Approximate measurement invariance and longitudinal confirmatory factor analysis: concept and application with panel data

Daniel Seddig
Institute of Sociology and Social Psychology
University of Cologne, Germany

and

Department of Psychology
University of Zurich, Switzerland

Heinz Leitgöb
Institute of Sociology
University of Eichstätt-Ingolstadt
Germany

This article addresses the approximate approach to assess measurement invariance with (longitudinal) confirmatory factor analysis. Approximate measurement invariance uses zero-mean, small-variance Bayesian priors to allow minor differences in estimated parameters across time, while still maintaining comparability of the underlying constructs. The procedure is illustrated for the first time with panel data on young peoples’ preferences to maximize pleasure and enjoy life. Results indicate whereas the traditional approach of exact measurement invariance failed to establish scalar invariance across time and precluded comparisons of latent means, it was possible to establish approximate scalar invariance. Based on a monitoring procedure for model fit and convergence, a rather small prior variance was deemed sufficient to account for minor deviations of cross-time intercept differences from zero.

Keywords: Confirmatory factor analysis, Bayesian structural equation modeling, approximate measurement invariance, panel data

1 Introduction

Comparative research is a core subject in different fields of social science. Research interests can be cross-national and cross-cultural or longitudinal and developmental. For example, it may be of interest to compare mean scores of variables, such as ethics, values, attitudes, and social or political trust across cultures, countries, or across time (Davidov, Schmidt, & Billiet, 2011). Regardless of whether the comparison is focused on different contexts or time points, the mere application of the same scale across groups or time does not ensure that the same concept is actually analyzed. Response characteristics to survey questions may be different across groups or over time which implies that people from different contexts may not understand the concept in the same way or that the understanding has changed. Thus, if the studied variables are latent variables measured by multiple indicators, a comparison across groups or time requires measurement invariance (hereafter: MI) to ensure that the concept remains comparable across groups or time. However, methodology reports of relevant large-scale panel studies, such as the Socioeconomic Panel (SOEP) in Germany, British Household Panel Survey (BHPS) in Great Britain, National Crime Victimization Survey (NCVS) in the United States, and European Union Statistics on Income and Living Conditions (EU-SILC) usually do not include systematic assessments of MI for key concepts across time. It seems that testing for MI usually remains within the area of responsibility of the individual data user. For this reason, the article aims at capacitating researchers to apply the innovative concept of Bayesian approximate MI for situations where conventional approaches turn out to be inappropriate.

Since traditional tests of MI build on very restrictive assumptions about the equivalence of measurement properties across groups or time, it cannot always be achieved to a satisfactory degree. With the introduction of Bayesian analytic properties into confirmatory factor analysis (hereafter: CFA) and structural equation modeling (hereafter: SEM), researchers are able to relax exact equality constraints by assuming that parameters are only approximately equal, but comparability of the underlying constructs is maintained. For this purpose, cross-group or cross-time parameter differences are expressed as Bayesian priors which are assumed to be shaped as a normal distribution with a mean of zero and a small variance (B. Muthén & Asparouhov, 2013; Van de Schoot, Schmidt, De Beuckelaer, Lek, and Zondervan-Zwijnenburg (2015) for a recent overview of recent developments in the analysis of MI.
Schoot et al., 2013).\(^2\)

In the following we will emphasize the properties of Bayesian approximate MI as an alternative to exact MI and demonstrate how it can be assessed within longitudinal CFA. Further, we will demonstrate the use of Bayesian approximate MI in an empirical application with panel data. So far, to the best of our knowledge, Bayesian approximate MI has not been applied to longitudinal data. The variables of interest are young peoples’ preferences to maximize pleasure and enjoy life which are conceptualized as a latent variable called “hedonism”. We will conclude with a summary of our findings and will discuss the implications for future comparative research.

2 Measurement Invariance

The indication of MI is “whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute” (Horn & McArdle, 1992, p. 117). If MI does not hold, assuming differences in latent variable means or regression coefficients can be incorrect or actual differences may be concealed. For longitudinal modeling, MI guarantees that the same latent variable is measured at all times and avoids interpreting change in the meaning or understanding of a latent variable as change in terms of an underlying developmental process (Ferrer, Balluerka, & Widaman, 2008; Little, 2013; Stoel, van den Wittenboer, & Hox, 2004; Widaman, Ferrer, & Conger, 2010).

To illustrate the parameters that can be subject to MI constraints, recall a confirmatory factor model (Bollen, 1989; Brown, 2015; Jöreskog, 1979; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000):

\[
y’ = \tau’ + \Lambda’ \eta’ + \epsilon’
\]

(1)

where \(y’\) is a \((M\times1)\) vector of \(M\) observed indicators (for time \(t\) consecutively), \(\eta’\) is a \((N\times1)\) vector of \(N\) latent unobserved variables, \(\tau’\) is a \((M\times1)\) vector of indicator intercepts, \(\Lambda’\) is a \((M\timesN)\) matrix of factor loadings linking observed indicators and unobserved latent variables, and \(\epsilon’\) is a \((M\times1)\) vector of measurement errors with \(E(\epsilon’) = 0\; \text{and} \; \text{COV}(\eta’ , \epsilon’) = 0.\)

MI is typically assessed by imposing different equality restrictions on the parameters across time (Hertzog & Nesselroade, 2003; Horn & McArdle, 1992; Meredith, 1993; Meredith & Teresi, 2006; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). Configural MI assumes structural equivalence. This means that only the patterns of factors and factor loadings are constrained to be equal across time. Metric MI requires the factor loadings to be equal across time (factor loading invariance), indicating that the meaning of a latent variable is the same across time. Metric MI is considered necessary to compare factor covariances or unstandardized regression coefficients. Scalar MI additionally requires indicator intercepts to be equal across time (intercept invariance), indicating the similar use or origin of a scale across time. Scalar MI is considered necessary to compare latent means.\(^4\)

3 Approximate Measurement Invariance

Factor loadings and intercepts are usually constrained to be exactly equal across time. When exact MI does not hold, one can reduce the number of time points that are compared. However, this may be rather unsatisfactory. Another alternative is partial MI (Byrne, Shavelson, & Muthén, 1989) where non-invariant parameters are freely estimated and the remaining are held exactly equal. However, it has been shown that approximate MI provides a better alternative when relatively small parameter differences occur (Van de Schoot et al., 2013).

Bayesian approximate MI is based on less strict assumptions about parameter differences, i.e. parameter differences are allowed some leeway or “wiggle room” (Van de Schoot et al., 2013, p. 1). Differences are assumed to be almost zero, but not exactly. However, differences are still kept at a minimum to ensure that concepts remain approximately comparable. The degree of leeway can be expressed in terms of a prior distribution, usually a normal distribution with a mean of zero and a small variance, i.e. \(N(0, \nu)\). The prior variance \(\nu\) can be changed to express different levels of confidence in the assumption that the differences are (close to) zero. The larger \(\nu\), the more a prior is uninformative and reflects uncertainty. Values of less than 1 for \(\nu\) are considered small (Asparouhov & Muthén, 2017). Figure 1 shows cross-time differences for a measurement parameter \(r\). Panel a displays exact MI with no differences and panel b shows Bayesian approximate MI with differences in the shape of a normal distribution with a mean of zero and a small variance. The prior variance in panel b implies that 95% of the variation of the differences is between -0.2 and 0.2 (B. Muthén & Asparouhov, 2012, p. 316).

4 In Brief: Principles of Bayesian Analysis

The principles of Bayesian statistics are described in great detail by several authors (Gelman, Carlin, Stern, & Rubin, 2013).\(^2\) Alternative approaches exist that estimate the variance of parameter differences from the data. These are random effect (Fox & Verhagen, 2017) or multilevel models (Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012; Jak, Oort, & Dolan, 2014).

\(^3\)Note that this terminology can be used compatibly for cross-group and cross-time analysis. From here on, we refer to the cross-time case.

\(^4\)Additional restrictions can be tested for other parameters, for example, equal disturbances of the same items across time, equal factor variances and covariances across time, and equal latent factor means across time (Vandenberg & Lance, 2000). However, for our purposes scalar MI is sufficient.
APPROXIMATE MEASUREMENT INVARIANCE

Bayesian approximate measurement invariance (zero mean, small variance prior)

Figure 1. Exact (a) and Bayesian approximate (b) MI (see B. Muthén & Asparouhov, 2012)

Prior beliefs about the parameter of interest are updated with observed data (i.e., the likelihood) resulting in the posterior distribution, which is a trade-off between priors and data (i.e., between expectations and evidence). A strong prior will have more influence on the posterior and a weak prior will give way for the data to have a stronger influence on the posterior. The posterior distribution carries the usually desired information about a parameter. The posterior mean, mode, or median can serve as a point estimate and the posterior standard deviation is an indication of precision.

To determine the posterior parameter distribution, Bayesian estimation procedures use Markov chain Monte Carlo (hereafter: MCMC) sampling methods (Geyer, 2011; Lynch, 2007). MCMC algorithms (e.g., Gibbs sampler, Metropolis-Hastings) iteratively sample from multivariate posterior distributions or conditional distributions of each parameter. Previous parameter values are repeatedly updated each time the algorithm passes through the process until the joint posterior distribution can be approximated. Since MCMC sampling methods directly generate the posterior distribution, asymptotic arguments (e.g., normality of a parameter distribution) are not needed.

The iterative process is called a chain and it is possible to use more than one chain to obtain a posterior solution. The first half of iterations in a chain is usually discarded as “burn-in phase” and the second half is used to approximate the joint posterior distribution. Convergence of the iterative process can be assessed with the potential scale reduction factor (hereafter: PSR) (Gelman & Rubin, 1992). The PSR assesses within- and between-chain variation where between-chain variation should be small compared to within-chain variation (see also Asparouhov and Muthén, 2010b; Gelman and Shirley, 2011; Kruschke, Aguinis, and Joo, 2012).

Models can be compared by means of the deviance infor-

Prior distributions are not necessarily uniform. Commonly used conjugate prior distributions are discussed in Lynch (2007, :68–70)
The discrepancy between generated and observed data is approximated in the posterior, a synthetic data set is generated and the discrepancy (Lynch & Western, 2004; Scheines et al., 1999). Based on model fit can be assessed with posterior predictive checks (Ando, 2010, p. 215; see also Ward, 2008). Approximately equivalent given uninformative prior information (Ellison, 2004) and approximately equivalent given uninformative prior information (Ando, 2010, p. 215; see also Ward, 2008).

The predictive quality of the posterior estimates (i.e., model fit) can be assessed with posterior predictive checks (Lynch & Western, 2004; Scheines et al., 1999). Based on the posterior, a synthetic data set is generated and the discrepancy between generated and observed data is approximated as a distribution of differences of test statistics $T$ (e.g., $\chi^2$) over a number of iterations. The proportion of test statistics $T$ from generated data ($y^{\text{gen}}$) exceeding $T$ from observed data ($y$) is calculated as a posterior predictive $p$ value (hereafter: PPP), indicating the "probability that a future observation would exceed the observed data, given the model" (Lynch, 2007, p. 156):

$\text{PPP} = p(T(y^{\text{gen}}) \geq T(y)|y)$ (4)

Values around .50 imply equally distributed discrepancies and thus good fit. Extreme values, such as PPP less than .05, indicate poor fit.

It should be noted that the DIC and PPP are not suited to evaluate the adequacy of small variance priors (Hoijtink & Van de Schoot, 2017). However, we will use both criteria in conjunction with the procedure of Asparouhov et al. (2015, Appendix A) who propose to run a series of models with different prior variances until increasing the prior variance is considered unnecessary (see below).

The use of Bayesian methodology has steadily increased in recent decades (Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2016). Several studies have covered diverse substantive and methodological areas of research on MI and some successfully used approximate MI in situations where exact parameter constraints have been too strict and suggested incomparability of the variables under study (Bujacz, Vittersø, Huta, & Kaczmarek, 2014; Chan et al., 2015; Chiorri, Day, & Malmberg, 2014; Cieciuch, Davidov, Algesheimer, & Schmidt, 2017; Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Davidov et al., 2015; De Bondt & Van Petegem, 2015; Falkenström, Hatcher, Skjulsvik, Holmqvist Larsson, & Holmqvist, 2015; Jackson, Gucciardi, & Dimmock, 2014; Van de Schoot et al., 2013; Zercher, Schmidt, Cieciuch, & Davidov, 2015). However, to the best of our knowledge there exists no application with panel data.

### 5 Empirical Application

#### 5.1 Data

Data for this analysis is taken from the German sociological and criminological panel study “Crime in the modern city (Crimoc)”. The study examines the onset of delinquency during early adolescence and the development of delinquency and explanatory dimensions during adolescence and young adulthood (Boers, Reinecke, Seddig, & Mariotti, 2010; Seddig, 2014, 2016). The study examines causes, consequences, and the development of delinquent behavior from adolescence to adulthood. In 2002 the study was first implemented in the city of Duisburg, Germany. A cohort of approximately 3,500 seventh-grade students (mean age 13) from different schools was interviewed annually in schools using paper-and-pencil questionnaires or via a postal questionnaire survey. Since 2009 data collection is conducted biennially. In 2017 the 12th survey was conducted.

The current analysis uses 4 waves of panel data covering ages 14, 16, 18, and 20. Although more data points are available, we chose this particular data set for two reasons. First, to observe substantial change in the latent means of hedonism we considered a longer time period spanning from early adolescence until young adulthood. Second, however, to keep the application simple and parsimonious, we decided 4 data points are adequate to illustrate the approximate MI procedure. Hence, we accepted two year gaps between measurement occasions.

The 4 waves of panel data contain information on $N = 1,002$ respondents (64% females; 36% males). Obviously, panel dropout is an issue in the Crimoc study (Reinecke, 2013). This can be explained by the code-based and anonymized panel construction procedure. A self-generated code based on six time-invariant personal characteristics has to be filled out by the respondents prior to each assessment. The code is used to identify data of the same respondent across time. However, many respondents do not reconstruct it correctly. Thus, compared to the separate cross-sectional study the code can be found at http://www.crimoc.org.
samples, the panel contains fewer males, fewer students from secondary schools, and fewer delinquents (Seddig, 2016). Consequently, the levels of hedonistic orientation may be underestimated.

### 5.2 Measures

To illustrate the application of approximate MI with Bayesian CFA we focus on adolescents’ preferences to maximize pleasure and enjoy life. Such a hedonistic orientation is characterized by an affection for enjoyment, excitement, and consumption pronouncing the immediate gratification of short-term desires. As one of the ten universal values captured in Schwartz’s theory of basic human values (Schwartz, 1992, 1994; Schwartz et al., 2012, 4), the concept of hedonism plays not only a prominent role in social psychology and cross-cultural research, but also in sociology (e.g., Hitlin & Piliavin, 2004). Hedonistic orientations are widely accepted among many adolescents and serve as a relevant factor in the explanation of various phenomena, such as youth delinquency (e.g., Boers et al., 2010; Seddig, 2014) or nonparticipation in higher education (e.g., Leitgöb, Paseka, Bacher, & Altrichter, 2012). Hence, we consider the development of hedonistic orientations as an appropriate example for a broad audience of researchers from various disciplines.

We will analyze hedonism with multiple indicators at each time point. Three indicator variables measure the concept of hedonism. Students were asked to rate their “understanding for people who do what they desire” ($y_1$) as well as their personal desires for “living a life of pleasure” ($y_2$) and “excitement” ($y_3$). These items were assessed on a 5-point scale, where 1 indicates the lowest and 5 indicates the highest level of agreement. Thus, we treated the scale as continuous (Finney & DiStefano, 2013; Rhemtulla, Brosseau-Liard, & Savalei, 2012). Table 1 shows item means and standard deviations. Figure 2 shows the longitudinal confirmatory factor model.

### 5.3 Analytical Procedure

We first test exact MI with the traditional (full information) maximum likelihood approach (Finkbeiner, 1979). Therefore, we follow a “bottom-up” strategy and begin with the least restrictive model (configural model). We then consecutively introduce cross-time equality restrictions on the factor loadings (metric model) and intercepts (scalar model). Finally, we replace exact equality constraints for parameters found to be non-invariant across time points with approximate equality constraints by using zero-mean, small-variance priors in the Bayesian approach.

We follow the outline of (Asparouhov et al., 2015, Appendix A) to assess which prior variance allows sufficient wiggle room for parameter differences, but is small enough so that the differences can be considered approximately zero. The procedure requires to begin with a very small prior variance (e.g., $\nu = 0.001$) and then gradually increase it in small steps to achieve identified solutions that allow to separate minor parameter differences from model misspecifications. The process is monitored with regard to model fit (DIC, PPP, and the 95% credibility intervals for the difference between observed and replicated chi-square values) and convergence. Increasing the prior variance is considered unnecessary when

---

**Table 1**

*Means and standard deviations (n=1,002)*

<table>
<thead>
<tr>
<th>Age</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>3.00</td>
<td>1.19</td>
<td>3.00</td>
<td>1.17</td>
<td>3.00</td>
<td>1.23</td>
</tr>
<tr>
<td>16</td>
<td>2.80</td>
<td>1.14</td>
<td>2.81</td>
<td>1.13</td>
<td>2.80</td>
<td>1.17</td>
</tr>
<tr>
<td>18</td>
<td>2.49</td>
<td>1.12</td>
<td>2.45</td>
<td>1.13</td>
<td>2.57</td>
<td>1.11</td>
</tr>
<tr>
<td>20</td>
<td>2.29</td>
<td>1.03</td>
<td>2.21</td>
<td>1.01</td>
<td>2.44</td>
<td>1.05</td>
</tr>
</tbody>
</table>

$y_1$ (Desires)

$y_2$ (Excitement)

$y_3$ (Pleasure)
model fit differences between models become irrelevant.

The program Mplus Version 8 (L. Muthén & Muthén, 1998-2017) was used with the Gibbs sampler and two MCMC chains in this study. A process is assumed to be converged when the second half of the iterations has PSR values lower than defined by the default of the Mplus command “bconvergence” (Asparouhov & Muthén, 2010b). The treatment of missing data in Bayesian estimation is similar to full information maximum likelihood, that is, the full information available from the data is used assuming missing at random (Asparouhov & Muthén, 2010a, 2010b).

### 5.4 Results

The results based on maximum likelihood CFA suggest that at least exact metric MI is supported (Table 2). Global fit statistics for scalar MI are also within the acceptable range (Hu & Bentler, 1999). However, the $\chi^2$-difference to the metric model is very large. Further, Chen (2007) developed criteria for other fit statistics to indicate when MI is not given. According to these criteria, change in RMSEA should be less than .015, change in SRMR should be less than .010, and change in CFI should be less than -.010. The scalar model does not meet these criteria. This implies a misspecification in the scalar model that leads to a substantial decrease in model fit, but not to an overall rejection of the model. However, although the global fit of the scalar model is tolerable, the metric model may be the better choice. We admit that this is a very strict interpretation of model fit differences, which should be used with caution and may only serve illustrative purposes in this example. However, the Bayesian exact scalar MI model by far has the worst fit (upper section in Table 3).\footnote{In all models, autocorrelated disturbances were considered to fit the data structure appropriately (see the Mplus input in Figure 3).}

According to the maximum likelihood modification indices, the misspecification in the scalar model is located at the intercept level. The intercepts of items $y_{1}$ and $y_{3}$ are non-invariant at time points 1 and 4 (age 14 and 20). Therefore,
Figure 3. Mplus input

Table 3

Bayesian CFA Model Fit (N = 1,002)

<table>
<thead>
<tr>
<th>Prior</th>
<th>DIC</th>
<th>PPP</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural</td>
<td>32976</td>
<td>0.350</td>
<td>[−28.4, +40.5]</td>
</tr>
<tr>
<td>Metric</td>
<td>32972</td>
<td>0.306</td>
<td>[−26.0, +42.5]</td>
</tr>
<tr>
<td>Scalar</td>
<td>33069</td>
<td>0.000</td>
<td>[+76.0, +144.1]</td>
</tr>
</tbody>
</table>

**Approximate MI**

<table>
<thead>
<tr>
<th>Prior</th>
<th>DIC</th>
<th>PPP</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalar</td>
<td>N(0,0.001)</td>
<td>33014</td>
<td>0.002</td>
</tr>
<tr>
<td>Scalar</td>
<td>N(0,0.005)</td>
<td>32979</td>
<td>0.174</td>
</tr>
<tr>
<td>Scalar</td>
<td>N(0,0.010)</td>
<td>32974</td>
<td>0.249</td>
</tr>
<tr>
<td>Scalar</td>
<td>N(0,0.100)</td>
<td>32972</td>
<td>0.302</td>
</tr>
<tr>
<td>Scalar</td>
<td>N(0,0.500)</td>
<td>32973</td>
<td>0.293</td>
</tr>
</tbody>
</table>

DIC = deviance information criterion; PPP = posterior predictive p-value; CI = credibility interval; MI = measurement invariance.

we specified zero-mean, small-variance priors for the differences of all intercepts across time. We began the Bayesian analyses with a very small prior variance and successively increased it while we monitored DIC, PPP, and 95% credibility intervals for the difference between observed and replicated chi-square value. Although this strategy is not a test of the adequacy of the prior variances (Hoijtink & Van de Schoot, 2017), it is possible to decide if model fit substantially improves with larger prior variances. An extract of the Mplus input for the approximate MI model with prior variance \( \nu = 0.010 \) is given in Figure 3.

The results show that the smallest prior variance \( \nu = 0.001 \) does not fit the data, i.e. \( \nu \) is too close to zero. By gradually adjusting \( \nu \) we were able to achieve good model fit while maintaining convergence and identifiability. The less strict prior variances \( \nu = 0.100 \) and \( \nu = 0.500 \) reveal no major improvement of model fit compared to \( \nu = 0.010 \). Moreover, models with \( \nu = 0.100 \) and \( \nu = 0.500 \) are very slow to converge, need much more iterations, and the quality of model identification diminishes. Further, the DIC, PPP, and limits of 95% credibility interval for the difference between observed and replicated chi-square values for \( \nu = 0.010 \) do not substantially differ from the exact metric MI model. Thus, \( \nu = 0.010 \) is deemed sufficient to consider minor deviations from exact intercept equivalence and we assume that approximate scalar MI holds. Thus, it is reasonable to compare the latent means of hedonism across time.

Figure 4 shows a part of the Mplus results output for the model with prior variance \( \nu = 0.010 \). It can be seen that the factor loadings are held exactly equal across time and small differences exist for the intercepts across time. The latent means of hedonism across time imply that a hedonistic orientation declines in the process of maturation from youth to young adulthood.

The latent means for hedonism across time can be compared with regard to the use of exact or approximate MI (Table 4). The decline of hedonism can be observed regardless of the model. However, two differences are noteworthy.

---

8The model with \( \nu = 0.500 \) has poor convergence even after 500,000 iterations. 100,000 were sufficient for \( \nu = 0.010 \). 100,000 iterations were used for all other models, except 200,000 were used for \( \nu = 0.100 \).

9In this specific case (but not necessarily in other cases), the
Table 4
Hedonism Latent Mean Comparison (n=1,002)

<table>
<thead>
<tr>
<th>Age</th>
<th>Exact MI (ML) Mean</th>
<th>Std. Err.</th>
<th>Exact MI (Bayes) Mean</th>
<th>Std. Err.</th>
<th>Appr. MI (Bayes) Mean</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>16</td>
<td>-0.161</td>
<td>0.049</td>
<td>-0.167</td>
<td>0.051</td>
<td>-0.144</td>
<td>0.113</td>
</tr>
<tr>
<td>18</td>
<td>-0.574</td>
<td>0.056</td>
<td>-0.600</td>
<td>0.062</td>
<td>-0.556</td>
<td>0.116</td>
</tr>
<tr>
<td>20</td>
<td>-0.777</td>
<td>0.061</td>
<td>-0.813</td>
<td>0.070</td>
<td>-0.738</td>
<td>0.123</td>
</tr>
</tbody>
</table>

MODEL RESULTS

<table>
<thead>
<tr>
<th>hedo1 BY</th>
<th>Posterior Estimate</th>
<th>S.D.</th>
<th>One-Tailed P-Value</th>
<th>95% C.I.</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>y11</td>
<td>0.668</td>
<td>0.044</td>
<td>0.000</td>
<td>0.581</td>
<td>0.754</td>
</tr>
<tr>
<td>y21</td>
<td>0.523</td>
<td>0.037</td>
<td>0.000</td>
<td>0.450</td>
<td>0.596</td>
</tr>
<tr>
<td>y31</td>
<td>0.514</td>
<td>0.035</td>
<td>0.000</td>
<td>0.444</td>
<td>0.583</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hedo2 BY</th>
<th>Posterior Estimate</th>
<th>S.D.</th>
<th>One-Tailed P-Value</th>
<th>95% C.I.</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>y12</td>
<td>0.668</td>
<td>0.044</td>
<td>0.000</td>
<td>0.581</td>
<td>0.754</td>
</tr>
<tr>
<td>y22</td>
<td>0.523</td>
<td>0.037</td>
<td>0.000</td>
<td>0.450</td>
<td>0.596</td>
</tr>
<tr>
<td>y32</td>
<td>0.514</td>
<td>0.035</td>
<td>0.000</td>
<td>0.444</td>
<td>0.583</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hedo3 BY</th>
<th>Posterior Estimate</th>
<th>S.D.</th>
<th>One-Tailed P-Value</th>
<th>95% C.I.</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>y13</td>
<td>0.668</td>
<td>0.044</td>
<td>0.000</td>
<td>0.581</td>
<td>0.754</td>
</tr>
<tr>
<td>y23</td>
<td>0.523</td>
<td>0.037</td>
<td>0.000</td>
<td>0.450</td>
<td>0.596</td>
</tr>
<tr>
<td>y33</td>
<td>0.514</td>
<td>0.035</td>
<td>0.000</td>
<td>0.444</td>
<td>0.583</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hedo4 BY</th>
<th>Posterior Estimate</th>
<th>S.D.</th>
<th>One-Tailed P-Value</th>
<th>95% C.I.</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>y14</td>
<td>0.668</td>
<td>0.044</td>
<td>0.000</td>
<td>0.581</td>
<td>0.754</td>
</tr>
<tr>
<td>y24</td>
<td>0.523</td>
<td>0.037</td>
<td>0.000</td>
<td>0.450</td>
<td>0.596</td>
</tr>
<tr>
<td>y34</td>
<td>0.514</td>
<td>0.035</td>
<td>0.000</td>
<td>0.444</td>
<td>0.583</td>
</tr>
</tbody>
</table>

Means

| hedo1    | 0.000  | 0.000  | 1.000  | 0.000  | 0.000  |
| hedo2    | -0.144 | 0.113  | 0.099  | -0.369 | 0.075  |
| hedo3    | -0.556 | 0.116  | 0.000  | -0.792 | -0.334 |
| hedo4    | -0.738 | 0.123  | 0.000  | -0.991 | -0.511 |

Intercepts

| y11      | 2.986  | 0.037  | 0.000  | 2.913  | 3.058  | *          |
| y21      | 2.448  | 0.035  | 0.000  | 2.380  | 2.516  | *          |
| y31      | 3.125  | 0.035  | 0.000  | 3.057  | 3.194  | *          |
| y12      | 2.301  | 0.076  | 0.000  | 2.753  | 3.052  | *          |
| y22      | 2.486  | 0.062  | 0.000  | 2.364  | 2.608  | *          |
| y32      | 3.195  | 0.061  | 0.000  | 3.076  | 3.315  | *          |
| y13      | 2.865  | 0.076  | 0.000  | 2.717  | 3.016  | *          |
| y23      | 2.425  | 0.063  | 0.000  | 2.305  | 2.549  | *          |
| y33      | 3.386  | 0.061  | 0.000  | 3.187  | 3.426  | *          |
| y14      | 2.793  | 0.080  | 0.000  | 2.642  | 2.954  | *          |
| y24      | 2.397  | 0.064  | 0.000  | 2.274  | 2.524  | *          |
| y34      | 3.421  | 0.062  | 0.000  | 3.303  | 3.544  | *          |

Figure 4. Mplus model results output (v = 0.010; point estimates are medians)
First, the decline appears to be less pronounced in the approximate MI solution. Second, the standard errors are larger in the approximate MI solution. For example, the change in latent means of hedonism between age 14 and 16 is considered not significant with approximate MI, but significant with both exact MI solutions.

Finally, the non-invariant parameters can be identified based on the approximate MI solution (B. Muthén & Asparouhov, 2013). Non-invariance is determined as the difference of a particular parameter at a time point from the average of estimates for the particular parameter across time. If a difference of zero is outside of the 2.5% and 97.5% quantiles of the posterior distribution of differences, the difference is assumed to be significant and the parameter is assumed non-invariant. Figure 5 shows the difference output produced by Mplus for the Bayesian CFA model with prior variance \( \nu = 0.010 \). The column “Average” refers to the average of estimates for parameters across time. Significant deviations from the average (i.e., non-invariant parameters) are labeled with an asterisk in the columns “Deviations from the Mean”. The results indicate that the intercepts of items 1, 2, 3, and 4 significantly deviate from the cross-time average at time points 1 and 4.

### 6 Conclusion

The analyses reported in this paper used Bayesian approximate MI to test the comparability of students’ preferences to maximize pleasure and enjoy life (“hedonism”). Since comparability is important for longitudinal research and the study of processes and change of latent variables, a careful assessment of MI has to be conducted before drawing substantive conclusions. The Bayesian approach offers a flexible alternative to the exact approach where all model parameters are held exactly equal across time points. Approximate MI allows for a small degree of deviation from exact zero constraints while the differences across time points are kept at a minimum.

Regardless of using maximum likelihood or Bayesian estimation, the results of testing the longitudinal MI of hedonism showed that exact scalar MI may not be the best solution. However, the use of zero-mean, small-variance priors lead to support for approximate scalar MI. A small prior variance (\( \nu = 0.010 \)) was deemed sufficient to account for minor differences of item intercepts across time. These minor differences already led to a substantial deterioration of fit of the exact scalar MI model and substantive conclusions may not be trustworthy. The Bayesian approach showed that the latent means are comparable at least based on approximate MI. Although the differences in the latent means from exact and approximate MI models were small, they may be more severe in other studies. In any case, a careful test of invariance properties is recommended over alternative strategies, such as using observed composite scores (Steinmetz, 2013).

For practical researchers it may not be easy to decide when and how to use the Bayesian approach. However, considering what can be achieved by using small variance priors for testing MI may open up a new perspective. When model fit of more restrictive (e.g., scalar) models is worse than that of less restrictive (e.g., metric) models, researchers may take a closer look at the actual parameter differences that elicit the misfit. Sometimes already very small parameter differences across time may lead to a decision against a particular model in the traditional (maximum likelihood) approach. In this case, using the Bayesian approach to reduce the stringency of parameter equality may be very reasonable and comparability can be justified, even if the parameter differences precluded a comparison using the traditional approach. However, if the differences are actually large and not regarded as negligible, too much leeway for the parameter differences is not advisable. Very large prior variances may be too vague and the estimation is dominated entirely by the data. Consequently, assuming comparability may be reasonable, even if model fit is good. In contrast, very small prior variances
may be too close to exact zero differences and result in poor model fit. A rule of thumb for limits of the prior variance does not seem to be useful. Many times a balance is needed between the desired condition (i.e., exact parameter equality) and the situation found in the data. This balance can be achieved by the Bayesian posterior distribution of parameter differences: while the irregularities in the data are considered the differences are kept as close to zero as possible.

The results presented in this paper do not rely on a strict test of the hypothesis that parameter differences defined by the prior variance are approximately zero. We applied the strategy outlined by Asparouhov et al. (2015) where no single prior is preferred over another. A test of the particular hypothesis that parameters are approximately zero is the prior-posterior predictive p-value (Asparouhov & Muthén, 2017; Hoijtink & Van de Schoot, 2017). However, Mplus does not yet provide the prior-posterior predictive p-value for approximate MI.

Further, the choice of a different than a zero-mean prior may be considered in cases where different types of response bias may give rise to assumptions about systematic reporting tendencies that contradict measurement equivalence (Billiet & Davidov, 2008; Cheung & Rensvold, 2000; Welkenuysen-Gybels, Billiet, & Cambré, 2003). If, for example, a shift in the intercept or factor loading of a particular item is expected across measurement occasions, this could be expressed with a non-zero mean for the prior.

Van de Schoot et al. (2013) pointed out that approximate MI may not be appropriate in all situations. They suggest partial approximate MI as a further approach that may lead to even better solutions. Partial approximate MI allows minor differences only for those parameters that are non-invariant. Further, in the presence of large parameter differences, non-invariant parameters tend to be pulled toward the average of the parameter estimates which causes bias in latent variable means (Van de Schoot et al., 2013). A method to compensate for this bias is alignment (Asparouhov & Muthén, 2014; Marsh et al., 2017; B. Muthén & Asparouhov, 2014) where a component loss function (Jennrich, 2006) is used to generate solutions with many approximately invariant parameters and very few non-invariant parameters. However, using strict, partial, or approximate MI, or even applying the alignment procedure must be judged on a case-by-case basis. The equality constraints imposed by any of these approaches are sensible to the unique data set that is analyzed.

Acknowledgements

The authors would like to thank the directors of the “Crime in the modern city (Crimoc)” study, Prof. Klaus Boers and Prof. Jost Reinecke, for granting us access to the data.

References


