5
Measurement Invariance Testing

This chapter introduces a Bayesian approach to assessing measurement invariance (MI). The Bayesian approximate MI approach can be used to examine group differences (or differences across time, see Chapter 8). This process does not assume exact equivalence across groups on a given parameter, as does the traditional ML-based approach. Instead, it implements a difference prior, which is centered at zero and has a narrowed variance hyperparameter. This near-zero (or approximate-zero) prior allows for flexibility and "wiggle" room for parameter differences across groups. The approach can be easily scaled to include multiple groups, and it can be implemented under relatively smaller sample sizes compared to traditional approaches for MI testing. In addition, Bayesian approximate MI may represent substantive interpretations closer t othe original intention than traditional approaches since it can avoid erroneous deletion of items from a scale or unnecessary freeing of constraints in the model. An example comparing the traditional and Bayesian approximate MI approaches is included.

5.1 A Brief Introduction to MI in SEM

A natural extension to the multiple-group approach in Chapter 4 is to examine MI. MI is typically of interest if the measurement model is going to be used to compare across groups (or time, as discussed in Chapter 8). If latent factor scores are to be compared, then the measurement model should hold as equivalent across groups. This equivalence includes all elements of the measurement model such as the factor loadings, intercepts, and factor covariances. If equivalence holds, then this indicates that relationships between the observed item indicators and the latent factors are not conditional on group membership. In other words, the measurement model holds across groups–thus, the groups are measurement *invariant*.

The current chapter highlights MI through a Bayesian perspective, and it is organized as follows. The remainder of this section covers traditional steps for assessing MI. The Bayesian approximate MI process is introduced (Section 5.2), which is followed by the basic model used here (Section 5.3), and the Bayesian form of the model (Section 5.4). An example of

implementation is presented (Section 5.5), as well as a guide for writing up results (Section 5.6). Finally, the chapter concludes with a summary, major take-home points, a map of all notation used throughout the chapter, an annotated bibliography for select resources pertinent to this topic, and sample M*plus* and R code for examples described in this chapter (Section 5.7).

5.1.1 Stages of Traditional MI Testing

This section briefly describes the traditional stages for testing MI in a multiple-group model. These stages are then implemented in an example and compared to the Bayesian approximate MI approach. For a full description of traditional measurement invariance, please see Millsap (2011).

The traditional approach to MI uses ML estimation and some index of model modification to aid in freeing parameters exhibiting non-invariance. There are two main approaches that can be implemented for assessing MI.

The first approach starts with a fully invariant measurement model and frees the invariance one item (or parameter) at a time. A likelihood-ratio chi-square test of invariance (or an index like the comparative fit index) is typically used in an iterative fashion (one restriction freed, the test of invariance is examined, then repeat the process). The second approach employs an opposite strategy in that the process begins with a fully non-invariant measurement model. Then one item (or parameter) is held invariant, the likelihood-ratio chi-square test of invariance is implemented, and then the process continues until the entire model is examined. Regardless of the approach taken, this traditional view of invariance testing works best with only a few groups (or time points, as I will discuss in Chapter 8). The next sections describe some of the main classifications for invariance.

Configural Invariance

Testing for configural invariance is typically the first step in the invariance testing process. This step ensures that the same basic pattern of loadings (free and fixed) exists across groups. If, for example, two groups have different CFA models (e.g., Group 1 is best represented by one factor, and Group 2 is best represented by two factors), then configural invariance does not hold. If this step does not hold, then this indicates the groups are associated with either different latent factors, or these latent factors take on different meanings across the groups. In order for configural invariance to hold, the groups must be associated with the same underlying latent factors.

Metric Invariance

If configural invariance holds, and the same basic latent factors exist across groups, then the next step can be examined in the MI process. The second step for assessing MI is to test for metric (or weak) invariance. This step entails examining the factor loadings by setting them to be invariant across groups. If metric invariance does not hold, then it implies that the strength of the relationship between the observed item indicator and the latent factor is not comparable across groups. If full metric invariance does not hold, then *partial* metric invariance can be tested. Within partial metric invariance, some factor loadings are allowed to be freely estimated (i.e., allowed to differ across groups). This is an iterative process when testing for MI in this step since it is typically done on a loading-by-loading basis (i.e., freeing one loading at a time and assessing for fit).

Scalar Invariance

The next step in the invariance testing process is to assess for scalar (strong) invariance. In this step, intercepts (or thresholds) are constrained across groups. If intercepts are found to be invariant, then it means that if two people (one from each of the two groups) have the same latent factor score, they would also have the same responses for the observed item indicators. After reaching scalar invariance, latent factor differences can be attributed to differences in observed item responses across the groups. In other words, latent factor means can be compared across groups. If full scalar invariance is not met, then *partial* scalar invariance can be examined by freeing certain intercepts in the model to differ across groups.

Unique Variances Invariance

The next step in the process is to test for unique variances (or strict) invariance. This step consists of constraining the error variances tied to observed item indicators to be equal across groups. This step examines whether variability associated with the observed item indicators is equal across groups after accounting for the latent factor. If full invariance is not achieved at this step, then *partial* unique variances invariance can be examined by freeing some error variances across groups.

Factor Variance Invariance

The next step of invariance testing that can be implemented is to assess whether latent factor variances are equal across groups. If invariance is

obtained, then equal variability in the latent factors is assumed across groups.

Factor Mean Invariance

The last step that can be implemented in traditional MI testing is to examine the factor means for invariance. In this step, latent factor means are constrained across groups. If the means are found to be invariant, then this indicates that the latent factor is measured equivalently across groups. This final step is sometimes not included in the traditional invariance testing process and is instead treated as a post-analysis of factor means. Most applications focus on a comparison of latent means, and this can be handled through comparing intercepts in the scalar invariance step described above.

5.1.2 Challenges within Traditional MI Testing

Full MI makes a strict assumption that the model parameters are exactly equivalent across groups. Take for example the CFA pictured in Figure 4.1. If the intercepts for Item 2 are different, but we constrain them to be equal in the MI process, then the difference between these intercepts is (incorrectly) assumed to be zero. A model specification error has been embedded since the parameter difference between the group intercepts is forced to zero when in fact it was non-zero (even if just slightly different from zero, it is still a mis-specification). This constraint may result in a poorly fitting model that prevents the researcher from interpreting model parameters. Even if the difference in the parameters across groups is negligible, setting it equal to zero could still result in a negative impact on fit and interpretation.

A large body of research has shown that this assumption of exact equivalence is not a reasonable assumption to make for measurement models. Many studies have shown that the latent factors, and associated model parameters, are not *exactly* equivalent across groups (see, e.g., Vandenberg & Lance, 2000; Millsap, 2011). In fact, departures in model parameters across groups can be linked to biased estimates when the parameters are held equivalent. One potential solution to this issue is to allow for small departures in model parameters across groups within the MI process. One such way of implementing this technique is through the use of Bayesian approximate MI, which implements near-zero priors akin to those described in Chapter 3.

5.2 Bayesian Approximate MI

In Chapter 3, we saw that Bayesian statistics can allow a certain amount of flexibility in how factor loadings are handled in a CFA. In particular, near-zero priors can be placed on parameters to avoid constraining potentially non-zero parameters to zero.

This same concept can be extended to the case of assessing for MI. In the traditional MI approach, parameters are held to be exactly equal across groups during the different steps of invariance testing. However, this equivalence may be overly restrictive in nature. It is unlikely that the researcher is interested in holding parameters to be *exactly equal* across groups. Instead, the interest is likely in *approximate equivalence*. Bayesian methods allow a more flexible treatment of the restricted MI approach by allowing for small differences in parameter estimates across groups. In other words, a factor loading does not have to be exactly equal across groups for invariance to hold (i.e., when the groups are exactly equal, then the difference in the loadings would be exactly zero across groups). Instead, the difference between the loadings would be *approximately* zero, adding some "wiggle" room or flexibility for what is considered *invariant*. This added flexibility is handled through the use of carefully specified priors placed on all parameter constraints tested throughout the MI steps.

Based on the description of the invariance testing steps above, MI implies that the measurement model, relationship between observed (continuous) item indicators and the latent factors, the factor covariances, and the intercepts are equal across groups. In other words, group membership does not dictate anything about the relationship between items and factors, or how the factors covary.

The flexibility that Bayesian methods afford regarding approximate zeros in the context of estimating measurement models (in the non-group setting) was nicely described in B. O. Muthén and Asparouhov (2012a). These concepts were extended to the case of MI testing in van de Schoot et al. (2013). Essentially, the same premise that was described in Chapter 3 is applied here. Narrow priors centered at zero are used to allow some "wiggle" room around zero. Instead of the difference between parameters being fixed to zero, it is allowed to vary slightly–within the bounds of the specified prior. The researcher would work to determine the optimal variance of the difference prior in order to pinpoint how narrow (or wide) it should be surrounding zero. This feature allows model results to be interpreted even if exact equivalence does not hold for model parameters across groups.

There are many benefits to using the Bayesian approximate MI approach, including: more accurate parameter estimates, the inclusion of small (non-zero) cross-loadings in the measurement model, and better performance than partial MI when parameter differences are small (B. O. Muthén & Asparouhov, 2013; Pokropek, Davidov, & Schmidt, 2019; van de Schoot et al., 2013).

However, there is also one assumption that must be met for application of this method, and it is tied to parameterization indeterminacies–also referred to as an *alignment issue*. B. O. Muthén and Asparouhov (2013) indicated that differences between parameters across groups must be small and non-systematic. The following example is based on one described in B. O. Muthén and Asparouhov (2013). Let's assume that Item 2 from Figure 4.1 is associated with invariance across multiple groups (> 2), with the exception of the last group, where there is a large positive deviation from the other groups. The near-zero prior will pull this deviating parameter toward the average value for that parameter across all groups. In effect, this causes the deviating parameter to be smaller in size, and the remaining invariant parameters are pulled to be larger than the true values. Essentially, the near-zero prior contributes to the model being mis-specified (through the prior) because it did not properly capture the group that deviated substantially from the other groups. When intercepts (or thresholds) or loadings are estimated with bias, then it follows that factor means and factor variances will also be biased. The substantive result is that comparing factor means across groups (which is likely a driving reason for conducting the MI process to begin with) will lead to incorrect interpretations because estimates are biased. The alignment issue can be resolved by combining approximate and partial MI to allow the systematic freeing of parameters that violate this assumption.

Another major issue to discuss within Bayesian approximate MI is the specification of the difference prior (i.e., the near-zero prior). Before delving into that issue, I present the model that will be used in a subsequent example. The presentation of the model will be followed by additional details surrounding the implementation of priors for approximate MI testing.

5.3 The Model and Notation

To illustrate the issues underlying MI and Bayesian approximate MI, consider the same model described in Chapter 4 for multiple-group modeling. The multiple-group CFA incorporating a mean structure analysis can be written out as a simple extension of the basic CFA such that

$$\boldsymbol{x}_{(g)} = \boldsymbol{\tau}_{x(g)} + \boldsymbol{\Lambda}_{x(g)}\boldsymbol{\xi}_{(g)} + \boldsymbol{\delta}_{(g)} \tag{5.1}$$

where the x's represent the observed indicators (e.g., the individual items on a questionnaire), which are linked to latent factors $\boldsymbol{\xi}$ through the factor loading matrix denoted as $\boldsymbol{\Lambda}_{x(g)}$. Akin to Equation 4.1 presented in the last chapter, $\boldsymbol{\tau}$ is a vector of intercepts with dimension $q \times 1$, where q is the number of observed x items. This vector is needed if the latent variable mean differences are to be compared across the groups. The g subscript is placed throughout the model to denote that the parameters are allowed to vary across the $g = 1, \ldots, G$ groups. All observed indicators also correspond to measurement errors $\boldsymbol{\delta}$, which are composed of specific variances and random components of observed indicators $\boldsymbol{x}$. We also assume that $\mathrm{E}(\boldsymbol{\delta} = 0)$, and that all errors are left uncorrelated with the latent factors ($\boldsymbol{\xi}$). The equation can be written out in the following form:

$$\begin{bmatrix} x_{1(g)} \\ x_{2(g)} \\ x_{3(g)} \\ x_{4(g)} \\ x_{5(g)} \\ x_{6(g)} \end{bmatrix} = \begin{bmatrix} \tau_{1(g)} \\ \tau_{2(g)} \\ \tau_{3(g)} \\ \tau_{4(g)} \\ \tau_{5(g)} \\ \tau_{6(g)} \end{bmatrix} + \begin{bmatrix} \lambda_{11(g)} & \lambda_{12(g)} \\ \lambda_{21(g)} & \lambda_{22(g)} \\ \lambda_{31(g)} & \lambda_{32(g)} \\ \lambda_{41(g)} & \lambda_{42(g)} \\ \lambda_{51(g)} & \lambda_{52(g)} \\ \lambda_{61(g)} & \lambda_{62(g)} \end{bmatrix} \begin{bmatrix} \xi_{1(g)} \\ \xi_{2(g)} \end{bmatrix} + \begin{bmatrix} \delta_{1(g)} \\ \delta_{2(g)} \\ \delta_{3(g)} \\ \delta_{4(g)} \\ \delta_{5(g)} \\ \delta_{6(g)} \end{bmatrix}$$

where, for example,

$$\boldsymbol{\Lambda}_{x(g)} = \begin{bmatrix} \lambda_{11(g)} = ? & \lambda_{12(g)} = 0 \\ \lambda_{21(g)} = ? & \lambda_{22(g)} = 0 \\ \lambda_{31(g)} = ? & \lambda_{32(g)} = 0 \\ \lambda_{41(g)} = 0 & \lambda_{42(g)} = ? \\ \lambda_{51(g)} = 0 & \lambda_{52(g)} = ? \\ \lambda_{61(g)} = 0 & \lambda_{62(g)} = ? \end{bmatrix} \tag{5.2}$$

In the case of the multiple-group model, these free parameters (marked with "?") are allowed to differ across groups.

The covariance structure for the CFA model can also be written in terms of multiple groups. The covariance form is as follows:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_{(g)}) = \boldsymbol{\Lambda}_{x(g)}\boldsymbol{\Phi}_{\xi(g)}\boldsymbol{\Lambda}'_{x(g)} + \boldsymbol{\Theta}_{\delta(g)} \tag{5.3}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ represents the covariance matrix of $\boldsymbol{x}$ as represented by $\boldsymbol{\theta}$, but it is allowed to vary across the g groups being examined. $\boldsymbol{\Lambda}_x$ still represents the factor loading matrix, and $\boldsymbol{\Phi}_\xi$ is the covariance matrix for the latent factors ($\boldsymbol{\xi}$). Finally, $\boldsymbol{\Theta}_\delta$ is the covariance matrix for the error terms ($\boldsymbol{\delta}$) linked to the item indicators ($\boldsymbol{x}$).

In a mean-structure situation, the following assumption is typically made:

$$\begin{aligned} E(x_g) &= \boldsymbol{\tau}_{x(g)} + \boldsymbol{\Lambda}_{x(g)} E(\boldsymbol{\xi}_{(g)}) \\ &= \boldsymbol{\tau}_{x(g)} + \boldsymbol{\Lambda}_{x(g)} \boldsymbol{\kappa}_{(g)} \end{aligned} \tag{5.4}$$

where $\boldsymbol{\kappa}_{(g)}$ is a k-dimensional vector of factor means for group g, where k represents the number of factors present in the model. Under the frequentist framework, there is need for an additional constraint to be added to this in order for model identification to be satisfied (Bollen, 1989; Kaplan, 2009). That constraint could be to set $\boldsymbol{\kappa} = \mathbf{0}$, which results in the factor mean estimates being interpreted as differences between the g groups (i.e., removes one restriction and allows factor means to be identified). Given that such identification issues are not necessary to address in the Bayesian framework, this constraint need not be added unless substantively desired.

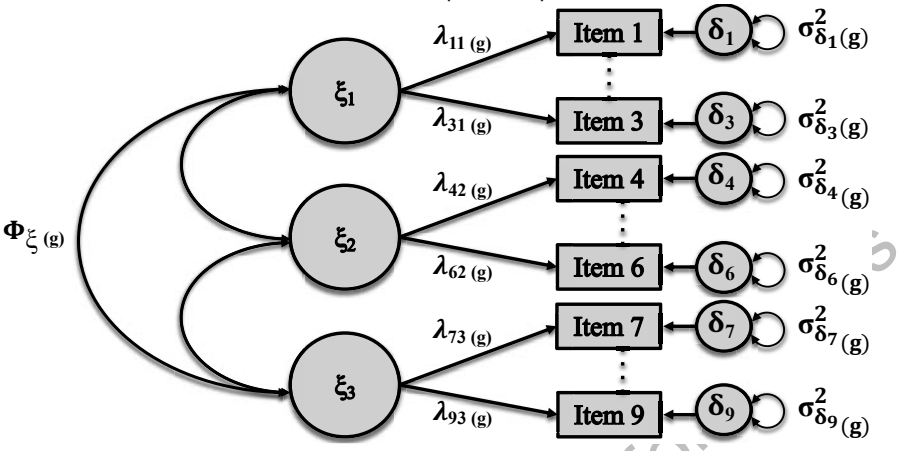
A basic form of the multiple-group CFA can be found in Figure 5.1, which was constructed to represent the example data explored below (and matches the model from Figure 4.1). This model contains three factors ($\boldsymbol{\xi}$), each comprising three items (with loadings contained in the $\boldsymbol{\Lambda}_x$ matrix). The factors are allowed to correlate via $\boldsymbol{\Phi}_\xi$. All item indicators correspond to error terms (δ), with variances denoted as σ^2_δ. In this model, there are no cross-loadings present, and all errors are left uncorrelated (although they need not be).

5.4 Priors within Bayesian Approximate MI

Now that the model has been presented, we can identify several different parameters that may be of interest in the approximate MI process. Depending on the researcher's goals and the level of invariance being examined, near-zero difference priors can be placed on a variety of model parameters (loadings, intercepts, etc.).

The near-zero prior is placed on a difference parameter that is specified for the difference between two groups on a single parameter. Take, for example, a factor loading for Item 2 on a factor. To set up a near-zero prior, the *difference* between the loading for Group 1 and the loading for Group 2 would be of interest. In traditional MI approaches, this difference would be set to zero, making a strict model constraint of exact equivalence between the two groups. However, in Bayesian approximate MI, this difference is allowed to vary (even just slightly so) from zero through the implementation of the near-zero prior. An example of this prior looks as follows:

FIGURE 5.1. The Multiple-Group CFA Model.



$$\lambda_{21}^{(G1)} - \lambda_{21}^{(G2)} \sim \mathcal{N}[0, 0.001] \tag{5.5}$$

where the loading for Factor 1, Item 2 (λ_{21}) is compared across groups (G1 and G2, in this case) by setting up a difference parameter. This parameter is assumed to be distributed normal ($\mathcal{N}$), with a mean hyperparameter of zero and a variance hyperparameter set to some predetermined value specified by the researcher (e.g., 0.001 in this example).

One of the main questions is what these difference priors should look like. In other words: What should the variance hyperparameter be for the difference prior? Given the context of Bayesian approximate MI testing, it is likely that the prior will be centered at zero to represent a parameter mean difference of zero across groups. It follows that the main issue regarding prior specification is tied to the variance hyperparameter.

Asparouhov et al. (2015) proposed a method that implements two different Bayesian fit and comparison indices to aid in selecting the optimal prior variance for the near-zero prior. Specifically, the method uses the deviance information criterion (DIC) and the posterior predictive *p*-value (PP*p*-value) to help select the variance.[1] Asparouhov et al. (2015) suggested estimating several models, each with a different variance hyperparameter specified. One approach would be to start with a relatively small variance hyperparameter value (e.g., 0.001) and then increase this value incrementally for the subsequent models estimated. The decision for which prior setting to use is based on: (1) the speed of convergence, (2) the PP*p*-value,

[1]Given that model fit is such an important element related to SEM in general, an entire chapter on Bayesian model fit related to SEM has been included. Chapter 11 includes much more information on these (and other) indices, as well as examples of implementation.

and (3) the DIC. When model fit differences between models becomes negligible, or reverse in direction (e.g., from a positive to a negative difference), then the prior variance need not be further increased. This approach was further explored by Pokropek et al. (2020). They also recommended using a combination of information from the DIC and PP*p*-value, but they placed more weight on the DIC for decision making based on simulation results.

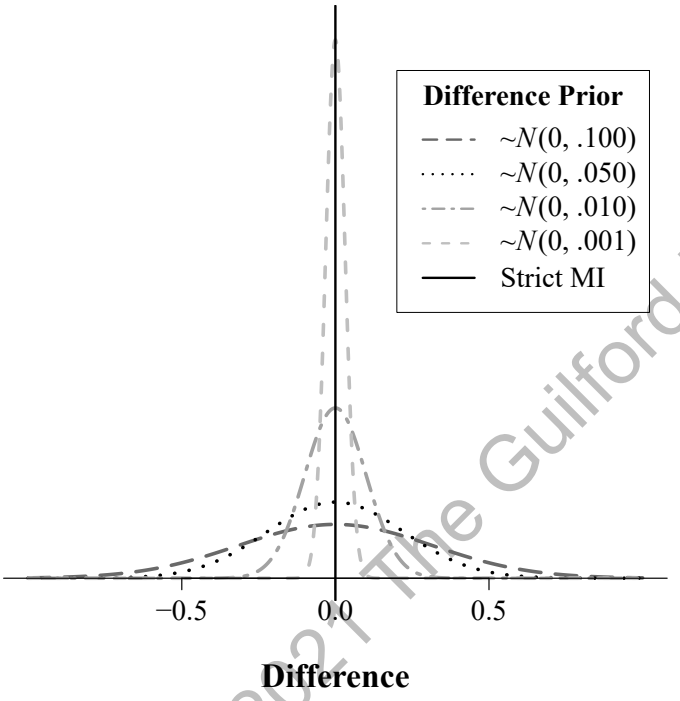
An example of these difference priors can be found in Figure 5.2. The solid line represents zero difference between the model parameters, and this is the strict assumption made in the traditional MI approach. The priors plotted in this figure represent four different options for the approximate MI prior setting. Each of the priors is centered at 0 but contains a different variance hyperparameter, ranging from 0.001 to 0.1. The researcher would decide on the optimal setting to implement and then proceed with the approximate MI process from there. In the next section, I demonstrate the steps needed for implementing Bayesian approximate MI.

5.5 Example: Illustrating Bayesian Approximate MI for School Differences

In this section, I present an example using the Holzinger-Swineford (1939) data, as implemented in Chapter 4. Here I examine a three-factor solution of nine items (three items per factor). The base form of this model can be found in Figure 5.1. The three factors are defined as follows:

- Factor 1: Spatial Ability
  - Item 1: Visual perception
  - Item 2: Cubes
  - Item 3: Lozenges
- Factor 2: Verbal Ability
  - Item 4: Paragraph comprehension
  - Item 5: Sentence completion
  - Item 6: Word meaning
- Factor 3: Task Speed
  - Item 7: Speeded addition
  - Item 8: Speeded counting of dots
  - Item 9: Speeded discrimination straight and curved capitals

FIGURE 5.2. Difference Prior Settings for Approximate MI.



Within the database, there is information about two different schools: Pasteur and Grant-White. This current example explores a multiple-group model of this factor structure for these two schools. The total sample size is n = 301, with 156 students from the Pasteur school (Group 1) and 145 students coming from the Grant-White school (Group 2). The main premise of assessing MI in this context is to examine if and where the two groups differ in the composition of the measurement model. Bayesian approximate MI adds flexibility to this assessment.

To illustrate the Bayesian approximate MI process, I followed these main steps:

1. I estimated invariance models following conventional MI methods (metric, scalar, etc.). These models were estimated via Bayesian estimation, but without the near-zero priors. For pedagogical purposes, I estimated all steps of invariance testing, ignoring whether or not the fit or comparison indices indicated tests should stop.

2. I estimated several versions of Bayesian approximate MI to assess the performance of different near-zero prior settings. Then I selected a final prior setting to use in further analyses.

3. I estimated two additional models (either combining Metric + Approximate MI for intercepts, or Metric + Partial for intercepts).

4. Finally, comparisons can be made for the latent factor means of the second school (Grant-White) across various measurement models.

Table 5.1 provides an overview of the different models estimated in this chapter. The table highlights which model parameters were constrained to be equal across groups or freely estimated. For example, the first model estimated was for configural invariance, and loadings, intercepts, and errors were freely estimated, with factor (co)variances and factor means constrained. The first six rows of this table represent the traditional MI steps, without the use of near-zero priors. The remaining rows represent the approximate MI approach, where near-zero priors were implemented.

TABLE 5.1. Example: Different MI Steps Examined

	Loadings	Intercepts	Errors	Factor (Co)Variances	Factor Means
Configural	Free	Free	Free	Constrain	Constrain
Metric	Constrain	Free	Free	Constrain	Constrain
Scalar	Constrain	Constrain	Free	Constrain	Constrain
Strict	Fixed	Constrain	Constrain	Constrain	Constrain
Factor Variances	Constrain	Constrain	Constrain	Free[a]	Constrain
Factor Means	Fixed	Constrain	Constrain	Constrain	Free[a]
Approximate	Approx.	Approx.	Constrain[b]	Constrain	Free
Approximate + Metric	Fixed	Approx.	Constrain[b]	Constrain	Free
Metric + Partial Scalar	Constrain	Constrain (Item 3 free)	Constrain[b]	Constrain	Free

[a] These models are compared to the "Strict" model to see if freeing the variances or means results in less model misfit (i.e., a lower DIC). [b] It is not possible to specify approximate invariance for error variances.

5.5.1 Results for the Conventional MI Tests

The first six models estimated represent the conventional steps for MI testing. Results for these analyses are presented in the top panel of Table 5.2 on page 183. The columns of results are the DIC, the PP*p*-value, and the 95% CI associated with the difference between the observed and replicated chi-square values. Although Chapter 11 covers these indices in more detail, I will provide a brief description of how to interpret them here. The DIC is an information criterion that is based on Bayesian deviance. It is interpreted comparably to traditional information criteria (e.g., the Bayesian information criterion and the Akaike information criterion). Typically, the model with the lowest DIC value is selected as optimal. However, if the difference between two models is less than 5.0 and the models are substantively different, then the researcher should not make the selection solely based on the lowest DIC (Lee, 2007). In the case of approximate MI, Pokropek et al. (2020) recommended values as low as 1 or 2 can be used for the DIC differences, but it would also be wise to use information from the PP*p*-value as a supplement.

Posterior predictive checks can be used to assess Bayesian model fit. The most common method is to examine the PP*p*-value (Gelman, Meng, & Stern, 1996). This process involves comparing the observed dataset to generated (or replicated) data. During each MCMC iteration, a dataset is generated based on current samples for the model parameters. The generated data are compared to the model implied covariance matrix, resulting in a discrepancy statistic. Then the observed data are compared to the model implied covariance matrix, resulting in a second discrepancy statistic. There are different discrepancy statistics that can be used, but a common one is the chi-square goodness-of-fit statistic. The PP*p*-value represents the proportion of chi-square values derived from the generated data that exceed those obtained from the observed data. PP*p*-values near 0.5 imply adequate fit, whereas values closer to zero indicate that the model does not fit the observed data well.

According to the results in Table 5.2, there is support for metric invariance with a DIC of 7475.87. Given that freeing the factor (co)variances did not result in a decrease from strict invariance, this could be taken as a sign that these parameters can stay fixed across the groups. However, when factor means were freed, the DIC dropped. This indicates that the factor means are not all equal across groups, which matches the substantive results obtained in Chapter 4. Notice that the PP*p*-values indicate that none of these models fit the observed data.

5.5.2 Results for the Bayesian Approximate MI Tests

The first step in the Bayesian approximate MI process is to figure out what the optimal variance hyperparameter is for the near-zero difference priors. To select the specific small variance prior specification, I will follow the iterative procedure outlined in Asparouhov et al. (2015). In addition, this example involves a relatively smaller dataset, so the PPp-value and the DIC should still reflect changes in the prior specification (Hoijtink & van de Schoot, 2018). There is an alternative test that can be used for assessing small variance priors that can outperform these indices when sample sizes are larger. It is called the prior-posterior predictive p-value (PPPP), and I describe this in more detail in Section 11.2.3.

Recall that Figure 5.2 showed four versions of the near-zero prior. The results for the models implementing these priors are in the middle panel of Table 5.2 in the rows labeled "Approximate." The DIC values are comparable for the three largest hyperparameter values, so there would likely not be much of a difference across them. To go with convention, I will select the approximate MI model implementing the near-zero prior of $\mathcal{N}(0, 0.05)$ since it is associated with the lowest DIC value (notice, again, that the PPp-value indicates none of these models fit the observed data well).

Results for the analysis using the approximate MI approach with the near-zero prior of $\mathcal{N}(0, 0.05)$ are presented in Table 5.3. This table presents results for the factor loadings and item intercepts. The first column represents the average estimate across groups, followed by the standard deviation. Then results for deviations from the mean are reported for each group. None of the factor loading estimates deviated significantly from the average factor loading–for either of the two groups. In contrast, there was one item intercept that deviated from its average item intercept across groups, and this was for Item 3 ("Lozenges"). Results indicated that the intercept within each of the groups differed significantly from the average intercept across groups. Item 3 loads onto the "Visualization" factor, and the Group 1 intercept was 2.454 with the Group 2 intercept slightly lower at 2.135. This indicates that the item was slightly "easier" for Group 2 in comparison.

TABLE 5.2. Example: Traditional and Approximate MI Model Comparison

Model	Prior	DIC	PP*p*-value	95% CI Lower	95% CI Upper
Configural		7482.82	0.000	29.43	103.36
Metric		7475.87	0.000	32.30	104.66
Scalar		7538.05	0.000	104.78	174.13
Strict		7536.34	0.000	113.07	179.57
Factor Variances Freed		7542.71	0.000	112.47	181.60
Factor Means Freed		7503.06	0.000	76.01	144.19
Approximate	$\mathcal{N}(0, 0.001)$	7496.72	0.000	67.32	137.35
Approximate	$\mathcal{N}(0, 0.010)$	7479.57	0.000	43.81	115.54
Approximate	$\mathcal{N}(0, 0.050)$	7478.05	0.000	37.56	108.77
Approximate	$\mathcal{N}(0, 0.100)$	7479.53	0.000	37.64	109.18
Metric + Approx	$\mathcal{N}(0, 0.050)$	**7475.03**	0.000	41.70	111.77
Metric + Partial		7497.51	0.000	69.60	137.40

Note. DIC = deviance information criterion; PP*p*-value = posterior predictive *p*-value; CI = 95% credible interval for the difference of observed and replicated χ^2 values. Bold indicates lowest DIC value.

TABLE 5.3. Example: Difference Prior Results.

	Average	*SD*	Deviations from Mean Group 1	Deviations from Mean Group 2
Loadings				
Item 1	0.869	0.084	0.030	−0.030
Item 2	0.519	0.082	0.005	−0.005
Item 3	0.701	0.078	0.037	−0.037
Item 4	0.967	0.056	0.013	−0.013
Item 5	1.060	0.061	0.080	−0.080
Item 6	0.895	0.052	−0.062	0.062
Item 7	0.624	0.078	−0.020	0.020
Item 8	0.728	0.077	−0.046	0.046
Item 9	0.669	0.078	−0.038	0.038
Intercepts				
Item 1	5.005	0.120	−0.057	0.057
Item 2	6.133	0.092	−0.116	0.116
Item 3	2.299	0.104	0.157*	−0.157*
Item 4	2.779	0.111	0.040	−0.040
Item 5	4.055	0.119	−0.053	0.053
Item 6	1.904	0.106	0.017	−0.017
Item 7	4.267	0.096	0.137	−0.137
Item 8	5.633	0.104	−0.063	0.063
Item 9	5.470	0.098	−0.047	0.047

Note. *Indicates a significant difference between the group estimate and the group average.

As a visual aid, Figure 5.3 shows the posterior densities for the Item 3 intercept, where non-invariance was obtained. The posteriors from both groups overlap, but there is also a clear distinction and higher proportion of the densities that do not overlap. In contrast, Figure 5.4 shows posteriors for another item intercept (Item 5, "Sentence Completion"), where invariance was obtained. These densities have a much more pronounced overlap compared to Figure 5.3, highlighting the substantive difference between the intercepts for the two items.

An additional set of models was estimated next. Some authors (see, e.g., van de Schoot et al., 2013) suggest constraining sets of parameters (e.g., all loadings) to equal if approximate MI testing reveals that there are no significant differences at that level. Given the results presented in Table 5.3, I estimated a follow-up model with all factor loadings constrained, while allowing for approximate invariance of the item intercepts (due to Item 3's non-invariance). The overall results for this model are presented in the lower panel of Table 5.2 under the row heading of "Metric + Approx." The DIC obtained for this model was the lowest of all models estimated, even slightly lower than the conventional metric invariance model in the top panel of the table. An additional approach recommended (see, e.g., B. O. Muthén & Asparouhov, 2013) is to use the Bayesian approximate MI findings to specify a partial MI model. In this final model, all factor loadings and item intercepts (with the exception of Item 3) were constrained across groups; error variance and factor variances were also constrained. Results for this model are presented in Table 5.2 under the row heading of "Metric + Partial." The DIC for this model resulted in a value between the conventional metric and scalar models in the upper panel, and it closely matched the approximate MI model with a variance hyperparameter of 0.001. Overall, results indicated that the "Metric + Approx" option is optimal based on the DIC, but none of the models fit according to the PP*p*-value.

5.5.3 Results Comparing Latent Means across Approaches

Finally, it is worthwhile to highlight what some of these findings mean in a substantive sense. Table 5.4 on page 186 presents the Group 2 latent factor means for the three factors (the factor means are fixed to zero in Group 1 during the MI process). The first column of results ("Approximate") is from the model where the factor loadings and item intercepts are modeled as approximately MI through near-zero priors; note that error variances and factor variances were constrained here. The second column of results ("Metric + Approximate") represents the model that combines metric MI with approximate MI for the item intercepts; again, constraining error vari-

ance and factor variances. The third column of results ("Strict") model that assumed strict MI, with constrained loadings, intercepts, error variances, and factor variances. The final column of results ("Metric + Partial") represents the last model estimated, where all factor loadings and item intercepts (with the exception of Item 3) were constrained across groups; error variance and factor variances were also constrained.

Compared to the selected model ("Metric + Approximate"), the other three models tended to overestimate the mean difference (compared to zero) of the "Visualization" factor; this was especially the case for the "Metric + Partial" model. In addition, these same three models slightly underestimated the mean of the "Verbal Intelligence" and "Speed" factors. In terms of substantive conclusions–namely, whether factor means are different across groups–there are no differences in the methods being compared. In other words, regardless of the invariance model selected, we still conclude that the two schools only differ in terms of their verbal intelligence.

FIGURE 5.3. Posterior Densities for Item 3: Lozenges (Showing Non-Invariance).

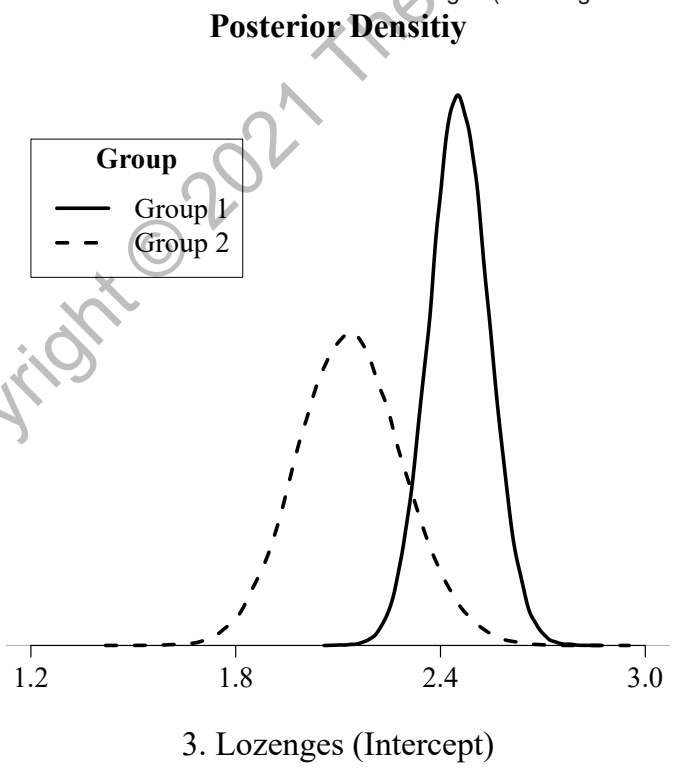


FIGURE 5.4. Posterior Densities for Item 5: Sentence Completion (Showing Invariance).

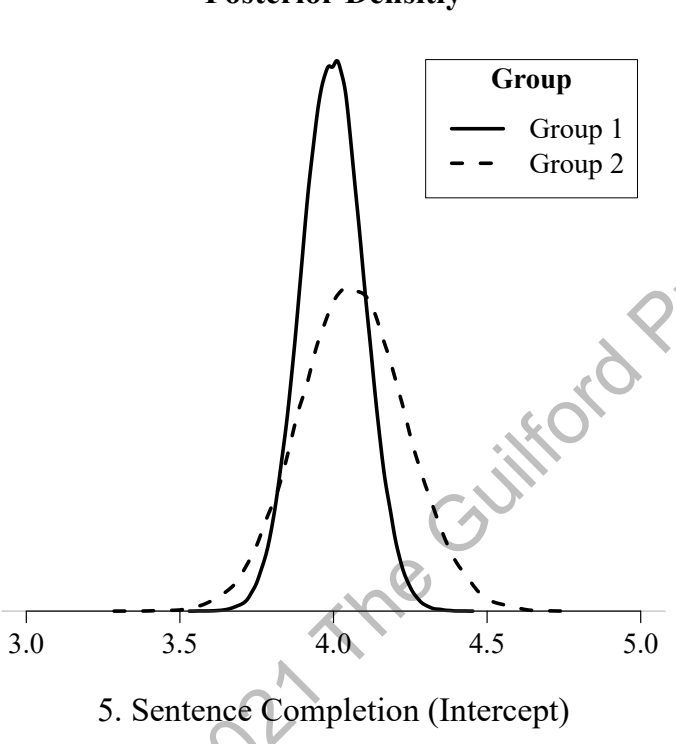


TABLE 5.4. Example: Latent Factor Mean Estimates for Group 2

	Approximate	Metric + Approximate	Strict (Fixed Factor Variance) + Free Means	Metric + Partial
Visualization	−0.172 (0.24)	−0.152 (0.23)	−0.170 (0.15)	−0.302 (0.16)
Verbal Intelligence	0.612 (0.19)*	0.616 (0.18)*	0.603 (0.13)*	0.604 (0.13)*
Speed	−0.282 (0.23)	−0.305 (0.24)	−0.271 (0.14)	−0.272 (0.14)

Note. Values in parentheses are standard deviations. *Indicates significant group difference.

5.6 How to Write Up Bayesian Approximate MI Results

In the current study, we were specifically interested in whether there were school differences between the measurement model illustrated in Figure 5.1. We believe identifying potential differences will help us to better un-

derstand the disparities between...[*Authors could go on to explain location, race/ethnicity, income, and so forth. Factors such as these would likely be the driving reason for being interested in MI.*] In order to assess these differences, we set up a multiple-group CFA and made group comparisons across schools (School 1: n = 156, and School 2: n = 145).

5.6.1 Hypothetical Data Analysis Plan

[*A data analysis plan should be constructed prior to data analysis. In cases in which data are collected (e.g., as opposed to secondary data analysis situations), the data analysis plan should be in place prior to data collection. The goal of the plan is to solidify the variables, model, and priors that will be examined at the analysis stage.*]

We are interested in examining differences across two schools regarding ability level, with the specific goal to test for measurement invariance across groups. The schools represent different areas of the district that are of interest because of funding differences. [*Go on to describe the rationale underlying the groups that are going to be compared.*] We plan to collect data from School A (representing a lower-funded school) and School B (representing a school from the highest tier of funding) using the following data collection process. [*Details about the selection of classrooms and children should be included here, as well as the target number of children to collect data from. Additional justifications or details may be provided in the case of secondary data analysis. For primary data collection situations, the population of interest should be thoroughly described.*]

Ability will be defined using the ability scale described in Author et al. (20xx). This scale includes nine items that are theorized to form three factors of: Spatial Ability, Verbal Ability, and Speed. [*Include more detail as to why the scale was selected, as well as why these specific factors are of substantive interest in terms of the groups being examined.*] In order to compare the two schools based on these ability types, we are proposing a Bayesian approximate measurement invariance process. Measurement invariance will allow us to examine whether there are any measurement model differences between School A and School B regarding the three latent factors proposed above.

The Bayesian approach will allow for *approximate* equivalence rather than *strict* equivalence through the use of near-zero priors. This approach was described as a more flexible treatment for assessing measurement invariance in ability by Author et al. (20xx). [*Next, go through and describe all of the priors that will be implemented, making sure to provide details for how hyperparameters will be specifically defined.*] The analysis plan has been preregistered at the following site: [*include link*].

5.6.2 Hypothetical Analytic Procedure

For all stages of MI testing, we used the three-factor CFA pictured in Figure 5.1. To identify this model, the first factor loading for each factor was set to 1.0. Factors were allowed to correlate freely. Prior to implementing the Bayesian approximate MI process, we examined the model across groups using the traditional approach to MI testing with full information ML estimation. (*Note that the traditional approach need not be included if the desired focus is only on the Bayesian implementation.*) We tested configural, metric, and scalar invariance. We then decided to explore partially invariant models only as applicable.

Next, we estimated the model using the Bayesian approximate MI approach for factor loadings and item intercepts. [*Depending on the journal audience, the authors may want to add a few sentences of justification for why a Bayesian approach was included. It may be helpful to include prose about the added flexibility of allowing for small differences through the use of the prior, rather than assuming exact equivalence through the traditional approach.*] We have followed the general guidelines presented in B. O. Muthén and Asparouhov (2013) for implementation of Bayesian approximate MI. Within the approximate MI process, difference priors were placed across the two schools (i.e., groups) for the factor loadings and intercepts. The difference priors took on this form: difference $\sim \mathcal{N}(0, \sigma^2)$, where the variance hyperparameter σ^2 was determined by incrementally testing several values. We then identified invariant and non-invariant parameters. The model was re-estimated with invariance parameters specified through near-zero priors. We used the M*plus* software version 8.4 (L. K. Muthén & Muthén, 1998-2017), and all code is presented in the online appendix.

For the traditional MI approach, we used the robust ML estimator. For the Bayesian implementation, we used the Gibbs sampler with two chains containing 50,000 burn-in iterations and 50,000 post-burn-in iterations. Convergence was monitored using the PSRF, or $\widehat{R}$, a convergence criterion developed by Gelman and Rubin and extended upon in later research (Brooks & Gelman, 1998; Gelman & Rubin, 1992a, 1992b; Vehtari et al., 2019). In order to ensure convergence was obtained, we used a stricter cutoff for the PSRF than the default software setting. We used a value of 1.01 rather than the default of 1.05. In addition to using the PSRF, we also visually examined all trace-plots for signs of non-convergence or other issues. To ensure that convergence was obtained, and that local convergence was not an issue, we estimated the model again with double the number of iterations (and double the length of burn-in). The PSRF criterion was satisfied and trace-plots still exhibited convergence. Next, we computed the percent of relative deviation, which can be used to assess how similar

results are across multiple analyses. To compute this deviation, we used the following equation for each model parameter: [(estimate from expanded model) − (estimate from initial model)/(estimate from initial model)] ∗ 100. We found that results were comparable across the two analyses, with relative deviation levels less than |1%|. After conducting these checks, we were confident that convergence was obtained for the final analysis. Aside from the small variance difference priors, default prior specifications in M*plus* were used for all parameters in the model (L. K. Muthén & Muthén, 1998-2017).

5.6.3 Hypothetical Results Section

Table 5.2 shows results for the traditional MI approach, with the first six rows representing model estimation using robust ML. There is support for metric invariance with a DIC of 7475.87. Given that freeing the factor (co)variances did not result in a decrease from strict invariance, this could be taken as a sign that these parameters can stay fixed across the groups. However, when factor means were freed, the DIC dropped. This indicates that the factor means are not all equal across groups. Notice that the PP*p*-values indicate that none of these models fit the observed data well.

Next, we implemented Bayesian approximate MI. We examined four potential settings for the near-zero prior variance hyperparameter setting. Results for these analyses are in Table 5.2 and, based on these results, we selected the value of 0.05 for the variance hyperparameter. There was no distinguishable difference between the three lowest variance values examined according to the DIC. Results for the analysis implementing the $\mathcal{N}(0, 0.05)$ difference prior setting on factor loadings and intercepts are reported in Table 5.3.

The first column represents the average estimate across groups, followed by the standard deviation. There was only one item intercept that deviated significantly across groups, and it was for Item 3 ("Lozenges"). Item 3 loads onto the "Visualization" factor, and the Group 1 intercept was 2.454 with the Group 2 intercept slightly lower at 2.135. This indicates that the item was slightly "easier" for Group 2 in comparison; a visual depiction of the group differences for this item intercept can be found in Figure 5.3. Otherwise, results were comparable across groups.

Following van de Schoot et al. (2013), we then constrained all loadings across groups. These parameters did not yield any significant differences through the Bayesian approximate MI process. These results are in the lower panel of Table 5.2 under the row heading of "Metric + Approx." Overall, results indicated that the "Metric + Approx" option is optimal based on the DIC, but none of the models fit according to the PP*p*-value.

5.6.4 Discussion Points Relevant to the Analysis

The Bayesian approximate MI approach was implemented here in order to introduce the added flexibility of allowing for "wiggle" room in the difference parameters rather than assuming exact equivalence across groups. It may not always be a viable approach to assume exact equivalence across groups. This Bayesian approach also works well when sample sizes within the groups are relatively small. One drawback of the approximate MI approach is that it can lead to biased results in latent factor means and variances when parameter differences across groups are large or systematic.

[*The researcher may go on to describe substantive differences that were obtained.*]

[*There are also issues tied to model fit that are further discussed in Chapter 11, which can be included in a discussion section for Bayesian approximate MI.*]

5.7 Chapter Summary

The Bayesian approximate MI approach allows for added flexibility in implementing "wiggle" room surrounding parameter differences. The ability to allow parameters to differ an amount that is not substantively meaningful can have broader impact on how MI is assessed. The traditional approach fixes group differences to be exactly zero. If certain fit criteria are not met, then the researcher may be left to relax the invariance specification altogether or even delete items from a scale that are deemed non-invariant. These actions could result in substantively altering the scale being examined, or treating negligible group differences as non-invariant. The Bayesian approximate MI approach allows for researchers to address these situations in a way that minimizes restrictions and improves flexibility of the modeling process.

An important component of the Bayesian approximate MI approach, just as with the traditional ML-based approach, deals with the assessment of model fit. In this chapter, I introduced the DIC and PP*p*-value as comparison and fit measures, respectively. Bayesian fit is a much larger issue than how it was presented in the current chapter. As a result, I have included additional information relevant to this topic in Chapter 11 regarding model fit and comparison.

5.7.1 Major Take-Home Points

The Bayesian approximate MI approach is highly flexible and circumvents the traditional requirement of assuming parameters are exactly equal across groups. Instead, this approach allows for a reasonable amount of "wig-

gle" room surrounding the parameter difference across groups. The idea here is that model results obtained are a more accurate representation of the substantive findings. In turn, the approach avoids possible model mis-specifications, where non-equivalent parameters are constrained to be equal. Here are some final points to consider surrounding Bayesian approximate MI:

1. Be aware of the *alignment issue*, which is linked to parameterization indeterminacies within the Bayesian approximate MI approach (B. O. Muthén & Asparouhov, 2013). Parameter differences must be small and non-systematic across groups in order for the near-zero priors to be properly implemented. If the assumption is violated, then approximate and partial MI should be combined to allow for systematic freeing of parameters in violation.

2. This approach can be easily scaled to handle many groups (or time points, as described in Chapter 8) and latent variables, and it works well when sample sizes are relatively small and traditional ML-based approaches fail (see, e.g., Winter & Depaoli, 2019).

3. The guidelines for selecting the variance hyperparameter value for the near-zero difference prior are still rather loose. Researchers should be mindful to carefully select the variance hyperparameter value, justify the selection, and potentially follow up with a sensitivity analysis examining the impact of different variance hyperparameter settings.

5.7.2 Notation Referenced

- $\boldsymbol{x}$: vector of observed indicators (e.g., items on a questionnaire)
- g: subscript of g denotes the parameter is allowed to vary across g groups
- $\boldsymbol{\tau}$: vector of intercepts tied to the x indicators
- q: the number of observed x variables
- $\boldsymbol{\Lambda}_x$: factor loading matrix for the x indicators
- $\boldsymbol{\xi}$: vector of latent factors
- $\boldsymbol{\delta}$: vector of measurement errors associated with x item indicators
- $\boldsymbol{\Sigma}(\boldsymbol{\theta})$: covariance matrix of x, as represented by $\boldsymbol{\theta}$
- $\boldsymbol{\Phi}_\xi$: covariance matrix for the latent factors ($\boldsymbol{\xi}$)
- $\boldsymbol{\Theta}_\delta$: covariance matrix for the error terms ($\boldsymbol{\delta}$)
- $E(\ldots)$: expected value
- $\boldsymbol{\kappa}$: vector of factor means
- $\mathcal{N}$: the normal prior distribution
- $\lambda_{21}^{(G1)}$: factor loading for Factor 1, Item 2, Group 1
- $\lambda_{21}^{(G2)}$: factor loading for Factor 1, Item 2, Group 2
- difference $\sim \mathcal{N}(0, \sigma^2)$: difference prior across two groups for a model parameter (e.g., a factor loading)
- σ^2: variance hyperparameter for the difference prior

5.7.3 Annotated Bibliography of Select Resources

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

- This book provides a comprehensive treatment of issues within the traditional approach to measurement invariance testing. It covers all of the steps researchers would take, as well as problems that can arise during the testing process.

Muthén, B. O., & Asparouhov, T. (2013). *BSEM measurement invariance analysis. Mplus Web Notes: No. 17.* Retrieved from https://www.statmodel.com/examples/webnotes/webnote17.pdf

- This unpublished webnote provides details surrounding the implementation and theory underlying Bayesian approximate measurement invariance. Along with examples and simulations, it provides explanations of issues such as the parameterization indeterminacies that can arise during Bayesian implementation of the process. It is a great resource for researchers wanting to implement these methods.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology: Quantitative Psychology and Measurement, 4*, 1-15.

- This paper introduces a thorough application of the Bayesian approximate measurement invariance process. It covers the benefits of the approach and then walks the reader through an example, highlighting the use of approximate-zero priors. It is a nice introduction to some of the issues that arise during Bayesian approximate MI testing.

5.7.4 Example Code for M*plus*

This is an example of partial M*plus* code for Bayesian approximate measurement invariance testing. In this case a difference prior of $\mathcal{N}(0, 0.05)$ is being implemented on factor loadings and intercepts. Arguments denoting estimation, number of chains, burn-in, and so forth, can be added to this base code.

```
MODEL:

%OVERALL%
f1 BY x1-x3*;
f2 BY x4-x6*;
f3 BY x7-x9*;
[x1-x9];

!  Labeling is crucial for invariance testing
!  Be sure to hold parameters free/constrained as needed

%c#1%
f1 BY x1-x3* (lam11-lam13);
f2 BY x4-x6* (lam14-lam16);
f3 BY x7-x9* (lam17-lam19);
[x1-x9] (nu11-nu19);

f1@1;
f2@1;
f3@1;
[f1@0];
[f2@0];
[f3@0];

%c#2%
f1 BY x1-x3* (lam21-lam23);
f2 BY x4-x6* (lam24-lam26);
f3 BY x7-x9* (lam27-lam29);
[x1-x9] (nu21-nu29);

f1@1;
f2@1;
f3@1;
[f1*0];
[f2*0];
```

```
[f3*0];

!f1 with f2 f3;
!f2 with f3;

MODEL PRIORS: !These are the near-zero, difference priors
DO(1,9) DIFF(lam1#-lam2#)~N(0,0.05);
DO(1,9) DIFF(nu1#-nu2#)~N(0,0.05);
```

For more information about these commands, please see the L. K. Muthén and Muthén (1998-2017) sections on CFA, multiple-group, invariance testing, and Bayesian analysis.

5.7.5 Example Code for R

Here is an example of basic measurement invariance using `blavaan` in R, but it does not include the use of difference priors akin to the M*plus* code provided. In this case, model fit is compared across a model with free loadings across groups (fit1) and a model with loadings held equal across groups (fit2).

```
library(blavaan)

HS.model <- ' visual =~ x1 + x2 + x3
textual =~ x4 + x5 + x6
speed =~ x7 + x8 + x9 '

fit1 <- bcfa(HS.model, data = HolzingerSwineford1939,
group = "school")

fit2 <- bcfa(HS.model, data = HolzingerSwineford1939,
dp = dpriors(...),
n.chains = 2,
burnin = 10000,
sample = 10000,
inits = "prior",
group = "school", group.equal = "loadings")
```

There are many helpful commands in the `blavaan` package, and this example code highlights the key features. The command `dp = dpriors(...)` can be used to override the default prior settings and list user-specified priors. The `n.chains` command controls the number of chains used in the

analysis. In this case, two chains have been specified for each model parameter. The `burnin` command is used to specify the number of iterations to be discarded in the burn-in phase. The `sample` command dictates the number of post-burn-in iterations (i.e., the number of iterations comprising the estimated posterior). The `inits` command can be used to specify the initial values for each model parameter. There are several different options that can be used here: "simple," "Mplus," "prior," and "jags." The default setting in `blavaan` is "prior," which determines the starting parameter values based on the prior distributions specified in the model. The command `group` indicates the grouping variable is *school* for this analysis. Finally, the command `group.equal` can be used for the invariance testing process.

For more information on using the `bcfa` command in `blavaan`, see Merkle and Rosseel (2018) for a tutorial.

Guilford Publications
370 Seventh Avenue
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com