model and investigate the benefits of selecting the proper DIF model in terms of the precision of parameter estimates and classification output. Second, we evaluate how effective is BIC in detecting the best model in terms of the most suitable scale-level DIF specification. In order to provide a clear picture of the importance of detecting measurement non-invariance in latent Markov models for longitudinal data and give clear guidelines to applied users, we run an extensive simulation study considering the following target measures: absolute bias, relative mean squared error and standard errors of the structural parameter estimates. As the number of the parameters involved is quite large, we adopt a more compact notation. When we mention the parameters α and β, we are referring to all the parameters α_{1k} and β_{1k}, i.e. those at time $t = 1$. Similarly, for $t > 1$, we use the notation α_1 and β_1 for the parameters α_{tk} and β_{tkj} respectively.

The simulation design is structured as follows. In each scenario we generate $n = 500$ observations, observed in $T = 4$ time occasions on which $H = 10$ dichotomous items are measured. The observations are divided in $K = 3$ latent states whose separation depends on a "separation parameter", namely *sep*, that is set equal to 0.9, 0.8 and 0.7. This is used for calibrating intra class separation as $\gamma_{hk} = \log(sep/(1-sep))$, where γ_{hk} represents, as in equations (2),(3), (4) and (5), the intercept of the linear model linking the logit of the item specific probabilities with the covariates available. Trivially, the higher is *sep*, the higher is the intra class separation. Finally we generate for simplicity a single numerical covariate from a standard Gaussian distribution. The five DIF scenarios are described below.

1. No DIF. The covariate only affects transition probabilities but it does not affect item specific probabilities. This is labelled as DIF 0.

2. Full DIF. The covariate affects both the transition probabilities and the item specific probabilities. In particular, the η_{htk} parameters vary across items, time and class. This is labelled as DIF 1.

3. Time-Constant DIF. The η_{htk} parameters vary across items and class but remain fixed across time. This is labelled as DIF 2.

4. State-Constant DIF. The η_{htk} parameters vary across time and item but remain fixed across latent states. This is labeled as DIF 3.

5. State- and Time- constant DIF. The η_{htk} parameters are fixed across time and latent states. This is labelled as DIF 4.

By combining all the parameters listed above we obtain 15 different scenarios and generate the data 500 times in each scenario. The following competing approaches are compared in our simulation study.

1. Simultaneous estimator simply neglecting DIF: the covariate only affects initial and transition probabilities and has no direct effect on items (labelled as "No DIF").

2. Two-Step estimator (labelled as "Two–Step"): No DIF model estimated by following the two step procedure proposed in Di Mari & Bakk (2018)

3. The LM model selected by the BIC (labelled as "S.L.M.").

Figure 2 summarizes the results and reports parameters' absolute bias, relative mean squared errors and standard errors (computed based on the numeric Hessian). The selected

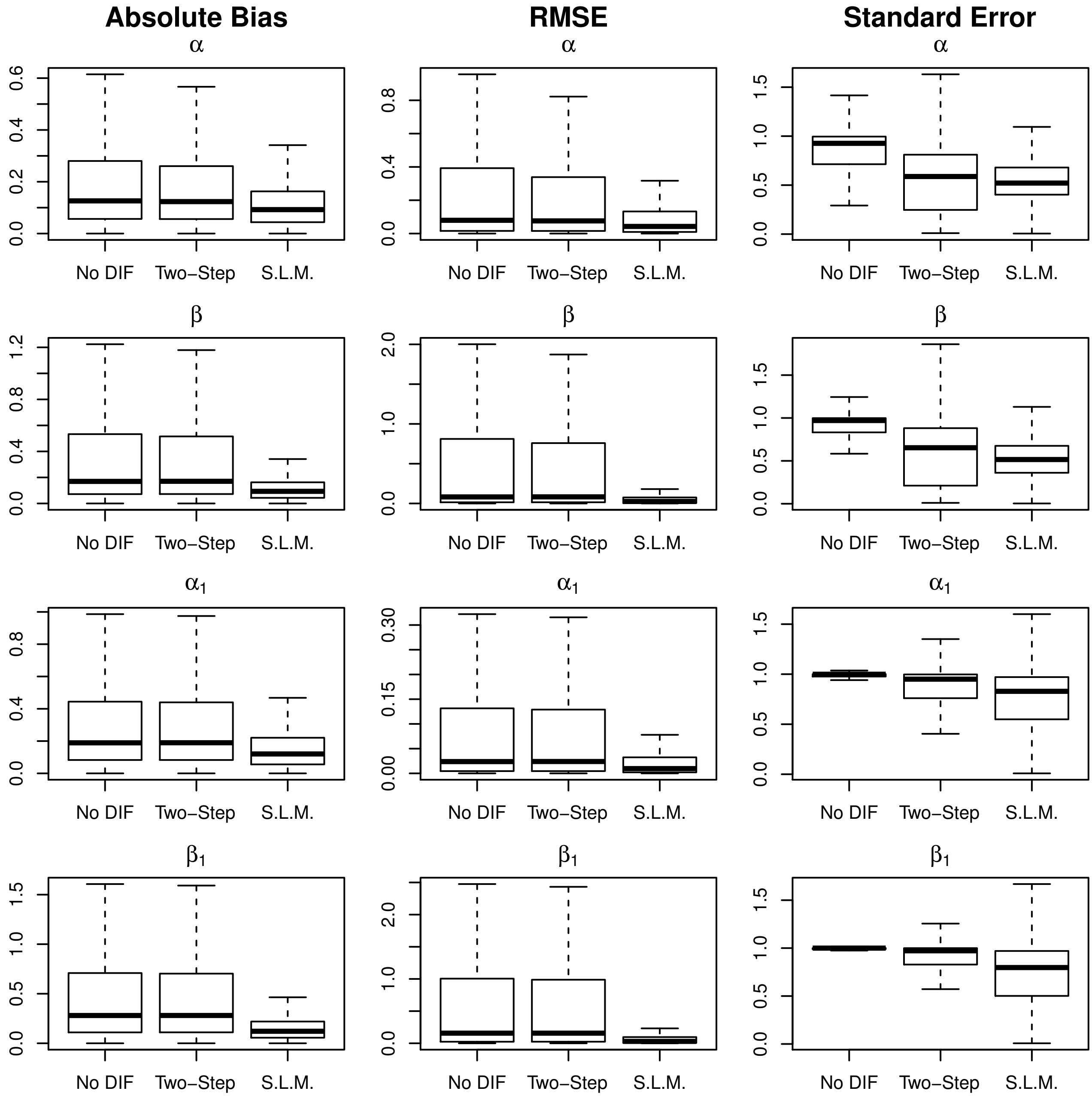

Figure 2. Absolute bias, Relative Mean Squared Error and Standard Error of the structural parameter estimates

model provides precise estimates of all the structural parameters while the two competing approaches provide severely biased parameter estimates in most of the scenarios. We also point out that, even if model selection may lead to use more complex model specifications, the associated standard errors are lower than the ones provided by possibly simpler models (see the third column of Figure 2). To show the impact of measurement non invariance on parameter estimates we report, in Figure 3, the value of the relative mean squared error of all the parameters as the type of DIF imposed varies. In absence of measurement invariance, as expected, all the procedures provide precise estimates, while, as differential item is imposed, standard approaches completely fail. On the contrary, the selected model provides estimates that are very accurate regardless of the type of underlying DIF. Even in the most complicate situation, the one in which the covariate affects both transition probabilities and item specific probabilities (Full DIF scenario- second panel of Figure 3) the parameters estimates show very low values in terms of relative mean squared error. Similarly as done with different type of DIF, we report, in Figure 4, the relative mean squared error of the parameter estimates as the intra class separation decreases. The selected model seems to be able to provide precise estimates regardless of the separation between the latent classes. On the contrary, even when the classes are well separated, the “No DIF” and the “Two-Step” models’ estimates look severely biased (and clearly the performances get worse and worse as the separation decreases). Finally, we observe that the usage of the BIC leads to select the proper model in most of the cases with probability very close to one - see the confusion matrix in Table 1. Furthermore, the full DIF case, probably the most problematic and least interpretable situation for applied researchers and practitioners, is detected with probability 1. Another interesting aspect regards state prediction. To measure the performances of in terms of state prediction - which is a classification task - we compute the Adjusted Rand Index (ARI, Hubert & Arabie, 1985) between the estimated latent class membership and the true classification labels. Table 2 reports the average ARI in each scenario for each of the procedures compared and confirms that in all of the proposed scenarios the selected model completely outperforms the other methods regardless of the type of DIF and the level of separation among the latent classes.

In conclusion, results from this extensive simulation study point out that the proposed model selection strategy is effective for choosing the scale level DIF specification, guaranteeing correct parameter estimates and model interpretability.

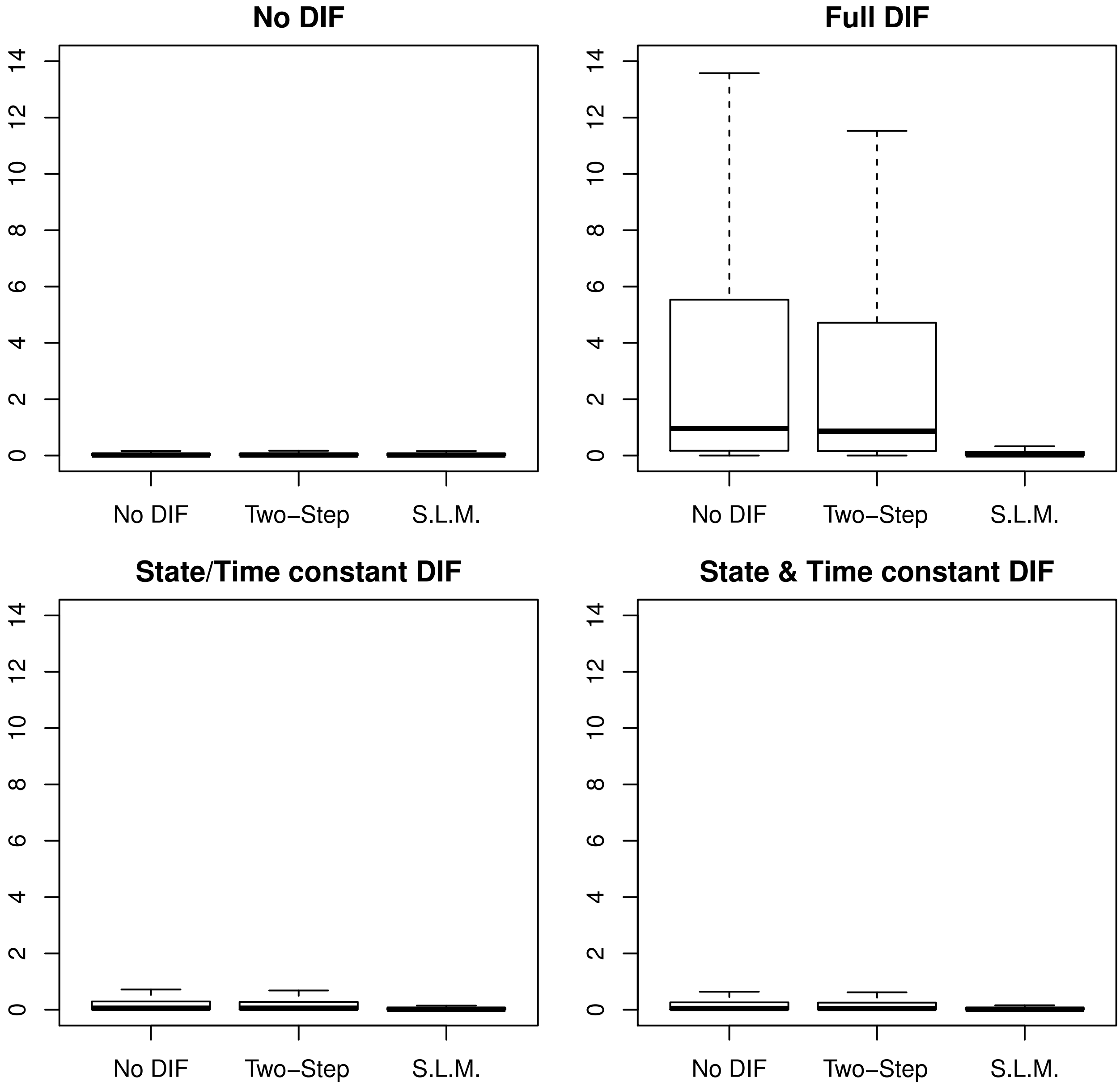

Figure 3. Relative Mean Squared Error as the DIF scenario varies

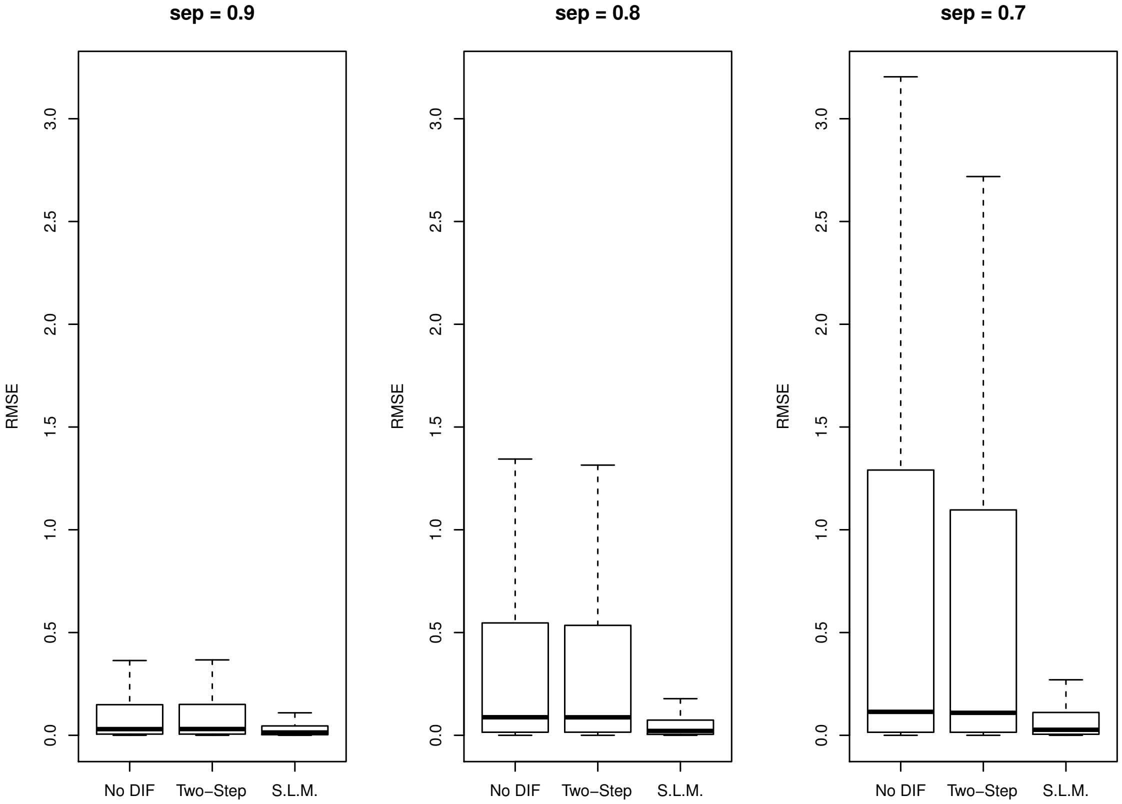

Figure 4. Relative Mean Squared Error as the intra-class separation varies

		True Model				
		No Dif	Full Dif	Time Constant Dif	State Constant Dif	State Time Constant Dif
BIC	No Dif	1.00	0.00	0.01	0.00	0.00
	Full DIF	0.00	1.00	0.00	0.00	0.00
	Time constant DIF	0.00	0.00	0.99	0.00	0.00
	State constant DIF	0.00	0.00	0.00	1.00	0.00
	State and Time constant DIF	0.00	0.00	0.00	0.00	1.00

Table 1
Confusion matrix normalized by column to evaluate the BIC performance.

Table 2

Average Adjusted Rand Index of the classification in each scenario

No DIF	Two-Step	S.L.M	Dif	Sep
0.96	0.95	0.96	0	0.9
0.43	0.43	0.94	1	0.9
0.96	0.95	0.98	2	0.9
0.86	0.86	0.98	3	0.9
0.92	0.92	0.96	4	0.9
0.92	0.92	0.92	0	0.8
0.18	0.18	0.73	1	0.8
0.79	0.79	0.93	2	0.8
0.61	0.61	0.93	3	0.8
0.75	0.75	0.93	4	0.8
0.87	0.86	0.88	0	0.7
0.24	0.24	0.68	1	0.7
0.66	0.68	0.90	2	0.7
0.49	0.50	0.85	3	0.7
0.62	0.63	0.86	4	0.7

5 Measuring tolerance towards minorities

The data are taken from the American General Social Survey (GSS), which is a nationally representative survey of the English-speaking, non-institutionalized adult population of the United States. We analyze a similar set of items as in McCutcheon (1985), who showed how age and education groups differ with respect to their tolerance toward minorities on data taken from the 1976 and 1977 GSS samples. The original set of items used by McCutcheon measured whether the respondents would allow members of different out-groups to speak in a public space. The items are formulated as follows: "Suppose this ...wanted to make a speech in your community. Should he be allowed to speak?" with response options "(Yes/No)" and are referred to communists, atheists, militarists, homosexuals, and racists. Newer versions of the survey include an additional item measuring tolerance towards Muslims. We used data from the panel version of the same survey, from the years 2010, 2012 and 2014 ($T = 3$), that are openly available[1]. We select sample units that were observed for all 3 waves, yielding a sample size $n = 697$.

We consider the extended set of items ($H = 6$), and focus on the effect of education and cohort separately on tolerance types. The same data set has been analyzed also in Di Mari & Bakk (2018). To make sure results are comparable with previous literature, we set the number of latent states to 3. Consistently with the rest of the paper, a time-invariant specification for the parameters of the transition probabilities is adopted.

As in McCutcheon (1985), we re-coded Education into three categories - less than grade 12 (1), completed high school (2) and higher educated (3). Cohort instead was coded into four categories: young (18-24 years old, coded as 1), young adults (25-42 years old, coded as 2), middle age (43-61 years old, coded as 3) and old (62 years and older, coded as 4). To avoid parameter proliferation, both covariates are taken as numeric in all model

[1] `https://gss.norc.org/`

equations. Following the setup of the simulation study, we fit DIF 0 to DIF 4 LM models, and the no DIF (DIF 0) model with the two–step estimator.

5.1 Education

Table 3 presents overall model statistics for each DIF type (simultaneous estimator). The lowest BIC is attained at time- and state- constant DIF (DIF 4). Interestingly, the direct effect of Education on items (Table 4) shows that a higher education corresponds on average to a higher probability to allow “Atheists”, “Communists”, “Homosexuals”, “Militarists”, and “Muslims” to speak in public. The effect of Education on the fifth item (allow “Racists” to speak in public) is not significant. For this reason, in what follows we will consider also the DIF 4 model with no DIF on item 5 - “DIF 4 (-)” henceforth. This configuration turns out to have the lowest BIC (8935.36, with 39 parameters and an entropy R^2 of 0.79).

	BIC	# par	entr. R^2
No DIF (DIF 0)	8973.67	34	0.81
Full DIF (DIF 1)	9198.68	88	0.80
Time-constant DIF (DIF 2)	8997.86	52	0.84
State-constant DIF (DIF 3)	9008.30	52	0.83
Time- State- constant DIF (DIF 4)	**8941.46**	40	0.79

Table 3

BIC, number of parameters and entropy R^2 for the simultaneous estimator of DIF 0 - DIF 4 models. DIF with respect to education.

	time- state- constant DIF (DIF 4)
$\eta_{1\cdot\cdot}$	0.76^{***}
(“Atheists”)	(0.17)
$\eta_{2\cdot\cdot}$	1.12^{***}
(“Communists”)	(0.20)
$\eta_{3\cdot\cdot}$	0.85^{***}
(“Homosexuals”)	(0.15)
$\eta_{4\cdot\cdot}$	0.39^{***}
(“Militarists”)	(0.14)
$\eta_{5\cdot\cdot}$	0.09
(“Racists”)	(0.14)
$\eta_{6\cdot\cdot}$	1.16^{***}
(“Muslims”)	(0.23)

Table 4

*Estimates of the direct effect of the covariate “Education” on the items according to the time- state- constant DIF (DIF 4) - the best according to BIC. Standard errors in parentheses (based on the observed Information matrix). *** p-value<0.01, ** p-value<0.05, * p-value<0.1*

The item–class (tolerance) profiles averaged across covariate levels from the DIF 4

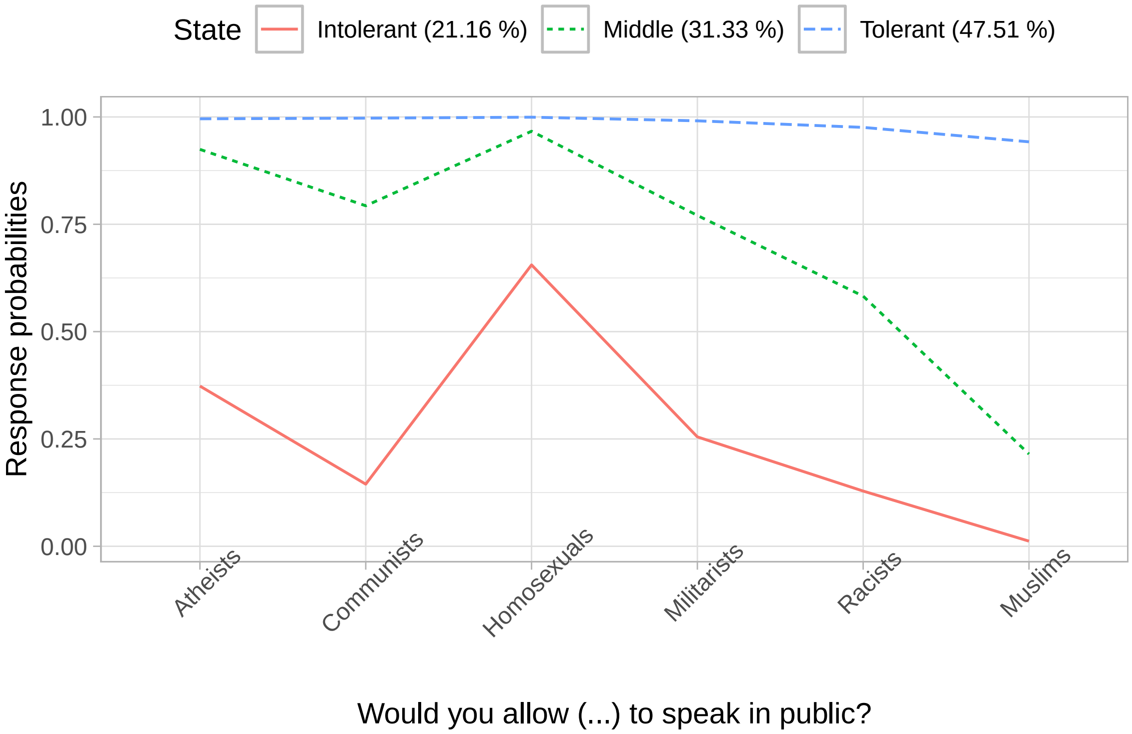

Figure 5. Response probabilities to answer "yes" given state membership. The covariate Education is loaded on the state variable and has a time and state constant direct effect on all indicators.

model are displayed in Figure 5. Overall, the tolerance type profiles are quite separated with each other, naturally giving the latent states an ordinal interpretation. The "Tolerant" types have a high probability of scoring "yes" on all items. For the "Middle" and the "Intolerant" types, the probability of scoring "yes" on item 3 (allow "Homosexual" to speak in public) is relatively higher than for other items, whereas the item with the lowest probability of scoring "yes" is the sixth (allow "Muslims" to speak in public). More in general, the "Intolerant" types have the lowest probabilities of scoring "yes" on all items compared to the other tolerance types. "Middle" has an overall profile that is in between the "Intolerant" and the "Tolerant" ones.

In Table 5 we display the overall transition probability matrix - averaged over sample units and time. The "Intolerant" state has, on average, the highest persistence. We note that "Intolerant" units have, on average, a probability of 0.25 to move to the "Middle" state - with more tolerant positions towards minorities. In general transitions are more likely to occur towards contiguous states.

Results for the structural model parameters, relating Education to state membership, are reported in Table 6. We report, for comparison, results for the simultaneous estimator for all 5 types of DIF and for the DIF 4 (time- state- constant DIF) with no DIF for item 5, and for the two–step estimator for the no DIF.

The signs of significant coefficients are in line, although we observe that the magni-

	State ($t = 1$):		
	“Tolerant”	“Middle”	“Intolerant”
	0.48	0.31	0.21
	State ($t - 1$):		
	“Tolerant”	“Middle”	“Intolerant”
State (t):			
“Tolerant”	0.90	0.17	0.08
“Middle”	0.08	0.72	0.25
“Intolerant”	0.02	0.11	0.67

Table 5
Overall initial (on the top) and transition (on the bottom) probabilities, conditional on Education, obtained averaging over sample units and time, for the simultaneous estimator of the time- state- constant DIF (DIF 4) model. For the transition probabilities, past is on the columns, present is on the rows.

tudes of the estimates for the fit of DIF 4 model are somewhat smaller in absolute value than for the other DIF types. With higher level of education corresponds, on average, a higher probability of starting in the “Tolerant” state. Those that are in state “Tolerant” have lower probability to transition to the other states. Interestingly, higher levels of education in general correspond to a lower probability to be/ to move to less tolerant states. In the appendix we report also the average initial and transition probabilities for each Education level.

	DIF 0	DIF 1	DIF 2	DIF 3	DIF 4	DIF 0 (2-step)	DIF 4 (-)
				Initial state			
$\alpha_{1,2}$	1.74***	1.43***	1.57***	1.25**	1.31***	0.85*	1.35***
	(0.53)	(0.51)	(0.51)	(0.48)	(0.51)	(0.45)	(0.51)
$\alpha_{1,3}$	3.11***	1.71***	1.85***	1.51***	1.70***	1.24***	1.81***
	(0.47)	(0.50)	(0.51)	(0.52)	(0.53)	(0.43)	(0.48)
$\beta_{1,2}$	-0.79***	-0.72***	-0.73***	-0.64***	-0.65***	-0.97***	-0.66***
	(0.19)	(0.19)	(0.19)	(0.18)	(0.19)	(0.19)	(0.19)
$\beta_{1,3}$	-1.50***	-0.97***	-0.95***	-0.91***	-0.89***	-1.31***	-0.95***
	(0.18)	(0.19)	(0.19)	(0.20)	(0.20)	(0.19)	(0.18)
				Transitions			
α_{21}	0.20	-0.62	-0.95	-0.54	-0.51	-4.72	-0.48
	(0.77)	(0.77)	(0.94)	(0.77)	(0.82)	(4.87)	(0.82)
α_{31}	-0.91	-0.65	-0.72	-0.44	-0.45	0.51	-0.52
	(1.71)	(1.18)	(1.18)	(1.05)	(1.08)	(0.53)	(1.10)
α_{12}	-0.55	-0.83	-1.02	-0.58	-0.66	-0.03	-0.69
	(0.70)	(0.74)	(0.68)	(0.73)	(0.71)	(0.69)	(0.71)
α_{32}	0.83	0.23	0.42	0.10	0.41	-1.19	0.48
	(0.65)	(0.62)	(0.62)	(0.70)	(0.70)	(1.47)	(0.68)
α_{13}	-4.62***	-3.37***	-3.29***	-3.09***	-3.05***	-1.79**	-3.14***
	(1.69)	(1.22)	(1.19)	(1.03)	(1.03)	(0.71)	(1.04)
α_{23}	-0.97**	-0.17	-0.26	-0.29	-0.36	-2.57**	-0.39
	(0.46)	(0.52)	(0.50)	(0.55)	(0.52)	(1.14)	(0.51)
β_{21}	-0.96***	-0.58**	-0.57	-0.68**	-0.76**	-0.37	-0.76**
	(0.29)	(0.29)	(0.35)	(0.30)	(0.32)	(1.79)	(0.32)
β_{31}	-1.35**	-1.28**	-1.26**	-1.33***	-1.36***	-1.68***	-1.34***
	(0.66)	(0.51)	(0.50)	(0.46)	(0.48)	(0.26)	(0.48)
β_{12}	-0.31	-0.40	-0.18	-0.46	-0.32	-0.41	-0.31
	(0.27)	(0.33)	(0.27)	(0.31)	(0.28)	(0.31)	(0.28)
β_{32}	-1.12***	-0.80***	-0.98***	-0.74**	-0.97***	-0.31	-1.00***
	(0.29)	(0.29)	(0.29)	(0.31)	(0.32)	(0.62)	(0.31)
β_{13}	0.96	0.50	0.48	0.43	0.40	0.56*	0.43
	(0.61)	(0.45)	(0.44)	(0.39)	(0.39)	(0.33)	(0.39)
β_{23}	0.04	-0.42*	-0.31	-0.38	-0.28	0.83*	-0.27
	(0.21)	(0.24)	(0.22)	(0.25)	(0.23)	(0.50)	(0.22)

Table 6

*Structural model parameter estimates for no DIF (DIF 0), full DIF (DIF 1), time-constant DIF (DIF 2), state-constant DIF (DIF 3), time- state- constant DIF (DIF 4), time- state-constant DIF (DIF 4 (-)) with one no DIF item ("Racists") and models with 3 latent states and Education as covariate. The numbering of the states is 1="Tolerant", 2="Middle", and 3="Intolerant". The first state ("Tolerant") is taken as reference for the initial state multinomial regression, whereas the no-change state is taken as reference for the transition probability multinomial regressions. Standard errors in parentheses (based on the inverse hessian). *** p-value<0.01, ** p-value<0.05, * p-value<0.1*

5.2 Cohort

Table 7 presents overall model statistics for each DIF type (simultaneous estimator). The model with no DIF has the lowest BIC.

	BIC	# par	entr. R^2
No DIF (DIF 0)	**9085.79**	34	0.81
Full DIF (DIF 1)	9110.50	88	0.83
Time-constant DIF (DIF 2)	9162.68	52	0.83
State-constant DIF (DIF 3)	9319.95	52	0.82
Time- State- constant DIF (DIF 4)	9093.29	40	0.81

Table 7

BIC, number of parameters and entropy R^2 for DIF 0 - DIF 4 models. DIF with respect to cohort.

The item–class (tolerance) profiles are displayed in Figure 6. Overall, we observe similar profiles as above: the tolerance type profiles are quite separated with each other. Interestingly, the overall size of the "Tolerant" state is slightly larger compared to what reported in the previous subsection - where Education has a direct effect on the items.

In Table 8 we display the overall transition probability matrix - averaged over sample units and time. Also here, the "Intolerant" state has, on average, the highest persistence.

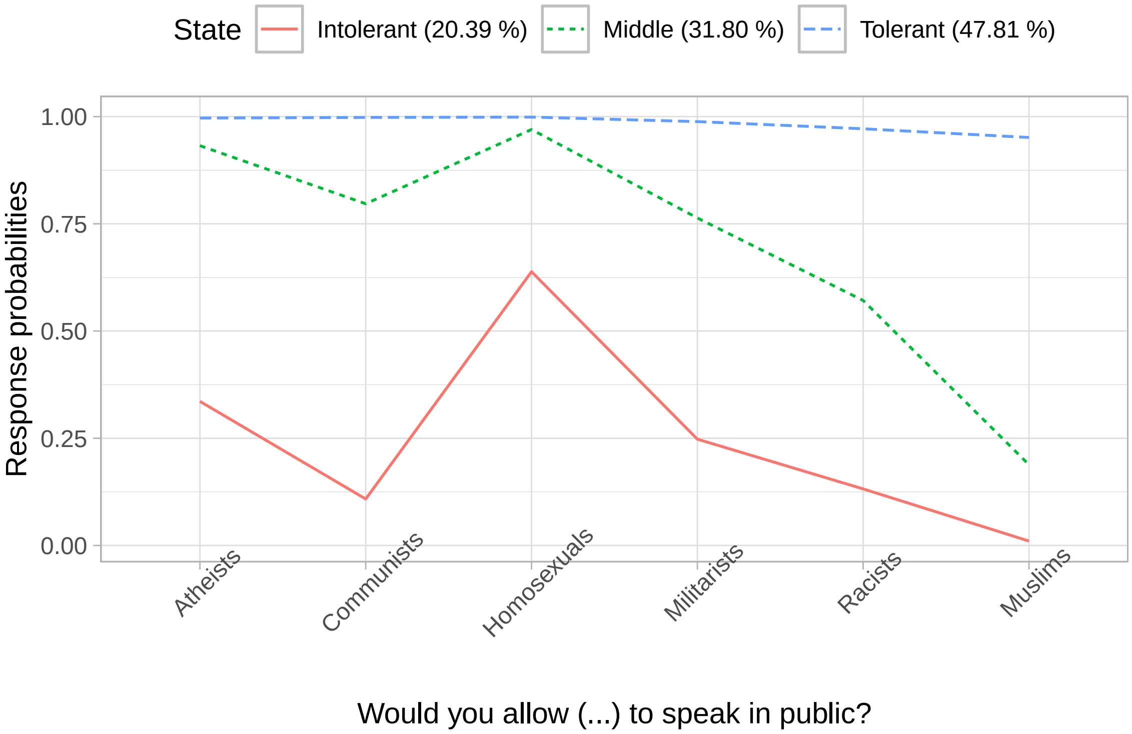

Figure 6. Response probabilities to answer "yes" given state membership. The covariate Cohort is loaded on the state variable but has no direct effect on the indicators.

We note that "Intolerant" units have, on average, a probability of 0.27 to move to the "Middle" state - with more tolerant positions towards minorities - which is 2 percentage points higher than in Table 8.

	State ($t = 1$):		
	"Tolerant"	"Middle"	"Intolerant"
	0.48	0.32	0.20
	State ($t - 1$):		
	"Tolerant"	"Middle"	"Intolerant"
State(t):			
"Tolerant"	0.91	0.17	0.06
"Middle"	0.08	0.71	0.27
"Intolerant"	0.01	0.12	0.67

Table 8

Overall initial (on the top) and transition (on the bottom) probabilities, conditional on Cohort, obtained averaging over sample units and time, for the simultaneous estimator of the time- state- constant DIF (DIF 4) model. For the transition probabilities, past is on the columns, present is on the rows.

Results for the structural model parameters are displayed in Table 9. We note that, the larger the cohort the lower on average the probability of moving from "Intolerant" to "Tolerant" - this is the only effect that is significant in all DIF specifications and with both simultaneous and two-step estimators.

In the appendix we report also the average initial and transition probabilities for each Education level.

	DIF 0	DIF 1	DIF 2	DIF 3	DIF 4	DIF 0 (2-step)
			Initial state			
α_{12}	-0.47	-1.11***	-1.14***	-0.78**	-0.77**	-0.67*
	(0.36)	(0.40)	(0.40)	(0.36)	(0.37)	(0.34)
α_{13}	-1.53***	-1.31***	-1.33***	-1.75***	-1.77***	-2.20***
	(0.39)	(0.41)	(0.40)	(0.45)	(0.44)	(0.46)
β_{12}	0.04	0.23	0.29**	0.13	0.15	0.05
	(0.13)	(0.14)	(0.14)	(0.13)	(0.13)	(0.12)
β_{13}	0.32**	0.22	0.27**	0.36**	0.40***	0.34**
	(0.13)	(0.14)	(0.14)	(0.16)	(0.15)	(0.15)
			Transitions			
α_{21}	-1.31*	-2.00***	-2.10**	-1.49**	-1.32*	-1.89***
	(0.76)	(0.74)	(0.86)	(0.68)	(0.71)	(0.68)
α_{31}	-3.61**	-2.20*	-2.35*	-3.53**	-3.59**	-4.04***
	(1.48)	(1.27)	(1.36)	(1.42)	(1.44)	(1.37)
α_{12}	-1.39**	-1.02	-1.04	-1.05*	-1.01*	-1.29**
	(0.61)	(0.77)	(0.72)	(0.63)	(0.61)	(0.53)
α_{32}	-1.55**	-0.67**	-0.47	-1.31**	-1.26**	-2.56***
	(0.64)	(0.74)	(0.72)	(0.64)	(0.64)	(0.73)
α_{13}	0.88	1.73	1.63	1.80	1.63	2.17
	(1.25)	(1.09)	(1.04)	(1.41)	(1.36)	(1.67)
α_{23}	-0.08	-0.20	-0.22	-0.09	-0.05	0.22
	(0.66)	(0.64)	(0.62)	(0.70)	(0.69)	(0.91)
β_{21}	-0.45	-0.14	-0.17	-0.37	-0.47	-0.27
	(0.31)	(0.27)	(0.32)	(0.29)	(0.30)	(0.27)
β_{31}	-0.29	-0.78	-0.74	-0.31	-0.31	-0.14
	(0.56)	(0.54)	(0.57)	(0.53)	(0.55)	(0.51)
β_{12}	-0.01	-0.19	-0.11	-0.16	-0.14	-0.04
	(0.22)	(0.26)	(0.24)	(0.23)	(0.22)	(0.19)
β_{32}	-0.08	-0.38	-0.46*	-0.16	-0.19	-0.27
	(0.23)	(0.26)	(0.26)	(0.24)	(0.23)	(0.25)
β_{13}	-1.28**	-1.67***	-1.62***	-1.71**	-1.61**	-1.61**
	(0.57)	(0.52)	(0.49)	(0.67)	(0.64)	(0.76)
β_{23}	-0.28	-0.25	-0.24	-0.30	-0.29	-0.29
	(0.22)	(0.21)	(0.21)	(0.23)	(0.22)	(0.29)

Table 9

Structural model parameter estimates for no DIF (DIF 0), full DIF (DIF 1), time-constant DIF (DIF 2), state-constant DIF (DIF 3), time- state- constant DIF (DIF 4), time- state-constant DIF (DIF 4 (-)) with one no DIF item ("Racists") and models with 3 latent states and Cohort as covariate. The first state ("Tolerant") is taken as reference for the initial state multinomial regression, whereas the no-change state is taken as reference for the transition probability multinomial regressions. Standard errors in parentheses (based on the inverse hessian). *** *p-value<0.01*, ** *p-value<0.05*, * *p-value<0.1*

6 Conclusion

Model selection at scale level is crucial for correct assessment of differential item functioning and accurate parameter estimation.

In the simulation study we have investigated the impact of various types of differential item functioning. Results indicate that, if selection of the appropriate DIF structure at scale level is correctly carried out, parameter estimates are quite precise under very wide conditions and classification output is very precise. We report also that BIC was able to select the correct DIF structure more than 95% of times.

In the empirical application, we have specified a time- and state- constant DIF for a LM model with education as covariate, and a no DIF when using cohort as predictor of class membership. The selection was based on the proposed procedure. In the first case (education as predictor), the time- and state- constant DIF specification has a number of extra parameters compared to the no DIF model. At item level, the direct effect of Education on the probability of scoring a “Yes” on one item (allow “Racists” to speak in public) was not significant. Regarding the structural model, although signs of the coefficients were in line with each other, we have observed that estimates of the unselected models were slightly greater in magnitude (in absolute value) than those of the selected model. Concerning the second case (cohort as predictor), we found that older cohorts are less likely, on average, to transition from the intolerant state to the tolerant state.

We have considered the effect of education and cohort separately on tolerance types. However, future research might explore education and cohort effects (separately and interacted) jointly.

Model specifications where the covariate(s) are allowed to have direct effects on the indicators might be difficult to interpret. However, as our results have pointed out, the most profitable strategy to get accurate parameter estimates seems the one taking into account possible measurement non-invariance, rather than simply leaving DIF under the hood.

Although our selection approach has been based on the simultaneous estimator only, its usefulness goes beyond that. Indeed our specifications and results offer an opportunity to amend also stepwise estimators (Di Mari & Bakk, 2018; related to this, see the discussion in Vermunt & Magidson, 2020).

As a general recommendation for applied users, whenever a DIF specification with less parameters than the full DIF is selected, standard parameter testing can be used to further simplify the model. However, note that selecting the DIF specification only based on directional tests can be misleading. This is partly due to the fact that standard errors might be inflated in overparameterized models. On this point, see also Battauz (2019).

Selection of the number of states, when model assumptions hold as well as when they do not, is still an unresolved issue in model-based clustering in general. We have adopted guidelines that give some guarantees (see, for instance, Bacci et al., 2014; Pohle et al., 2017). More in general, the choice of the number of latent states can be done by combining information criteria, researchers’ specifications and desires, and *a priori* information that may be available. Investigating the properties of information criteria under different (realistic) scenarios is however an important topic that deserves further research.

7 References

Altman, R. M. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, *102*(477), 201–210.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2017). Dynamic latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 257–269.

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling*, *21*, 329–341.

Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent markov model for longitudinal data. *Advances in Data Analysis and Classification*, *8*(2), 125–145.

Bakk, Z., & Kuha, J. (2020). Relating latent class membership to external variables: An overview. *British Journal of Mathematical and Statistical Psychology.*

Bartolucci, F., & Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, *104*, 816–831.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent Markov models for longitudinal data.* Chapman and Hall / CRC Press.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates (with discussion). *TEST*, *23*, 433-486.

Bartolucci, F., Pennoni, F., & Vittadini, G. (2011). Assessment of school performance through a multilevel latent markov rasch model. *Journal of Educational and Behavioral Statistics*, *36*(4), 491–522.

Battauz, M. (2019). On Wald tests for differential item functioning detection. *Statistical Methods & Applications*, *28*, 103-118.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2018). Testing for approximate measurement invariance of human values in the european social survey. *Sociological Methods & Research*, *47*(4), 665–686.

Cohen, J. (1992). Things i have learned (so far). In *Annual convention of the american psychological association, 98th, aug, 1990, boston, ma, us; presented at the aforementioned conference.*

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). Wiley.

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

Davidov, E., Muthen, B., & Schmidt, P. (2018). *Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones.* SAGE Publications Sage CA: Los Angeles, CA.

De Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis*, *10*, 157-169.

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

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

Di Mari, R., Oberski, D. L., & Vermunt, J. K. (2016). Bias-adjusted three-step latent Markov modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 649–660.

Dotto, F., Farcomeni, A., Pittau, M. G., & Zelli, R. (2019). A dynamic inhomogeneous latent state model for measuring material deprivation. *Journal of the Royal Statistical Society: Series A*, *182*, 495-516.

Farcomeni, A. (2015). Generalized linear mixed models based on latent Markov heterogeneity structures. *Scandinavian Journal of Statistics*, *42*, 1127-1135.

Farcomeni, A., Ranalli, M., & Viviani, S. (2020). Dimension reduction for longitudinal multivariate data by optimizing class separation of projected latent Markov models. *TEST*, available online.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology*, 79–259.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Ip, E. H., Snow Jones, A., Heckert, D. A., Zhang, Q., & Gondolf, E. D. (2010). Latent Markov model for analyzing temporal configuration for violence profiles and trajectories in a sample of batterers. *Sociological Methods & Research*, *39*(2), 222–255.

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., & Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*, 193–209.

Janssen, J. H., Van Laar, S., De Rooij, M. J., Kuha, J., & Bakk, Z. (2019). The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(2), 280–290.

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for measurement invariance with latent class analysis. *Cross-cultural analysis: Methods and applications*, 359–384.

Kankaraš, 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.

Kim, M., Vermunt, J. K., Bakk, Z., Jaki, T., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 601–614.

Lorah, J., & Womack, A. (2019). Value of sample size for computation of the bayesian information criterion (bic) in multilevel modeling. *Behavior Research Methods*, *51*(1), 440–450.

Magidson, J., & Vermunt, J. (2004). Latent variable models. *???*, *?*, ??

Maruotti, A. (2011). Mixed hidden Markov models for longitudinal data: an overview. *International Statistical Review*, *79*(3), 427–454.

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

McCoach, D., & Black, A. (2008). Evaluation of model fit and adequacy. In A. O'Connel & D. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age.

McCutcheon, A. L. (1985). A latent class analysis of tolerance for nonconformity in the american public. *Public Opinion Quarterly*, *49*(4), 474–488.

Montanari, G. E., Doretti, M., & Bartolucci, F. (2018). A multilevel latent markov model for the evaluation of nursing homes' performance. *Biometrical Journal*, *60*(5), 962–978.

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth across europe in 1999 and 2009: The alignment method applied to iea cived and iccs. *Sociological Methods & Research*, *47*(4), 687–728.

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, *47*(4), 637–664.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, *46*, 252–282.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2017). Power and type I error of local fit statistics in multilevel latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 216–229.

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 782–797.

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.

Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A monte carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*, 267–279.

Oberski, D. L., & Vermunt, J. K. (2018). The expected parameter change (EPC) for local dependence assessment in binary data latent class models. *arXiv preprint arXiv:1801.02400*.

Paas, L. J., Vermunt, J. K., & Bijmolt, T. H. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(4), 955–974.

Pennoni, F., & Genge, E. (2020). Analysing the course of public trust via hidden Markov models: a focus on the Polish society. *Statistical Methods and Applications*, *29*, 399-425.

Pohle, J., Langrock, R., van Beest, F. M., & Schmidt, N. M. (2017). Selecting the number of states in hidden markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(3), 270–293.

Poulsen, C. S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, *7*, 5–19.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Van de Pol, F., & De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research*, *15*(1-2), 118–141.

Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, 213–247.

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

Vermunt, J. K., Langeheine, R., & B ockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*, 179–207.

Vermunt, J. K., & Magidson, J. (2016). Technical guide for latent gold 5.1: Basic, advanced, and syntax. *Belmont, MA: Statistical Innovations Inc.*.

Vermunt, J. K., & Magidson, J. (2020). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–9.

Vogelsmeier, L. V., Vermunt, J. K., Keijsers, L., & De Roover, K. (2020). Latent markov latent trait analysis for exploring measurement model changes in intensive longitudinal data. *Evaluation & the Health Professions.*

Vogelsmeier, L. V., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 557–575.

Vriens, I., Moors, G., Gelissen, J., & Vermunt, J. K. (2017). Controlling for response order effects in ranking items using latent choice factor modeling. *Sociological Methods & Research*, *46*(2), 218–241.

Weakliem, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods & Research*, *33*(2), 167–187.

Wiggins, L. M. (1973). *Panel analysis: latent probability models for attitude and behaviour processes.* Elsevier, Amsterdam.

Yang, Y. (2005). Can the strenghts of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, *92*, 937-950.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R.* Chapman and Hall / CRC Press.

Appendix

Additional results for the real-data application

In Table A1 we report the predicted initial and transition probabilities for Education levels, for Type-4 model. In Table A2 we report the predicted initial and transition probabilities for Cohort levels, for the measurement invariant model.

7.1 Cohort

		State ($t = 1$):	
	“Tolerant”	“Middle”	“Intolerant”
Education =			
1	0.20	0.38	0.42
2	0.34	0.35	0.31
3	0.53	0.28	0.19
Education = 1			
		State (t):	
	“Tolerant”	“Middle”	“Intolerant”
State($t - 1$):			
“Tolerant”	0.69	0.20	0.11
“Middle”	0.19	0.52 0.29	
“Intolerant”	0.04	0.33	0.63
Education = 2			
		State (t):	
	“Tolerant”	“Middle”	“Intolerant”
State($t - 1$):			
“Tolerant”	0.85	0.11	0.04
“Middle”	0.18	0.67	0.15
“Intolerant”	0.07	0.26	0.67
Education = 3			
		State (t):	
	“Tolerant”	“Middle”	“Intolerant”
State($t - 1$):			
“Tolerant”	0.93	0.06	0.01
“Middle”	0.15	0.78	0.07
“Intolerant”	0.11	0.20	0.69

Table A1

Predicted initial and transition probabilities for Education levels for Type-4 model.

		State ($t = 1$):	
	“Tolerant”	“Middle”	“Intolerant”
Cohort =			
1	0.51	0.33	0.15
2	0,48	0.32	0.20
3	0.44	0.31	0.25
4	0.40	0.29	0.31
Cohort = 1			
		State (t):	
	“Tolerant”	“Middle”	“Intolerant”
State($t-1$):			
“Tolerant”	0.84	0.14	0.02
“Middle”	0.17	0.69	0,14
“Intolerant”	0.28	0.30	0.42
Cohort = 2			
		State (t):	
	“Tolerant”	“Middle”	“Intolerant”
State($t-1$):			
“Tolerant”	0.89	0.10	0.01
“Middle”	0.17	0.70	0.13
“Intolerant”	0.11	0.31	0.58
Cohort = 3			
		State (t):	
	“Tolerant”	“Middle”	“Intolerant”
State($t-1$):			
“Tolerant”	0.93	0.06	0.01
“Middle”	0.17	0.71	0.12
“Intolerant”	0.04	0.27	0.69
Cohort = 4			
		State (t):	
	“Tolerant”	“Middle”	“Intolerant”
State($t-1$):			
“Tolerant”	0.95	0.04	0.01
“Middle”	0.17	0.72	0.11
“Intolerant”	0.01	0.23	0.76

Table A2

Predicted initial and transition probabilities for Cohort levels for the no DIF model.