# BSEM Measurement Invariance Analysis

*Bengt Muthén & Tihomir Asparouhov*

Mplus Web Notes: No. 17

January 11, 2013

**Abstract**

This paper concerns measurement invariance analysis for situations with many groups or time points. A BSEM (Bayesian Structural Equation Modeling) approach is proposed for detecting non-invariance that is similar to modification indices with maximum-likelihood estimation, but unlike maximum-likelihood is applicable also for high-dimensional latent variable models for categorical variables. Under certain forms of non-invariance, BSEM gives proper comparisons of factor means and variances using only approximate measurement invariance and without relaxing the invariance specifications or deleting non-invariant items. To ensure correct estimation, a two-step Bayesian analysis procedure is proposed, where step 1 uses BSEM to identify non-invariant parameters and step 2 frees those parameters.

# 1 Introduction

Muthén and Asparouhov (2012a) introduced a special type of Bayesian structural equation modeling that provides a more flexible confirmatory factor analysis (CFA) and structural equation model (SEM) analysis. Instead of translating hypotheses into fixed zero parameters which are characteristic of such analyses, the authors proposed the use of approximately zero parameters using zero-mean, small-variance informative priors via Bayesian analysis. The approach is referred to as BSEM and is mainly intended for cross-loadings and residual correlations in CFA and for direct effects representing measurement non-invariance in CFA with covariates, also referred to as MIMIC modeling. An application of BSEM to competing factor analysis models for the Wechsler Intelligence Scale is given in Golay et al. (2012). The current paper generalizes the BSEM approach to the analysis of measurement invariance across several groups or several time points, applying the zero-mean, small-variance prior idea to differences in measurement parameters, thereby introducing the concept of approximate measurement invariance.

Measurement invariance CFA using maximum-likelihood estimation often takes as a baseline model the specification of fully-invariant measurement parameters and then uses modification indices as guidance to improve the model by freeing parameters that strongly violate invariance. This procedure is suitable for cases with only a few groups/time points, typically only two, and where only a small number of measurement parameters are non-invariant. Modification indices have not been used in item response analysis, presumable for computational reasons due to the numerical integration necessary for maximum-likelihood estimation with categorical variables. For categorical variables, maximum-likelihood estimation with likelihood ratio chi-square testing of item parameters across groups was discussed in Thissen et al. (1993). In this context, backward and forward maximum-likelihood approaches have been studied in (Kim & Yoon, 2011; Lee et al., 2010; Stark et al., 2006). The backward approach starts with a fully invariant model and then releases the invariance specification for one item at a time, producing a likelihood-ratio chi-square test of invariance for each item. The forward approach starts with a fully non-invariant model, comparing it to models where one item at a time is held invariant. Depending on the type of application, non-invariant items are either deleted or their invariance specification relaxed. Cai et al. (2011; pp. 239-243) discussed a multi-step Wald procedure with a designated set of invariant items. De Jong et al. (2007) and Fox (2010) proposed a novel random parameter approach to non-invariance where a measurement parameter for different groups is seen as randomly drawn from a population with shared mean and variance and where a significant variance corresponds to non-invariance. Bou and Satorra (2010) criticize the assumption of random draws in this approach in favor of a multiple-group approach.

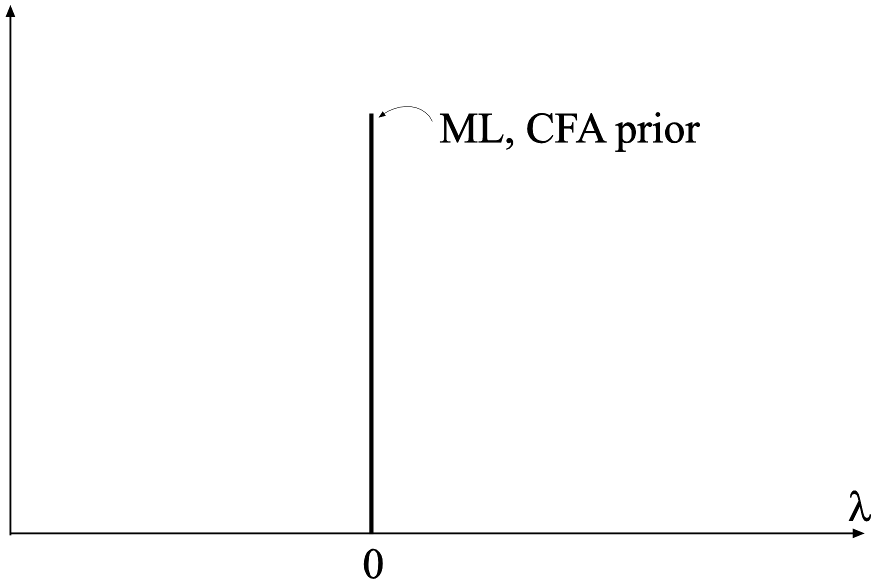
In many applications, measurement invariance needs to be studied for many groups or many time points where there is possibly a large number of non-invariant item parameters. The case of many groups/time points is the focus of this paper. Typical applications are country comparisons such as with the achievement studies of PISA (Program for International Student Assessment) conducted by the Organization for Economic Co-operation and Development, with cross-cultural studies such as the European Social Survey (see, e.g., Davidov et al., 2011), as well as with growth modeling with many time points. In these cases, an item parameter may be invariant across some groups/time points and non-invariant across some other groups/time points. Here, proposed backward and forward maximum-likelihood approaches and multi-step Wald procedures may be too cumbersome to be practical. Furthermore, although maximum-likelihood modification index procedures can be developed for categorical items, many applications involve high-dimensional latent variable models where the computations would be too heavy.

CFA is characterized by using strict parameter constraints. In single-group CFA these are fixed zero loadings corresponding to items hypothesized to not load on certain factors (so called cross-loadings). In multiple-group and multiple-time point CFA these are equalities of loadings and intercepts across groups and time point. As illustrated in the top part of Figure 1, from a Bayesian perspective the zero cross-loading case can be seen as CFA using a very strong prior with mean zero and zero variance for certain factor loadings, here represented by a parameter $\lambda$. The bottom part of Figure 1 shows the more flexible BSEM approach of using a zero-mean, small-variance prior for the parameter as discussed in Muthén and Asparouhov (2012a). In multiple-group and multiple-time point modeling considered in this paper, the parameter in question concerns the difference in a certain measurement parameter across two groups or time points. This BSEM approach implies the specification of approximate measurement invariance. Appendix, Section 9.1 presents the Mplus specification of such zero-mean, small-variance priors for differences between parameters.

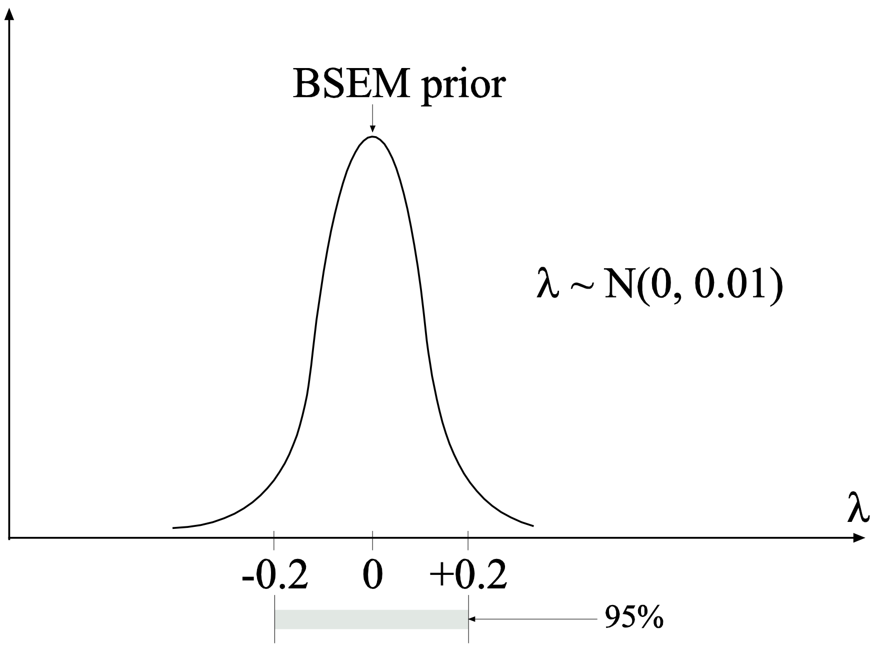
In this paper it is shown that the BSEM approach provides a tool for detecting non-invariance that serves the same purpose as modification indices with maximum-likelihood estimation, but unlike maximum-likelihood is applicable also for high-dimensional latent variable models for categorical variables. Under

Figure 1: ML versus Bayes priors

(a)




(b)



certain forms of non-invariance, proper comparisons of factor means and variances are obtained using only approximate measurement invariance and without relaxing the invariance specifications or deleting non-invariant items. To ensure correct estimation, a two-step Bayesian analysis procedure is proposed, where step 1 uses BSEM to identify non-invariant parameters and step 2 frees those parameters.

Section 2 discusses an alignment issue which reflects an indeterminacy in measurement invariance analysis. Section 3 presents the BSEM approach for detecting non-invariance. Section 4 discusses a simulation study of a multiple-group factor analysis model. Section 5 present an application with binary items measuring an achievement factor in 40 countries. Section 6 discusses a simulation study of a multiple indicator growth model. Section 7 presents a binary multiple indicator growth model application for eight time points. Section 8 concludes.

## 2 The alignment issue

As pointed out in Asparouhov and Muthén (2012b), the BSEM method for multiple-group, multiple-time point analysis, as well as other methods for such settings, faces a parameterization indeterminacy, or put differently an alignment issue. Briefly stated, the approximate invariance of BSEM essentially acts to find a solution where the variance across groups or time points for a measurement parameter is small. This is not the same simplicity criterion as for example seeking a solution that has many invariant parameters and few non-invariant parameters. The latter simplicity criterion may be preferred by researchers who hypothesize a large degree of invariance, while the BSEM criterion may be preferred by researchers who hypothesize that there is a large degree of minor non-invariance where different non-invariance may be in opposite directions and largely cancel each other out.

As a hypothetical example of the BSEM approach applied to the case with few non-invariant parameters, consider an item where there is invariance of a measurement intercept or threshold across all time points except for one time point where there is a large positive deviation. The BSEM small-variance prior for the parameter differences tends to pull the deviating parameter towards the average of the parameters for all the time points. This means that the deviating parameter will be smaller and the invariant parameters larger than their true values. The same thing occurs for loadings. With intercepts/thresholds and loadings misestimated, the factor means and factor variances are misestimated. This is the alignment issue. As Asparouhov and Muthén (2012b) point out, it does not mean that the model does not fit the data, but that an equally well-fitting solution with a possibly simpler interpretation due to another simplicity criterion may be available.

The alignment issue should be kept in mind in Monte Carlo simulation studies of the proposed approach. The BSEM analysis for multiple groups or time points is not expected to always recover parameter values used to simulate the data. Perfect recovery will only occur when the non-invariance is designed to be in line with BSEM, such as random, normally-distributed deviations from a parameter average over groups/time points. When non-invariant items have been detected and equalities relaxed, however, the parameter values will be recovered.

Given that in the above hypothetical example there is a majority of invariant time points, the underestimation of the non-invariant parameter will be smaller than the overestimation of the invariant parameters. Only the non-invariant parameter is therefore likely to be detected as significantly different from the average over time points. Such parameters can then be freed from their invariance restriction in a subsequent analysis that gives correct results. This procedure is discussed in the next section.

# 3 Detecting non-invariance

The BSEM analysis is augmented with estimation of the difference between each measurement parameter and its average across the groups or time points. Assuming that the parameter is approximately invariant for most of the groups/time points, these differences can point to the groups/time points that have significant non-invariance. In a follow-up analysis, the equality constraint for the parameter can then be relaxed for those groups/time points. This follow-up analysis would use Bayesian analysis but not specify BSEM approximate measurement invariance for any of the parameters but instead hold all but the non-invariant parameters exactly equal across groups/time points. This two-step approach serves the same purpose as working with modification indices with maximum-likelihood estimation. Instead of the modification index approach of relaxing one equality restriction at a time, however, it is here proposed that all significantly mis-fitting equalities are relaxed at the same time. One consideration is the risk of relaxing too many equality restrictions. This should not, however, have much effect on the point estimates, but only increase standard errors to a small extent. With too many equality restrictions relaxed, a non-identification problem arises, but this should be clearly seen in the Bayes plots (for Bayes non-identification detection, see the rejoinder discussion in Muthén & Asparouhov, 2012a).

The parameter differences across groups/time points are printed in an Mplus output section labeled DIFFERENCE OUTPUT. In Monte Carlo simulations, the output also contains information in parentheses about the proportion of replications for which the difference is significant, providing an estimate of the power to detect the non-invariance. With only two groups/time points, the difference relative to the average can be augmented by the difference across the two groups/timepoints which can be expressed in MODEL CONSTRAINT.

The choice of prior variance for the differences between parameters is important for detecting non-invariance. As the prior variance is increased, the non-invariance of a parameter is allowed to be more freely estimated, that is, the estimate can escape from the invariance value to a larger degree. At the same time, the standard error of the parameter increases as the prior variance is increased (see also Muthén & Asparouhov, 2012a). The significance of a parameter difference considers a ratio where the numerator is the invariance value and the denominator is the standard error. The numerator increases with increased prior variance and the denominator increases with increased prior variance so that it is difficult to predict how significance is going to be affected. A balance needs to be found for the best prior variance choice. Simulation studies which include these choices are presented in the next section and in Section 6.

# 4 A multiple-group factor analysis simulation

Consider a situation with continuous items $y$ measuring a single factor in different groups, using the factor model for individual $i$, variable $j$, and group $g$,

$$y_{ijg} = \nu_{jg} + \lambda_{jg}\,\eta_{ig} + \epsilon_{ijg}, \tag{1}$$

$$E(\eta_{ig}) = \alpha_g, V(\eta_{ig}) = \psi_g, \tag{2}$$

$$E(\epsilon_{ijg}) = 0, V(\epsilon_{ijg}) = \theta_{jg}, \tag{3}$$

where the residual $\epsilon$ is uncorrelated with the factor $\eta$. The measurement parameters are the intercepts $\nu_{jg}$ and the loadings $\lambda_{jg}$. The focus is on correctly estimating the factor means $\alpha_g$ and factor variances $\psi_g$. The factor metric is set as mean zero and variance one in one group.

The Appendix Section 9.2 shows the Mplus Monte Carlo simulation input using six items, 10 groups, and a small degree of non-invariance for seven of the 60 intercepts and for seven of the 60 loadings. Only item 1 is invariant across all groups. Groups 1 and 10 have no non-invariant items. The intercept non-invariance is an increase of 0.2 in group 2 for item 6, in group 3 for item 5, in group 4 for item 3, in group 5 for item 2, in group 7 for item 6, in group 8 for item 5, and in group 9 for item 3. The loading non-invariance is an increase of 0.2 in group 2 for item 3, in group 3 for item 2, in group 4 for item 6, in group 5 for item 5, in group 7 for item 3, in group 8 for item 2, and in group 9 for item 5. The intercept non-invariance corresponds to an average of about 0.2 standard deviations of the items and the loading non-invariance about 0.15 in a standardized metric. The small number of non-invariant items and the small magnitude of non-invariance correspond to a well-developed measurement instrument.

As a devil's advocate approach, the settings for the simulation are chosen as the least favorable for the proposed two-step BSEM procedure in two regards. First, there is limited non-invariance so that an approach that simply ignores the non-invariance is a strong contender. Second, the form of non-invariance is not of the type that is assumed for BSEM approximate invariance. The non-invariance is in only one direction with higher values than the average over time points in each non-invariant instance for both loadings and intercepts. This implies that the alignment will not be perfect and the population values used to generate the data will not be perfectly recovered. It is of interest to see to which degree recovery occurs and what the coverage values are. A determinant of this is the power to detect non-invariance and the study of how well an analysis with relaxed invariance for non-invariant items performs.

As a first step, three BSEM analyses are done to study the power to detect non-invariance. Prior variances for the differences are set at 0.01, 0.05, and 0.10, respectively. Using the three different prior variances, Table 1, Table 3, and Table 4 show power estimates in terms of the percentage of replications that show significant differences from the mean across groups, that is, the proportion of replications where the 95% credibility interval for the difference does not include zero. Power estimates for the non-invariant parameters are bolded. In Table 1 the prior variance 0.01 is used and all 14 non-invariant parameters have power estimates close to 100%. The largest power estimate for an invariant parameter is 52% for the loading of item 3 in group 3, showing that it is possible to make an incorrect diagnosis of non-invariance. Most power estimates for invariant parameter are, however, much smaller. It is therefore likely that the correct

Table 1: Multiple-group factor analysis simulation: Power estimates with prior variance 0.01

| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Loadings | | | | | | | | | | |
| Item 1 | 00 | 02 | 02 | 00 | 00 | 00 | 01 | 00 | 00 | 04 |
| Item 2 | 00 | 00 | **100** | 00 | 01 | 00 | 00 | **100** | 03 | 03 |
| Item 3 | 17 | **100** | 52 | 22 | 02 | 17 | **100** | 05 | 02 | 05 |
| Item 4 | 00 | 01 | 01 | 00 | 00 | 00 | 01 | 00 | 01 | 02 |
| Item 5 | 03 | 00 | 01 | 00 | **96** | 04 | 01 | 48 | **98** | 25 |
| Item 6 | 01 | 04 | 25 | **100** | 00 | 02 | 03 | 01 | 00 | 01 |
| Intercepts | | | | | | | | | | |
| Item 1 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 01 |
| Item 2 | 00 | 01 | 00 | 02 | **100** | 01 | 01 | 07 | 02 | 00 |
| Item 3 | 08 | 01 | 14 | **100** | 12 | 04 | 01 | 13 | **100** | 07 |
| Item 4 | 00 | 00 | 00 | 00 | 00 | 01 | 00 | 00 | 00 | 01 |
| Item 5 | 00 | 04 | **96** | 04 | 23 | 00 | 04 | **95** | 14 | 02 |
| Item 6 | 02 | **93** | 17 | 06 | 14 | 01 | **96** | 16 | 09 | 03 |

non-invariant parameters are identified.

Table 2 shows the Mplus output in the DIFFERENCE section for the intercept of the second item using prior variance 0.01. The fifth group has a difference estimate of 0.140 with power estimate 1.00. The other nine groups show smaller negative differences, so that all differences sum to zero.

Table 3 and Table 4 use the prior variances 0.05 and 0.10, respectively, showing decreasing power. With prior variance 0.10 several of the non-invariant parameters have power estimates smaller than the typical limit of 80%, indicating that a

Table 2: Difference output for the intercept of the second item using prior variance 0.01

| NU1_2 | NU2_2 | NU3_2 | NU4_2 | NU5_2 |
|---|---|---|---|---|
| -0.002(0.00) | -0.018(0.01) | -0.016(0.00) | -0.025(0.02) | 0.140(1.00) |
| NU6_2 | NU7_2 | NU8_2 | NU9_2 | NU10_2 |
| 0.000(0.01) | -0.013(0.01) | -0.033(0.07) | -0.024(0.02) | -0.008(0.00) |

larger sample size for each group is necessary for this size non-invariance when prior variance 0.10 is used. The low power estimate of 15% for item 5 in group 8 corresponds to an intercept that has a non-invariance of 0.16 of a standard deviation of the item. Put differently, for this small magnitude of non-invariance the smaller prior variance of 0.01 is more suitable.

To study how well the factor means and variances can be recovered by different approaches, five different types of Bayesian analyses are performed:

1. Exact invariance over groups for all loadings and intercepts. It is expected that this analysis gives biased results in this setting

2. BSEM analysis using approximate invariance of all loadings and intercepts across groups, assuming ignorance of where the non-invariance resides. This analysis provides information about non-invariance, that is, parameters that are significantly different from the average over groups for the respective item parameter. The analysis is not expected to recover the population values exactly due to the alignment issue. It is unknown if this analysis improves

Table 3: Multiple-group factor analysis simulation: Power estimates with prior variance 0.05

| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Loadings | | | | | | | | | | |
| Item 1 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| Item 2 | 00 | 00 | **94** | 00 | 00 | 00 | 00 | **100** | 01 | 00 |
| Item 3 | 02 | **94** | 12 | 09 | 00 | 00 | **98** | 00 | 01 | 17 |
| Item 4 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 01 |
| Item 5 | 01 | 00 | 00 | 00 | **78** | 00 | 00 | 02 | **78** | 00 |
| Item 6 | 00 | 03 | 05 | **98** | 00 | 00 | 00 | 00 | 00 | 06 |
| Intercepts | | | | | | | | | | |
| Item 1 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| Item 2 | 00 | 00 | 00 | 00 | **86** | 00 | 00 | 01 | 00 | 00 |
| Item 3 | 02 | 00 | 04 | **96** | 03 | 00 | 00 | 05 | **98** | 03 |
| Item 4 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| Item 5 | 00 | 00 | **73** | 00 | 03 | 00 | 00 | **53** | 00 | 00 |
| Item 6 | 00 | **89** | 09 | 02 | 03 | 01 | **92** | 02 | 02 | 00 |

Table 4: Multiple-group factor analysis simulation: Power estimates with prior variance 0.10

| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Loadings | | | | | | | | | | |
| Item 1 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 14 |
| Item 2 | 00 | 00 | **62** | 00 | 00 | 00 | 00 | **96** | 01 | 19 |
| Item 3 | 00 | **75** | 05 | 01 | 00 | 00 | **80** | 00 | 00 | 51 |
| Item 4 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 20 |
| Item 5 | 00 | 00 | 00 | 00 | **59** | 00 | 00 | 00 | **56** | 04 |
| Item 6 | 00 | 00 | 01 | **89** | 00 | 00 | 00 | 00 | 00 | 26 |
| Intercepts | | | | | | | | | | |
| Item 1 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| Item 2 | 00 | 00 | 00 | 00 | **47** | 00 | 00 | 00 | 00 | 00 |
| Item 3 | 01 | 00 | 01 | **79** | 00 | 00 | 00 | 00 | **80** | 01 |
| Item 4 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| Item 5 | 00 | 00 | **30** | 00 | 00 | 00 | 00 | **15** | 00 | 00 |
| Item 6 | 00 | **74** | 03 | 00 | 00 | 00 | **76** | 00 | 00 | 00 |

on analysis 1

3. Exact invariance for invariant items based on analysis 2. and BSEM approximate invariance for other items. It is unknown if utilizing invariant "anchor" items improves on analysis 2. The invariant item 1 is the anchor item in the simulation

4. Free non-invariant parameters based on analysis 2 and BSEM approximate invariance for other parameters. It is unknown if freeing the non-invariant parameters improves on analysis 2. The 14 non-invariant parameters are freed in the simulation

5. Exact invariance for only the invariant parameters and non-invariant parameters free. This analysis is expected to perform optimally given that this is how the data are generated, assuming that the correct non-invariance is diagnosed in analysis 2

Analysis 5 is the proposed two-step procedure. A sample size of $n = 500$ for each group (total sample size 5,000) and 100 replications are used in the five simulation analyses.

Table 5 shows the factor mean and variance estimates and coverage using the five approaches. The results for the 10th group are not shown because this is the metric-setting group with fixed values. The last column gives the average absolute bias over the first nine groups.

Analysis 1 specifies exact invariance across all 10 groups for all loadings and intercepts. The biases are not large in this case. The 95% coverage for the factor variances is, however, too low in four of the groups.

Analysis 2a shows the BSEM analysis with approximate measurement invariance and prior variance 0.10. This analysis gives worse bias than analysis 1. due to the alignment issue. Analysis 2b uses BSEM with prior variance 0.01 which was found to better match the generated non-invariance and to give better power to detect non-invariance. Analysis 2b improves on analysis 2a but still performs worse than analysis 1 in terms of bias.

Analysis 3 uses item 1 as an anchor item with exact invariance and does perform better than analysis 2b. It performs better than analysis 1 in terms of bias for factor means although not for factor variances.

Analysis 4 frees the non-invariant parameters and thereby improves on analysis 2b and analysis 3. It performs better than analysis 1.

Analysis 5 frees the non-invariant parameters and holds other parameter exactly invariant. It is the second step in the proposed approach and performs better than all the other analyses.

Although not considered here, a more thorough study of analysis 5 can be done where for each replication the parameters with significant difference from the mean in analysis 2 are freed in analysis 5 and the final estimates of this two-step procedure summarized. This takes into account that in some replications the wrong parameters will be diagnosed as non-invariant.

Other simulation settings with different choices of parameter values are of interest but are not pursued here. For example, data may be generated with many non-invariant parameters and where the non-invariance is in both directions and more in line with BSEM specifications. In such settings, the first step of BSEM analysis may be sufficient and not requiring the freeing of non-invariant parameters.

Table 5: Multiple-group factor analysis simulation: Factor mean and variance estimates (95% coverage) using five different methods

| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | Bias |
|---|---|---|---|---|---|---|---|---|---|---|
| Population values | | | | | | | | | | |
| Mean | 1.000 | 1.000 | 0.600 | 0.400 | 0.000 | 1.000 | 1.000 | 0.600 | 0.400 | |
| Variance | 2.000 | 1.800 | 1.600 | 1.400 | 1.000 | 2.000 | 1.800 | 1.600 | 1.400 | |
| 1. Exact invariance | | | | | | | | | | |
| Mean estimate | 1.027 | 1.104 | 0.680 | 0.444 | 0.041 | 1.015 | 1.099 | 0.631 | 0.428 | 0.052 |
| Mean coverage | 0.980 | 0.840 | 0.800 | 0.950 | 0.920 | 0.900 | 0.830 | 0.940 | 0.970 | |
| Variance estimate | 2.151 | 2.050 | 1.837 | 1.609 | 0.992 | 2.154 | 2.045 | 1.616 | 1.398 | 0.141 |
| Variance coverage | 0.920 | 0.770 | 0.820 | 0.700 | 0.930 | 0.900 | 0.750 | 0.980 | 0.970 | |
| 2a. BSEM, V=0.10 | | | | | | | | | | |
| Mean estimate | 0.832 | 0.904 | 0.562 | 0.354 | 0.029 | 0.823 | 0.900 | 0.518 | 0.346 | 0.088 |
| Mean coverage | 0.990 | 1.000 | 1.000 | 1.000 | 1.000 | 0.980 | 0.980 | 1.000 | 1.000 | |
| Variance estimate | 1.453 | 1.414 | 1.294 | 1.114 | 0.688 | 1.457 | 1.413 | 1.119 | 0.965 | 0.409 |
| Variance coverage | 0.960 | 0.980 | 1.000 | 1.000 | 0.940 | 0.930 | 0.990 | 0.960 | 0.940 | |
| 2b. BSEM, V=0.01 | | | | | | | | | | |
| Mean estimate | 1.040 | 1.124 | 0.698 | 0.445 | 0.040 | 1.028 | 1.119 | 0.647 | 0.432 | 0.064 |
| Mean coverage | 0.990 | 0.900 | 0.870 | 0.980 | 0.970 | 0.950 | 0.910 | 0.970 | 0.980 | |
| Variance estimate | 2.244 | 2.172 | 1.982 | 1.711 | 1.055 | 2.248 | 2.167 | 1.724 | 1.483 | 0.243 |
| Variance coverage | 0.960 | 0.820 | 0.830 | 0.750 | 0.980 | 0.950 | 0.810 | 1.000 | 1.000 | |
| 3. Exact invariance for one anchor item, BSEM V=0.01 for others | | | | | | | | | | |
| Mean estimate | 1.040 | 1.067 | 0.643 | 0.418 | 0.019 | 1.026 | 1.064 | 0.640 | 0.426 | 0.038 |
| Mean coverage | 0.940 | 0.910 | 0.910 | 0.960 | 0.950 | 0.870 | 0.930 | 0.960 | 0.960 | |
| Variance estimate | 2.226 | 2.018 | 1.814 | 1.595 | 1.071 | 2.225 | 2.008 | 1.757 | 1.520 | 0.182 |
| Variance coverage | 0.880 | 0.880 | 0.900 | 0.800 | 0.930 | 0.910 | 0.860 | 0.910 | 0.870 | |
| 4. Freeing non-invariants, BSEM V=0.01 for others | | | | | | | | | | |
| Mean estimate | 0.952 | 0.967 | 0.572 | 0.365 | -0.009 | 0.941 | 0.961 | 0.569 | 0.374 | 0.034 |
| Mean coverage | 0.980 | 0.980 | 0.980 | 0.990 | 0.990 | 0.930 | 0.980 | 0.980 | 1.000 | |
| Variance estimate | 1.894 | 1.705 | 1.521 | 1.342 | 0.954 | 1.898 | 1.701 | 1.546 | 1.337 | 0.078 |
| Variance coverage | 0.980 | 1.000 | 1.000 | 0.960 | 1.000 | 0.990 | 1.000 | 1.000 | 0.990 | |
| 5. Freeing non-invariants, exact invariance for others | | | | | | | | | | |
| Mean estimate | 1.019 | 1.021 | 0.610 | 0.406 | -0.002 | 1.015 | 1.020 | 0.610 | 0.410 | 0.013 |
| Mean coverage | 0.960 | 0.950 | 0.940 | 0.950 | 0.950 | 0.940 | 0.950 | 0.950 | 0.950 | |
| Variance estimate | 2.107 | 1.891 | 1.681 | 1.466 | 1.052 | 2.097 | 1.891 | 1.694 | 1.469 | 0.083 |
| Variance coverage | 0.930 | 0.940 | 0.930 | 0.930 | 0.910 | 0.910 | 0.920 | 0.910 | 0.920 | |

# 5 A multiple-group factor analysis application: Math item responses in 40 PISA countries

The two-step BSEM multiple-group approach is here applied to binary items from the PISA (Program for International Student Assessment) survey of 2003. As in Section 7.6 of Fox (2010), a one-factor model is considered for eight mathematics test items administered to a total of 9796 students from 40 countries. The model is a 2-parameter probit IRT model that accommodates country measurement non-invariance for all difficulty (threshold) and discrimination (loading) parameters as well as country-specific factor means and variances.

The analyses use Bayesian estimation with a probit link so that for subject $i$, item $j$, and group $g$ the probability of the binary variable $u_{ijg}$ can be expressed in terms of the factor $\eta$ using the standard normal distribution function $\Phi$ as

$$P(u_{ijg} = 1|\eta_{ig}) = \Phi[-\tau_{jg} + \lambda_{jg}\,\eta_{ig}], \tag{4}$$

where for each group it is assumed conditional independence among the $u$'s conditional on the factor and normality for the factor. The item response model can be expressed in terms of continuous-normal latent response variable $u^*$ underlying the binary outcomes,

$$u^*_{ijg} = \lambda_{jg}\,\eta_{ig} + \epsilon_{ijg}, \tag{5}$$

with the usual probit standardization of unit variance for the residual $\epsilon$. The factor metric can be set by fixing the factor mean in one group at 0 and by fixing either a factor loading at 1 for one item or by fixing the factor residual variance at 1 for one group. For these BSEM analyses, the choice is made to fix the factor mean and variance at 0 and 1 for the 40th group.

A posterior predictive p-value (PPP) is obtained for the fit to the $u^*$ structure as discussed in Asparouhov and Muthén (2010) and Muthén and Asparouhov (2012a). As described in Muthén and Asparouhov (2012a), the PPP is a useful guide when studying the sensitivity of the results to the choice of prior variance. If the prior variance is small relative to the magnitude of non-invariance, PPP will be lower than if the prior variance corresponds better to the magnitude of non-invariance.

For the PISA data a BSEM prior variance of 0.01 for the measurement parameters results in PPP = 0.136, a prior variance of 0.05 results in PPP = 0.449, a prior variance of 0.10 results in PPP = 0.489, and a prior variance of 0.20 results in PPP = 0.493. A prior variance 0.10 implies a prior belief that 95% of the distribution of the non-invariance lies in the range $[-0.62, +0.62]$. Using this prior variance the estimated magnitude of non-invariance is less than 0.55 in absolute value and is therefore within this range. Using prior variance 0.10 in the BSEM analysis, Table 6 shows the PISA countries with significant differences relative to the average across countries. Relatively few of the loadings and thresholds are non-invariant, only 31 out of the 640 parameters. Several countries have no non-invariant parameters and only countries 18 and 27 have more than one non-invariant parameter. Item 8 shows the least degree of non-invariance.

This analysis may be compared to that of Fox (2010) using the same data. Fox (2010) used a random parameter approach to non-invariance where a measurement

Table 6: PISA countries with significant differences relative to the average across countries (prior variance = 0.10)

| Item | Loading | Threshold |
|---|---|---|
| 1 | - | 2, 12, 18, 22, 28, 39 |
| 2 | 15, 35 | 29, 38 |
| 3 | 15 | 23, 34, 35 |
| 4 | - | 12, 27, 40 |
| 5 | 3 | 7, 37 |
| 6 | 3, 33 | 5, 18, 25, 27, 37 |
| 7 | - | 9, 24, 27 |
| 8 | 24 | - |

parameter for different groups is seen as randomly drawn from a population with shared mean and variance and where a significant variance corresponds to non-invariance. This technique finds non-invariance for all item parameters except the threshold for item 8. The analysis does not, however, pinpoint which countries contribute to this non-invariance. For example, as Table 6 shows, there are only five countries who have any non-invariant loadings.

The ordering of factor means across the countries can be compared between the different types of analysis using scatter plots. Figure 2 compares BSEM analysis (X axis) with analysis imposing exact invariance (Y axis). On the whole, the points in the figure align well along a line with only minor deviations, describing a similar ordering of countries. One notable exception is seen in the upper right part of the figure where the analysis under exact invariance ranks the country lower than using BSEM.

Figure 3 compares BSEM results with those of analysis 5, freeing the non-invariant parameters and imposing exact equality for the other parameters. There is no longer a notable deviation from the line as in Figure 2. Also, the two countries which in Figure 2 have zero factor means according to exact invariance analysis and BSEM analysis are separated along the Figure 3 Y axis using the proposed approach so that the two analyses disagree in this regard. According to the simulations, the ordering according to analysis 5 shown on the Y axis in Figure 3 is the most trustworthy.

Figure 2: Estimated factor means for 40 countries: Comparing BSEM analysis (X axis) with analysis imposing exact invariance (Y axis)

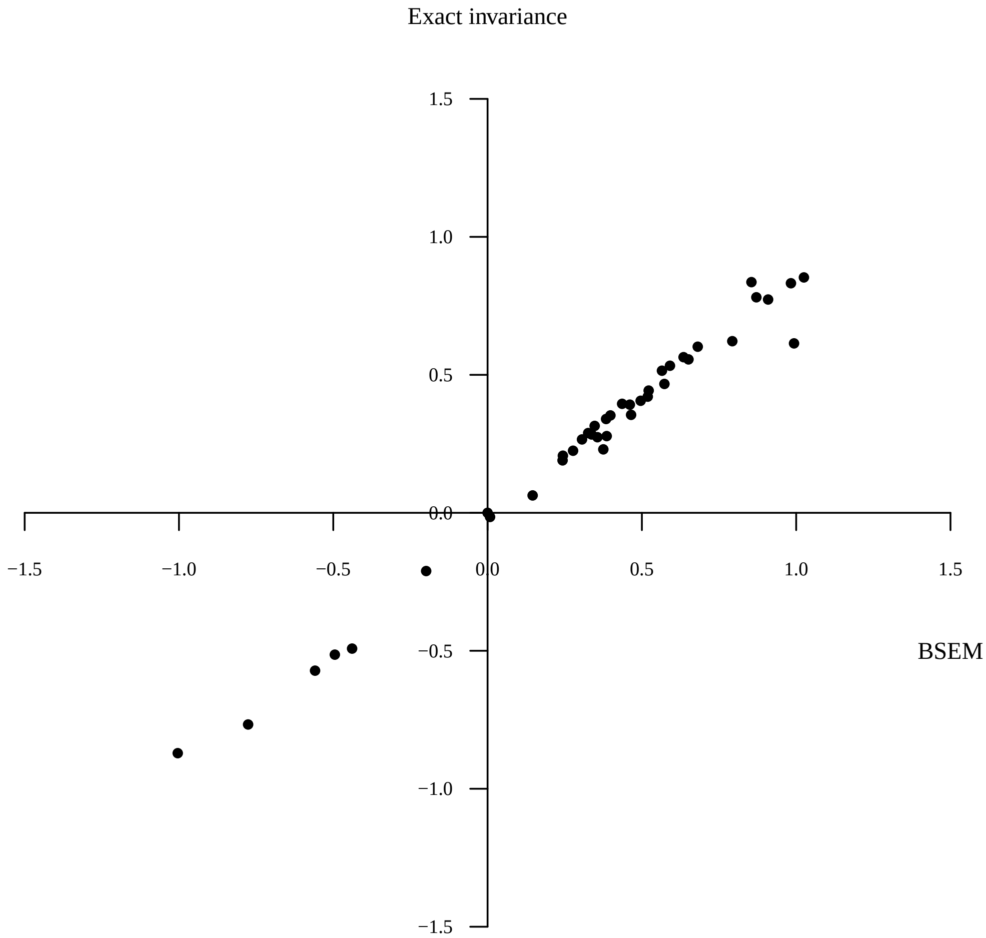
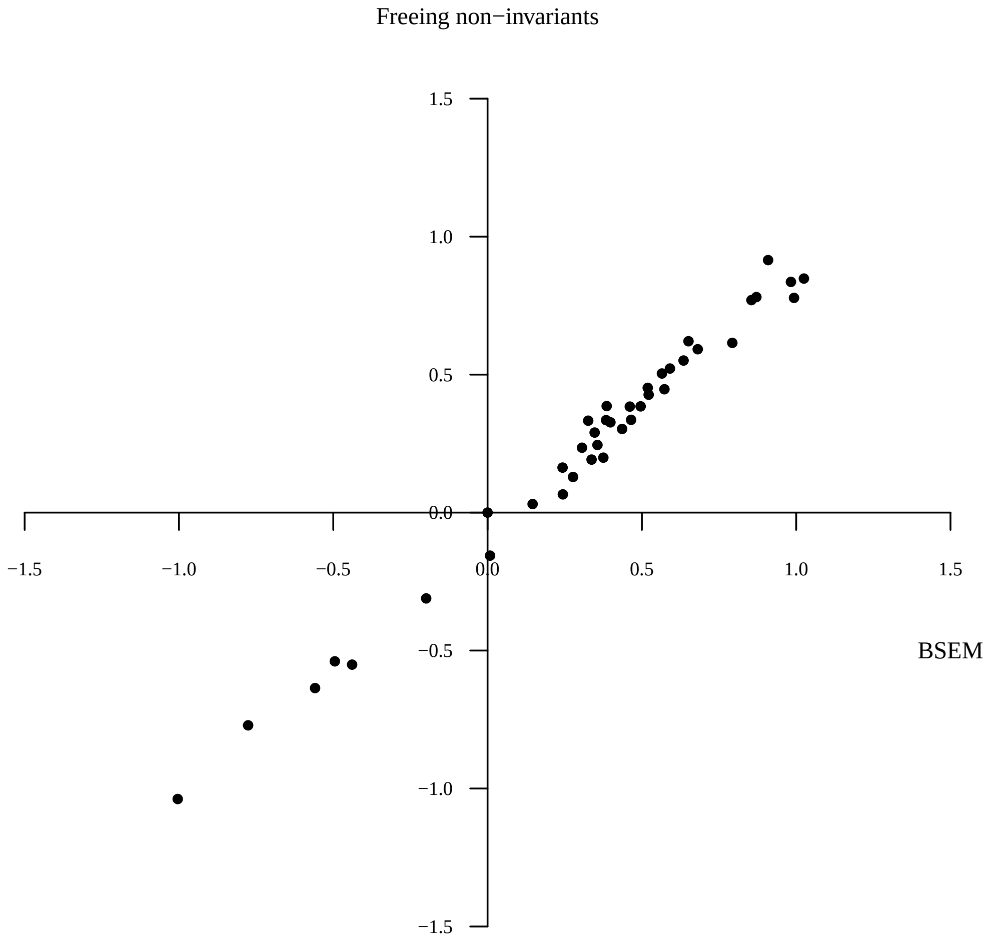


Figure 3: Estimated factor means for 40 countries: Comparing BSEM analysis (X axis) with analysis that frees non-invariant parameters and imposes exact equality for other parameters (Y axis)



## 6 A multiple indicator growth model simulation

Consider a case where there are 10 binary items measuring a single factor at five time points. A linear growth model is specified for the factor. The factor loadings and thresholds are invariant across time with some minor exceptions. The growth model parameterization holds the thresholds equal across time so that the intercept growth factor mean is zero. The focus is on correctly estimating the slope growth factor mean as well as the variances of the intercept and slope growth factors.

The analyses use Bayesian estimation with a probit link so that for subject $i$, item $j$, and time point $t$, the probability of the binary variable $u_{ijt}$ can be expressed in terms of the factor $\eta$, the intercept growth factor $\eta_0$ and the slope growth factor $\eta_1$ as

$$P(u_{itj} = 1 | \eta_{it}, \eta_{0i}, \eta_{1i}) = \Phi[-\tau_{tj} + \lambda_{tj}\ \eta_{it}], \tag{6}$$

$$\eta_{it} = \eta_{0i} + x_t\ \eta_{1i} + \zeta_{it}, \tag{7}$$

with conditional independence among the $u$'s conditional on the factors. The item response model can be expressed in terms of continuous-normal latent response variable $u^*$ underlying the binary outcomes,

$$u^*_{itj} = \lambda_{tj}\ \eta_{it} + \epsilon_{itj}, \tag{8}$$

with the usual probit standardization of unit variance for the residual $\epsilon$. The factor metric can be set either by a factor loading fixed at 1 for one item or by the factor residual variance fixed at 1 for one time point. With binary items, it turns out that the former approach is preferable due to making the Bayes iterations converge faster.

The time-invariant factor loadings are chosen as 1 while the time-invariant thresholds are chosen as 0. Following are the non-invariant loadings and thresholds: $\lambda_{11} = 1.70$, $\lambda_{22} = 1.70$, $\lambda_{33} = 1.35$, $\lambda_{44} = 1.35$, $\lambda_{55} = 1.35$, $\tau_{16} = 0.7$, $\tau_{17} = 0.7$, $\tau_{56} = -0.7$, $\tau_{57} = -0.7$. This means that out of the 100 measurement parameters, only nine are non-invariant, corresponding to a well-developed measurement instrument. Nevertheless, ignoring the non-invariance has biasing effects on key parameters as will be seen. The magnitude of the loading and threshold non-invariance can be related to the standardized metric of each $u^*$ variable, that is, dividing both loadings and thresholds by the $u^*$ standard deviation and multiplying the loadings by the factor standard deviation. Although changing slightly over the time points due to changing factor variances, the loading difference of 0.7 corresponds to about 0.60 in the standardized metric, the loading difference of 0.35 to half of that, and the threshold difference of 0.7 to about 0.35 in the standardized metric.

The variance of the prior for the non-invariance, that is the difference between each pair of loading or threshold parameters, is chosen as 0.10. This implies a prior belief that 95% of the distribution of the (unstandardized) non-invariance lies in the range $[-0.62, +0.62]$, a range which almost covers the 0.7 magnitude non-invariance. Using a prior variance of 0.15 corresponds to the range $[-0.76, +0.76]$ and gives similar results. Using the much smaller prior variance 0.01 corresponds to the narrower range $[-0.20, +0.20]$ and is not large enough to let non-invariance be sufficiently expressed for this example and does not give good BSEM point estimates or coverage. The non-invariance detection is still reasonably good also in this case, however, although indicating too many invariant parameters.

As in the multiple-group simulation, the form of non-invariance simulated in this example is not of the type that is assumed for BSEM approximate invariance. This implies that the alignment will not be perfect and the population values used to generate the data will not be perfectly recovered. It is of interest to see to which degree recovery occurs and what the coverage values are. Most importantly, however, is what the power is to detect non-invariance and how well an analysis with relaxed invariance for non-invariant items performs.

As in the multiple-group simulation, five different types of Bayesian analyses are performed:

1. Exact invariance over time for all loadings and thresholds. It is expected that this analysis gives strongly biased results in this setting

2. BSEM analysis using approximate invariance of all loadings and thresholds across time, assuming ignorance of where the non-invariance resides. This analysis is expected to perform better than analysis 1 and also provide information about non-invariance, that is, parameters that are significantly different from the average over time for the respective item parameter. The analysis is not expected to recover the population values exactly due to the alignment issue

3. Exact invariance for invariant items based on analysis 2. and BSEM approximate invariance for other items. It is unknown if utilizing invariant "anchor" items improves on analysis 2. The invariant items 8-10 are anchor items in the simulation

4. Free non-invariant parameters based on analysis 2 and BSEM approximate invariance for other parameters. It is unknown if freeing the non-invariant parameters improves on analysis 2. The nine non-invariant parameters are freed in the simulation

5. Exact invariance for only the invariant parameters and non-invariant parameters free. This analysis is expected to perform optimally given that this is how the data are generated, assuming that the correct non-invariance is diagnosed in analysis 2

A sample size of $n = 1000$ and 100 replications are used in the three simulation analyses. The approximate BSEM invariance is specified with prior variance 0.10 for both loadings and thresholds.

Appendix Section 9.3 shows the input for the Monte Carlo simulation of the second analysis, the BSEM analysis. A total of 50 items are analyzed, 10 items for each of the 5 time points. The metric of the factor is determined by fixing the loading of the last item at the first time point to 1. This approach has been found to perform better than setting the metric in the factor variance at one time point. As shown in the table, MODEL PRIORS therefore excludes the loadings of the last item so that the DO loop for the DIFF statement only goes to 9 instead of 10. Instead, a MODEL CONSTRAINT section is added to express the differences between the last item's loadings and their average over time points. In MODEL PRIORS the approximate invariance for the last item is specified for time points 2-5 using a prior with mean 1, that is, the same value as the fixed loading of the last item at the first time point.

Table 7 shows the results for the difference at each of the five time points for each of the 10 item's loading and threshold relative to their mean over time points. In parentheses are given the estimates of the power to detect non-invariance. All non-invariant parameters have a power estimate of at least 0.80. These power estimates are bolded. No invariant parameters has a power estimate larger than 0.28. This suggests that BSEM will most often diagnose the correct parameters as non-invariant.

Table 8 shows the results of the five analyses to be discussed. Five key parameters are considered. First the mean of the slope growth factor (the mean of the intercept growth factor is zero in this parameterization where the thresholds are held equal or approximately equal across time) and the variances of the intercept and slope growth factors. Next, the standardized slope mean is considered to include a scale-free quantity. A final quantity considered is the product of the last item's factor loading and the slope mean which relates to how quickly the last item probability changes over time.

Analysis 1 with exact invariance gives unacceptable results in that the bias and coverage is poor for the slope mean and the loading $\times$ slope mean product. Analysis 2 using BSEM with small-variance prior variance 0.10 gives improved results but due to the alignment issue it still gives a large bias for the slope mean and poor coverage for the loading $\times$ slope mean product. Analysis 3 with exact invariance for invariant items based on analysis 2 while maintaining BSEM approximate invariance for other items does not give uniform improvement over analysis 2. Analysis 4 with free non-invariant parameters based on analysis 2 while maintaining BSEM approximate invariance for other parameters also does not give uniform improvement over analysis 2. Analysis 5 that frees the non-invariant parameters while having exact invariance for the other parameters is the

Table 7: BSEM results for a growth model simulation with 10 binary items at 5 time points: Deviations from the mean (power estimates)

| | Time 1 | Time 2 | Time 3 | Time 4 | Time 5 |
|---|---|---|---|---|---|
| Loadings | | | | | |
| Item 1 | **0.405 (1.00)** | -0.124 (0.26) | -0.102 (0.22) | -0.105 (0.24) | -0.083 (0.08) |
| Item 2 | -0.139 (0.36) | **0.410 (1.00)** | -0.101 (0.19) | -0.087 (0.15) | -0.090 (0.16) |
| Item 3 | -0.090 (0.07) | -0.072 (0.10) | **0.251 (0.84)** | -0.059 (0.08) | -0.040 (0.02) |
| Item 4 | -0.098 (0.18) | -0.077 (0.11) | -0.054 (0.05) | **0.256 (0.90)** | -0.034 (0.04) |
| Item 5 | -0.094 (0.15) | -0.070 (0.07) | -0.054 (0.09) | -0.053 (0.04) | **0.263 (0.89)** |
| Item 6 | -0.060 (0.03) | -0.010 (0.02) | 0.013 (0.02) | 0.011 (0.01) | 0.038 (0.01) |
| Item 7 | -0.061 (0.04) | -0.013 (0.03) | 0.021 (0.01) | 0.011 (0.02) | 0.032 (0.00) |
| Item 8 | -0.037 (0.05) | -0.010 (0.01) | 0.004 (0.07) | 0.006 (0.01) | 0.029 (0.06) |
| Item 9 | -0.039 (0.05) | -0.006 (0.01) | 0.010 (0.02) | 0.008 (0.01) | 0.019 (0.01) |
| Item 10 | -0.052 (0.08) | -0.016 (0.00) | 0.028 (0.04) | 0.013 (0.05) | 0.019 (0.02) |
| Thresholds | | | | | |
| Item 1 | -0.108 (0.23) | -0.036 (0.05) | 0.001 (0.00) | 0.036 (0.04) | 0.106 (0.28) |
| Item 2 | -0.100 (0.22) | -0.027 (0.02) | -0.002 (0.03) | 0.023 (0.04) | 0.104 (0.25) |
| Item 3 | -0.100 (0.22) | -0.040 (0.04) | 0.027 (0.01) | 0.018 (0.03) | 0.092 (0.19) |
| Item 4 | -0.103 (0.24) | -0.053 (0.09) | 0.007 (0.00) | 0.041 (0.02) | 0.106 (0.27) |
| Item 5 | -0.108 (0.25) | -0.044 (0.10) | -0.003 (0.04) | 0.018 (0.02) | 0.135 (0.25) |
| Item 6 | **0.555 (1.00)** | -0.042 (0.03) | -0.002 (0.04) | 0.033 (0.03) | **-0.546 (1.00)** |
| Item 7 | **0.549 (1.00)** | -0.039 (0.05) | 0.012 (0.03) | 0.026 (0.04) | **-0.549 (1.00)** |
| Item 8 | -0.098 (0.25) | -0.027 (0.04) | -0.002 (0.06) | 0.026 (0.04) | 0.099 (0.21) |
| Item 9 | -0.092 (0.23) | -0.032 (0.04) | 0.002 (0.02) | 0.015 (0.01) | 0.105 (0.25) |
| Item 10 | -0.102 (0.24) | -0.044 (0.07) | 0.013 (0.02) | 0.033 (0.04) | 0.098 (0.17) |

preferred analysis and gives good bias results and coverage for all five quantities. The remaining bias seen in analysis 5 is due to sampling error and improves with a larger sample size.

As for the multiple-group simulation study, the outcome of the growth simulation suggests that the proposed two-step approach works well where the BSEM analysis (analysis 2) is followed by analysis 5. The strong power results shown in Table 7 are key to the good performance.

Table 8: A growth model simulation with 10 binary items at 5 time points: Mean estimates (95% coverage) for five analyses

| Model | Slope Mean | Intercept Variance | Slope Variance | Standard'd Slope Mean | Loading × Slope Mean |
|---|---|---|---|---|---|
| Population values | | | | | |
| | 0.5 | 0.5 | 0.2 | 1.118 | 0.5 |
| 1. Exact invariance | | | | | |
| | 0.595 | 0.538 | 0.217 | 1.310 | 0.595 |
| | (0.300) | (0.940) | (0.950) | (0.870) | (0.300) |
| 2. BSEM, V=0.10 | | | | | |
| | 0.553 | 0.516 | 0.187 | 1.314 | 0.594 |
| | (0.920) | (0.950) | (0.960) | (0.950) | (0.800) |
| 3. Exact invariance for anchor items, BSEM V=0.10 for others | | | | | |
| | 0.516 | 0.544 | 0.216 | 1.139 | 0.516 |
| | (0.890) | (0.930) | (0.940) | (0.970) | (0.890) |
| 4. Freeing non-invariants, BSEM V=0.10 for others | | | | | |
| | 0.439 | 0.457 | 0.171 | 1.096 | 0.491 |
| | (0.930) | (0.930) | (0.940) | (0.990) | (0.990) |
| 5. Freeing non-invariants, exact invariance for others | | | | | |
| | 0.504 | 0.537 | 0.224 | 1.094 | 0.504 |
| | (0.940) | (0.930) | (0.920) | (0.950) | (0.940) |

# 7 A growth model application

As a growth model example, consider a teacher-rated measurement instrument capturing aggressive-disruptive behavior among a sample of U.S. students in Baltimore public schools (Ialongo et al., 1999). A total of 1174 boys are observed in 41 classrooms at eight different time points, from the Fall of Grade 1 to the Spring of Grade 7. Nine items scored as 1 (Almost Never) through 6 (Almost Always) are considered. The item distribution is very skewed with a high percentage in the Almost Never category. The items are therefore dichotomized into Almost Never versus the other categories combined. A multiple-indicator growth model is considered for the eight time points. With maximum-likelihood estimation, this requires eight-dimensional numerical integration which is feasible with Monte Carlo integration, for instance using 5000 integration points. Maximum-likelihood, however, gives heavy computations with this many latent variable dimensions, has difficulty in obtaining precise likelihood values (Asparouhov & Muthén, 2012a), does not automatically produce indices of non-invariance, and does not scale to more time points as easily as Bayesian analysis. The multilevel aspect of students within classrooms is ignored in this analysis, but leads to even more latent variable dimensions; see Muthén and Asparouhov (2012b).

Appendix Section 9.4 shows the input for BSEM analysis with approximate measurement invariance across the eight time points. A quadratic growth model is used with zero quadratic growth factor variance. The aim of the BSEM analysis is to detect non-invariance in the measurement parameters. In line with the simulations of the previous section, the metric is set in the factor loading for

Table 9: Time points with significant differences relative to the average across time points for the aggression example (prior variance = 0.10)

| Item | Loading | Threshold |
|---|---|---|
| stubborn | 1, 3, 8 | 1, 2, 3, 6, 8 |
| breaks rules | 1, 2, 8 | 1, 5, 8 |
| harms others | 1, 7, 8 | 2, 8 |
| breaks things | 1, 7 | 2, 3, 8 |
| yells | 1, 3 | 2, 4, 6, 8 |
| takes property | 1, 8 | 1, 2, 5, 6, 7 |
| fights | 2, 8 | 1, 3, 4 |
| lies | 2, 8 | - |
| teases | 7, 8 | 1, 4, 6, 8 |

the first time point and the variance for the prior is chosen as 0.10. Variance choices ranging from 0.01 to 0.15 result in similar indications of non-invariant parameters, although somewhat fewer significant parameters are obtained with smaller variance values.

Table 9 shows the time points with significant differences relative to the average across time points for both loadings and thresholds. Among the 144 possible instances of non-invariance, 50 are detected as significant in this BSEM analysis. No item is invariant across all time point with respect to both loadings and thresholds. Time points one and eight show the largest number of non-invariant parameters which may be natural given that these time points span the time range. The table is informative about the invariance properties of the items. Of the nine items, the items stubborn and take property show the largest degree of non-invariance, while the item lies shows the least degree of non-invariance.

Table 10 shows the growth factor estimates using the three analyses: Exact

Table 10: Estimates (SEs) from three analyses of aggression example: Exact invariance, BSEM, and freeing non-invariants

| Analysis | PPP | Mean s | Mean q | Var i | Var s |
|---|---|---|---|---|---|
| Exact | 0.000 | 0.062* | -0.007 | 1.847* | 0.023* |
| invariance | | (0.025) | (0.004) | (0.143) | (0.005) |
| BSEM | 0.095 | 0.053 | -0.006 | 1.960* | 0.060* |
| | | (0.052) | (0.009) | (0.198) | (0.013) |
| Freeing | 0.089 | 0.150* | -0.017 | 7.392* | 0.258* |
| non-invariants | | (0.065) | (0.011) | (1.437) | (0.063) |

invariance, BSEM, and exact invariance except freeing non-invariants (analyses 1, 2, and 5 in Section 6). The results are quite different among the three analyses with a much steeper linear increase using the third analysis. The third analysis was found to be the better performing in the simulation studies of Section 6.

# 8 Conclusions

The multiple-group and multiple-time point simulations show that the BSEM approach with approximate measurement invariance is capable of detecting non-invariant parameters. It is a convenient way to identify the specific combination of item and group/time point that gives rise to the non-invariance. The simulations also show that the proposed two-step procedure of BSEM followed by a regular Bayes analysis that frees the non-invariant parameters works well. The procedure easily scales to many groups and time points, as well as many latent variable dimensions.

The proposed BSEM approach has the advantage that it can also be combined with the types of BSEM applications discussed in Muthén and Asparouhov (2012a). For example, small-variance priors can be used to accommodate small cross-loadings in multifactorial confirmatory factor analysis models. This flexibility may be necessary in for example cross-cultural studies where a confirmatory factor analysis model with exactly zero cross-loadings is often unrealistic.

# References


Asparouhov, T. & Muthén, B. (2010). Bayesian analysis of latent variable models using Mplus. Technical Report. Version 4. www.statmodel.com.

Asparouhov, T. & Muthén, B. (2012a). Comparison of computational methods for high-dimensional item factor analysis. Submitted for publication. www.statmodel.com.

Asparouhov, T. & Muthén, B. (2012b). Multiple group confirmatory and exploratory factor analysis. Mplus Web Notes: No. 18. www.statmodel.com

Bou, J.C. & Satorra, A. (2010). A multigroup structural equation approach: A demonstration by testing variation of firm profitability across EU samples. Organizational Research Methods, October 2010; vol. 13, 4: pp. 738-766., first published on January 26, 2010.

Cai, L., Yang, J.S. & Hansen, M. (2011). Generalized full-information item bifactor analysis. Psychological Methods, 16, 221-248.

Davidov, E., Schmidt, P., & Billiet, J. (2011). Cross-cultural analysis. Methods and applications. New York: Routledge.

De Jong, M.G., Steenkamp, J.-B.E.M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hi- erarchical IRT model. Journal of Consumer Research, 34, 260-278.

Fox, J.P. (2010). Bayesian item response theory. New York: Springer.

Golay, P., Reverte, I., Rossier, J., Favez, N. & Lecerf, T. (2012). Further insights on the French WISC-IV factor structure through Bayesian structural equation modeling. Psychological Assessment. Advance online publication. DOI: 10.1037/a0030676

Kim, E.S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. Structural Equation Modeling, 18, 212-228.

Lee, J., Little, T.D. & Preacher, K.J. (2010). Methodological issues in using structural equation models for testing differential item functioning. In Davidov, E., Schmidt, P., & Billiet, J. (eds.), pp. 55-84. Cross-cultural analysis. Methods and applications. New York: Routledge.

Muthén, B. & Asparouhov, T. (2012a). Bayesian SEM: A more flexible representation of substantive theory. Psychological Methods, 17, 313-335.

Muthén, B. & Asparouhov, T. (2012b). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-timepoint example. In Linden & Hambleton (2013). Handbook of item response theory: Models, statistical tools, and applications. www.statmodel.com.

Muthén, L. & Muthén, B. (1998-2012). Mplus User's Guide. Version 7. Los Angeles, CA: Muthén & Muthén.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. Journal of Applied Psychology, 91(6), 12921306.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67113). Hillsdale, NJ: Erlbaum.

# 9 Appendix

## 9.1 Using Mplus to specify priors for parameter differences

When choosing Bayesian priors, parameters are typically specified as independent, but can also be allowed to covary. In Mplus the covariation can be accomplished using the COVARIANCE option in MODEL PRIORS. Following is a regression example:

MODEL:
y ON x1(a)
x2(b);
MODEL PRIORS:
a $\sim$ N(10, 4);
b $\sim$ N(6, 1);
COV(a,b)=0.5;

which says that the prior bivariate distribution of a and b has a covariance of 0.5, which translates to a correlation of 0.25:

$$\frac{0.5}{\sqrt{(4)}\,\sqrt{(1)}} = 0.25$$

The COVARIANCE option can also be used to specify small differences between parameters. Note that

$$V(a-b) = V(a) + V(b) - 2\,Cov(a,b), \tag{9}$$

so that if a and b have non-informative priors with $V(a) = V(b) = 1000$, using $Cov(a,b) = 999.995$ gives $V(a-b) = 0.01$. With a normal distribution, this means that the difference has a 95% chance of being between $-0.196$ and $+0.196$, that is, in a small range around the zero mean. This is specified as follows in Mplus:

a ~ N(0, 1000);

b ~ N(0, 1000);

COV(a, b)=999.995;

The Mplus DIFF option is used with MODEL PRIORS to simplify specifying differences between parameters. The above example can be specified as:

DIFF(a, b) ~ N(0, 0.01);

DO DIFF is used to express parameter differences between large sets of parameters and groups/timepoints. Consider the example of group differences for the loadings on a factor measured by 4 variables in 3 groups. Let lamjk denote a factor loading for group j and variable k. Table 11 shows the Mplus input for this analysis. Here the MODEL PRIORS statement

DO(1,4) DIFF(lam1# – lam3#) ~ N(0,0.01);

results in approximate invariance across the 3 groups, so that for the first variable lam11 ≈ lam21, lam11 ≈ lam31, lam21 ≈ lam31, etc. for the lam parameters for variables 2-4.

Table 11: DO DIFF input for a 1-factor model with 4 indicators in 3 groups

| | |
|---|---|
| ANALYSIS: | TYPE = MIXTURE; ! 3 classes corresponding to 3 groups |
| | ESTIMATOR = BAYES; |
| | PROCESSORS = 2; |
| | MODEL = ALLFREE; |
| MODEL: | %OVERALL% |
| | f BY y1-y4* (lam#_1 – lam#_4); |
| | ! the above gives labels for all 3 groups (group is #) |
| MODEL PRIORS: | |
| | DO(1,4) DIFF(lam1_# – lam3_#)~N(0,0.01); |

The DO DIFF approach is used in Table 11 to analyze multiple groups, which for Bayes is carried out using KNOWNCLASS (not shown) and TYPE=MIXTURE. The ANALYSIS specification MODEL=ALLFREE lets parameter arrays be different across groups. The auto labeling statement

f BY y1-y4* (lam#_1 – lam#_4);

saves an increasingly large amount of typing with an increasing number of groups.

## 9.2 Mplus input for multiple-group simulation

Table 12: caption

```
TITLE:            This is an example of BSEM analysis of a multiple-group
                  model with approximate measurement invariance for a
                  single-factor model in 10 groups
MONTECARLO:
                  NAMES = y1-y6 u;
                  GENERATE = u (9);
                  CATEGORICAL = u;
                  GENCLASSES = c(10);
                  CLASSES = c(10);
                  NOBSERVATIONS = 5000;
                  NREPS = 100;
ANALYSIS:
                  TYPE = MIXTURE;
                  ESTIMATOR = BAYES;
                  PROCESSORS = 2;
                  BITERATIONS = (10000);
MODEL POPULATION:
                  same as for MODEL
MODEL:
                  %OVERALL%
                  f1 BY y1*1 y2*.7 y3*.5
                  y4*1 y5*.7 y6*.5;
                  [y1-y6*0];
                  [f1*1];
                  f1*2;
                  y1-y6*.5;
                  %c#1%
                  [u$1@15 u$2@16 u$3@17 u$4@18 u$5@19 u$6@20 u$7@21
                  u$8@22 u$9@23];
                  f1 BY y1*1 y2*.7 y3*.5 y4*1 y5*.7 y6*.5 (lam1_1-lam1_6);
                  [y1*0 y2*.5 y3*1 y4*1.5 y5*2 y6*2.5] (nu1_1-nu1_6);
                  [f1*1];
                  f1*2;
                  y1-y6*.5;
                  %c#2%
                  [u$1@-15 u$2@16 u$3@17 u$4@18 u$5@19 u$6@20 u$7@21
                  u$8@22 u$9@23];
                  f1 BY y1*1 y2*.7 y3*.7 y4*1 y5*.7 y6*.5 (lam2_1-lam2_6);
                  [y1*0 y2*.5 y3*1 y4*1.5 y5*2 y6*2.7] (nu2_1-nu2_6);
                  [f1*1];
                  f1*1.8;
                  y1-y6*.5;
```

```
%c#3%
[u$1@-16 u$2@-15 u$3@17 u$4@18 u$5@19 u$6@20 u$7@21
u$8@22 u$9@23];
f1 BY y1*1 y2*.9 y3*.5 y4*1 y5*.7 y6*.5 (lam3_1-lam3_6);
[y1*0 y2*.5 y3*1 y4*1.5 y5*2.2 y6*2.5] (nu3_1-nu3_6);
[f1*0.6];
f1*1.6;
y1-y6*.5;
%c#4%
[u$1@-17 u$2@-16 u$3@-15 u$4@18 u$5@19 u$6@20 u$7@21
u$8@22 u$9@23];
f1 BY y1*1 y2*.7 y3*.5 y4*1 y5*.7 y6*.7 (lam4_1-lam4_6);
[y1*0 y2*.5 y3*1.2 y4*1.5 y5*2 y6*2.5] (nu4_1-nu4_6);
[f1*0.4];
f1*1.4;
y1-y6*.5;
%c#5%
[u$1@-18 u$2@-17 u$3@-16 u$4@-15 u$5@19 u$6@20
u$7@21 u$8@22 u$9@23];
f1 BY y1*1 y2*.7 y3*.5 y4*1 y5*.5 y6*.5 (lam5_1-lam5_6);
[y1*0 y2*.7 y3*1 y4*1.5 y5*2 y6*2.5] (nu5_1-nu5_6);
[f1*0];
f1*1;
y1-y6*.5;
%c#6%
[u$1@-18 u$2@-17 u$3@-16 u$4@-15 u$5@-14 u$6@20
u$7@21 u$8@22 u$9@23];
f1 BY y1*1 y2*.7 y3*.5 y4*1 y5*.7 y6*.5 (lam6_1-lam6_6);
[y1*0 y2*.5 y3*1 y4*1.5 y5*2 y6*2.5] (nu6_1-nu6_6);
[f1*1];
f1*2;
y1-y6*.5;
%c#7%
[u$1@-18 u$2@-17 u$3@-16 u$4@-15 u$5@-14 u$6@-13
u$7@21 u$8@22 u$9@23];
f1 BY y1*1 y2*.7 y3*.7 y4*1 y5*.7 y6*.5 (lam7_1-lam7_6);
[y1*0 y2*.5 y3*1 y4*1.5 y5*2 y6*2.7] (nu7_1-nu7_6);
[f1*1];
f1*1.8;
y1-y6*.5;
```

```
                   %c#8%
                   [u$1@-18 u$2@-17 u$3@-16 u$4@-15 u$5@-14 u$6@-13
                   u$7@-12 u$8@22 u$9@23];
                   f1 BY y1*1 y2*.5 y3*.5 y4*1 y5*.7 y6*.5 (lam8_1-lam8_6);
                   [y1*0 y2*.5 y3*1 y4*1.5 y5*2.2 y6*2.5] (nu8_1-nu8_6);
                   [f1*0.6];
                   f1*1.6;
                   y1-y6*.5;
                   %c#9%
                   [u$1@-18 u$2@-17 u$3@-16 u$4@-15 u$5@-14 u$6@-13
                   u$7@-12 u$8@-11 u$9@23];
                   f1 BY y1*1 y2*.7 y3*.5 y4*1 y5*.5 y6*.5 (lam9_1-lam9_6);
                   [y1*0 y2*.5 y3*1.2 y4*1.5 y5*2 y6*2.5] (nu9_1-nu9_6);
                   [f1*0.4];
                   f1*1.4;
                   y1-y6*.5;
                   %c#10%
                   [u$1@-18 u$2@-17 u$3@-16 u$4@-15 u$5@-14 u$6@-13
                   u$7@-12 u$8@-11 u$9@-10];
                   f1 BY y1*1 y2*.7 y3*.5 y4*1 y5*.7 y6*.5 (lam10_1-lam10_6);
                   [y1*0 y2*.5 y3*1 y4*1.5 y5*2 y6*2.5] (nu10_1-nu10_6);
                   [f1@0];
                   f1@1;
                   y1-y6*.5;
MODEL PRIORS:
                   DO(1,6) DIFF(lam1_#-lam10_#)~N(0,0.10);
                   DO(1,6) DIFF(nu1_#-nu10_#)~N(0,0.10);
OUTPUT:
                   TECH8 TECH9;
```

## 9.3 Mplus input for growth model simulation

## 9.4 Mplus input for growth modeling of aggression items

Table 13: A BSEM growth model simulation with 10 binary items at 5 time points

| | |
|---|---|
| TITLE: | This is an example of BSEM analysis of a growth model with |
| | approximate measurement invariance. |
| MONTECARLO: | NAMES = u11-u60; |
| | GENERATE = u11-u60 (1); |
| | CATEGORICAL = u11-u60; |
| | NOBSERVATIONS = 1000; |
| | NREPS = 100; |
| MODEL | |
| POPULATION: | Same as for MODEL |
| ANALYSIS: | ESTIMATOR = BAYES; |
| | PROCESSORS = 2; |
| | BITERATIONS = 75000 (20000); |
| MODEL: | f1 BY u11*1.70 u12-u19*1 (lam1_1-lam1_9) |
| | u20@1; |
| | f2 BY u21*1 u22*1.70 u23-u30*1 (lam2_1-lam2_10); |
| | f3 BY u31-u32*1 u33*1.35 u34-u40*1 (lam3_1-lam3_10); |
| | f4 BY u41-u43*1 u44*1.35 u45-u50*1 (lam4_1-lam4_10); |
| | f5 BY u51-u54*1 u55*1.35 u56-u60*1 (lam5_1-lam5_10); |
| | [u11$1-u15$1*0 u16$1-u17$1*.7 u18$1-u20$1*0] (tau1_1-tau1_10); |
| | [u21$1-u30$1*0] (tau2_1-tau2_10); |
| | [u31$1-u40$1*0] (tau3_1-tau3_10); |
| | [u41$1-u50$1*0] (tau4_1-tau4_10); |
| | [u51$1-u55$1*0 u56$1-u57$1*-.7 u58$1-u60$1*0] (tau5_1-tau5_10); |
| | [f1-f5@0]; |
| | f1*3 f2*2.5 f3*2 f4*1.5 f5*1; |
| | i s \| f1@0 f2@.5 f3@1 f4@1.5 f5@2; |
| | [i@0 s*.5]; |
| | i*.5 s*.2; i WITH s*.1; |
| MODEL PRIORS: | DO(1,9) DIFF(lam1_#-lam5_#)~N(0,0.10); |
| | DO(1,9) DIFF(tau1_#-tau5_#)~N(0,0.10); |
| | lam2_10~N(1,0.10); |
| | lam3_10~N(1,0.10); |
| | lam4_10~N(1,0.10); |
| | lam5_10~N(1,0.10); |
| MODEL | |
| CONSTRAINT: | NEW(lam1_10*1 ave*1 diff1-diff5*0); |
| | lam1_10 = 1; |
| | ave = (lam1_10+lam2_10+lam3_10+lam4_10+lam5_10)/5; |
| | DO(1,5) diff#=lam#_10-ave; |
| OUTPUT: | TECH8 TECH9; |

Table 14: A growth model of aggressive-disruptive behavior with 9 items measured at 8 time points

| | |
|---|---|
| ANALYSIS: | ESTIMATOR = BAYES; |
| | PROCESSORS = 2; |
| | BITERATIONS = 100000 (50000); |
| MODEL: | f1f BY stub1f@1 |
| | bkrule1f-tease1f* (lam12-lam19); |
| | f1s BY stub1s-tease1s* (lam21-lam29); |
| | f2s BY stub2s-tease2s* (lam31-lam39); |
| | f3s BY stub3s-tease3s* (lam41-lam49); |
| | f4s BY stub4s-tease4s* (lam51-lam59); |
| | f5s BY stub5s-tease5s* (lam61-lam69); |
| | f6s BY stub6s-tease6s* (lam71-lam79); |
| | f7s BY stub7s-tease7s* (lam81-lam89); |
| | [stub1f$1-tease1f$1] (tau11-tau19); |
| | [stub1s$1-tease1s$1] (tau21-tau29); |
| | [stub2s$1-tease2s$1] (tau31-tau39); |
| | [stub3s$1-tease3s$1] (tau41-tau49); |
| | [stub4s$1-tease4s$1] (tau51-tau59); |
| | [stub5s$1-tease5s$1] (tau61-tau69); |
| | [stub6s$1-tease6s$1] (tau71-tau79); |
| | [stub7s$1-tease7s$1] (tau81-tau89); |
| | [f1f-f7s@0]; |
| | i s q \| f1f@0 f1s@0.5 f2s@1.5 f3s@2.5 f4s@3.5 f5s@4.5 f6s@5.5 f7s@6.5; |
| | q@0; |
| OUTPUT: | TECH1 TECH8; |
| MODEL PRIORS: | DO(2,9) DIFF(lam1#-lam8#)~N(0,0.10); |
| | DO(1,9) DIFF(tau1#-tau8#)~N(0,0.10); |
| | lam21~N(1,0.10); |
| | lam31~N(1,0.10); |
| | lam41~N(1,0.10); |
| | lam51~N(1,0.10); |
| | lam61~N(1,0.10); |
| | lam71~N(1,0.10); |
| | lam81~N(1,0.10); |
| MODEL CONSTRAINT: | NEW(lam11 ave diff1-diff8); |
| | lam11=1; |
| | ave = (lam11+lam21+lam31+lam41+lam51+lam61+lam71+lam81)/8; |
| | DO(1,8) DIFF#=lam#1-ave; |