Selecting an Optimal Measurement Model and Detecting Differential Item Functioning Using Bayesian Confirmatory Factor Analysis

By

Terrence D. Jorgensen

Submitted to the Department of Psychology and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

Chairperson: Wei Wu, Ph.D.

Pascal R. Deboeck, Ph.D.

Carol M. Woods, Ph.D.

William P. Skorupski, Ph.D.

Paul E. Johnson, Ph.D.

Date Defended: May 1st, 2015

The Dissertation Committee for Terrence D. Jorgensen

certifies that this is the approved version of the following dissertation:

Selecting an Optimal Measurement Model and Detecting Differential Item Functioning Using Bayesian Confirmatory Factor Analysis

_____

Chairperson: Wei Wu, Ph.D.

Date approved: May 1st, 2015

Abstract

I investigated the sampling behavior of DIC and WAIC in the context of selecting an optimal measurement model in Bayesian SEM, as well as the utility of highly constrained parameter estimates in detecting differential item functioning (DIF). I assessed the relative efficiency of WAIC compared to DIC, evaluated analytical WAIC *SE*s by calculating relative bias, and reported how often WAIC and DIC indicated a preference for each invariance model. I compared the power and Type I error rates for DIF detection across conditions, and assessed the quality of estimates by calculating bias and 95% CI coverage rates for key parameters. Results indicate that although WAIC has less sampling variability than DIC, their model preferences are similar. Both WAIC and DIC have greater power to detect that invariance constraints are untenable than AIC in using maximum likelihood (ML) estimation. In tests of null hypotheses that DIF parameters are zero, Bayesian credible intervals and ML modification indices have similar power, but Bayesian credible intervals have much lower Type I error rates.

Acknowledgements

I would like to thank each of my committee members—Drs. Wei Wu, Pascal Deboeck, Carol Woods, Billy Skorupski, and Paul Johnson—for their expertise, time, guidance, and support. I am grateful for the statistical resources and technological support provided by the University of Kansas's Center for Research Methods and Data Analysis. I would also like to thank my wife, Katharina Jorgensen, for excusing my antisocial behavior while undertaking this task, and my family for understanding why I spent most of the winter holidays working.

Table of Contents

Abstract iii
Acknowledgement iv
Table of Contents v
List of Tables and Figures vii
PART I: Literature Review 1
Defining a Model's Goodness-of-Fit 2
Defining Model Fit in SEM 5
Covariance structure analysis 6
Model fit in covariance structure analysis 9
Residuals-based fit measures 9
χ^2 fit statistic 10
χ^2-based fit indices 12
Sources of misfit 15
Application of Fit Measures in Traditional SEM 16
Model evaluation 17
Model modification 19
Tools for model modification 20
Model comparison 24
Nested model comparisons 25
Nonnested model comparisons 28
Bayesian SEM 32
Bayesian statistical inference 32
Estimating Bayesian models 35
Bayesian Model Fit 36
Model evaluation 37
Model modification 39
Model comparison 41
Bayes factors 42

Information criteria 43
Summary of Bayesian Model-Comparison Tools 47
PART II: Assessing Bayesian Tools for Selecting an Optimal Measurement Model . . . 49
Monte Carlo Design for Study 1 50
Procedure 54
Results and Discussion 55
Variability of information criteria 56
Impact of model misspecification 67
Model rankings and preferences 70
PART III: Assessing Bayesian Tools for Detecting DIF 78
Monte Carlo Design for Study 2 79
Procedure 80
Results and Discussion 83
Nonconverged Models 83
Variability of parameter estimates 84
Rejection Rates 89
PART IV: General Discussion 96
Limitations and Future Directions 99
Conclusions 103
References 106
Appendix: Prior Distributions for Model Parameters 116

List of Tables and Figures

Table 1: Manipulated Variables in Monte Carlo Design for Studies 1 and 2 51
Table 2: Effect Sizes (η^2) of Monte Carlo Factors on Information Criteria 57
Table 3: Effect Sizes (partial-η^2) of Monte Carlo Factors on Parameter Estimates 85
Figure 1: Population model(s) for data generation in Study 1 52
Figure 2: Mean $WAIC_1$ across conditions 58
Figure 3: Mean AIC across conditions 60
Figure 4: Mean DIC_1 across conditions 60
Figure 5: Mean DIC_2 across conditions 61
Figure 6: Standard deviations of four information criteria across conditions 62
Figure 7: Standard deviations of all five information criteria across conditions 62
Figure 8: Relative efficiency of DIC_1 to DIC_2 63
Figure 9: Relative efficiency of $WAIC_2$ to $WAIC_1$ 64
Figure 10: Relative efficiency of $WAIC_2$ to DIC_1 65
Figure 11: Relative *SE* bias of $WAIC_2$ 66
Figure 12: Effect of DIF, model type, and parsimony error on latent-mean bias 68
Figure 13: Effect of DIF, model type, and parsimony error on latent-variance bias 68
Figure 14: Effect of DIF, model type, and parsimony error on CFI 69
Figure 15: Effect of DIF, model type, and parsimony error on RMSEA 70
Figure 16: Model preferences based on ranked AIC, DIC_1, $WAIC_1$, and $WAIC_2$ 71
Figure 17: Model preferences based on ranked DIC_2 73
Figure 18: How often the lowest WAIC's 95% CI contains the next lowest WAIC 75
Figure 19: How often the lowest WAIC's 95% CI contains the highest WAIC 75
Figure 20: How often the second lowest WAIC's 95% CI contains the highest WAIC 76
Figure 21: Model preferences including the correct partial invariance model 78
Figure 22: Convergence rates for each model across conditions 83
Figure 23: Bias in the second latent mean grows in magnitude as DIF increases 86
Figure 24: Average posterior mean of the second latent *SD* by DIF 86
Figure 25: Average posterior mean of $\Delta\lambda$s by DIF, prior σ, and *N* 87

Figure 26: Average posterior mean of Δτs by DIF, prior σ, and *N* 88
Figure 27: Rejection rates for Δλs by DIF, prior σ, and *N* 89
Figure 28: Rejection rates for Δτs by DIF, prior σ, and *N* 91
Figure 29: Maximum-DIF rejection rates for Δτs by DIF, prior σ, and *N* 91
Figure 30: False discovery rates (FDR) by DIF, prior σ, and *N* 92
Figure 31: Power and Type I error rates for detecting DIF using modification indices 93
Figure 32: Type I error rates by DIF, prior σ, and *N* 95
Figure 32: Power by DIF, prior σ, and *N* 95

Selecting an Optimal Measurement Model and Detecting Differential Item Functioning Using Bayesian Confirmatory Factor Analysis

PART I: Literature Review

Bayesian methods have been incorporated into popular software packages to estimate structural equation models (SEM), such as Amos (Arbuckle, 2012) and M*plus* (Muthén & Muthén, 2012). This has resulted in increased popularity of such estimators in applied research (Andrews & Baguley, 2013). More frequent use of Bayesian estimation will be accompanied by a greater demand for methods to evaluate SEMs in a Bayesian context (Levy, 2011). Applied users might be motivated to use a Bayesian estimation technique to fit an SEM that is intractable to estimate with maximum likelihood (ML) or may be attracted to the Bayesian framework for its interpretational benefits (Gelman et al., 2014; Iversen, 1984). In either case, the scientific community can expect future journal articles to include the use of Bayesian methods to answer research questions about measurement equivalence (or "invariance") and differential item functioning (DIF), which are related topics of frequent methodological research in the context of latent variable models such as SEM and item-response theory (IRT).

I begin this section with an introduction to the concept of model fit, followed by a thorough literature review of methods for evaluating traditional SEMs in a frequentist framework. I then provide a conceptual introduction to the Bayesian approach to statistical inference and estimation, and I review existing methods for Bayesian SEM (BSEM) evaluation. I use common applications of model comparison and modification (e.g., testing measurement invariance, identifying misspecified parameters) to contrast the frequentist and Bayesian approaches of assessing model fit. I conclude the review with a discussion of gaps in the invariance testing literature. Finally, I propose a study to investigate the frequency properties of

tools for selecting an optimal measurement model and for detecting DIF in a Bayesian context.

Defining a Model's Goodness-of-Fit

The fit of a model in many contexts (e.g., regression, multilevel models) refers to how similar the observed values are to the predicted values implied by the model. For instance, in an "intercept-only" general linear model

$$Y_i = \hat{\beta}_0 + e_i \quad (1)$$

the estimate of β_0 is the sample mean $\bar{Y}$, and the residuals e_i are the mean-centered data, representing how much the i^{th} observation deviates from $\bar{Y}$. In this model, each observation's predicted value $\hat{Y}_i$ is the sample mean $\bar{Y}$, and the estimated residual variance of e_i is the total sample variance. If *X* is a variable correlated with *Y*, its inclusion in the model

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \quad (2)$$

will improve the predicted values $\hat{Y}_i$ in the sense that the discrepancies between predicted and observed values (i.e., residuals, $e_i = Y_i - \hat{Y}_i$) will be smaller, on average. The residual variance of e_i thus decreases by the amount of shared variance between *X* and *Y*.

In the ANOVA decomposition of a linear model (Maxwell & Delaney, 2004), the degree to which predicted values differ from observed values is estimated by the mean squared error (*MSE*), the square-root of which (*RMSE*) is the *SD* of the residuals. The magnitude of these discrepancies relative to the total sample variance $\left(\frac{SSE}{SST}\right)$ indicates the degree to which the model fails to perfectly predict each observed Y_i, and its complement is thus a measure of model goodness-of-fit $\left(1 - \frac{SSE}{SST}\right)$, more commonly known as R^2 (interchangeably known as η^2 in software such as SPSS). R^2 is most commonly interpreted as the proportion of variance in the outcome *y* that is explained by the linear combination of predictors. Another interpretation,

which is more related to model fit, is that R^2 quantifies the degree of correspondence between observed and predicted values of *y*—where predicted values are a linear combination of predictors.

Model fit may be evaluated in an absolute sense (without reference to any competing alternative models) using R^2, but applied researchers are often interested in how much a model improves by including additional predictors or covariates[1]. Because R^2 will nearly always increase with the inclusion of any additional predictor, regardless of its merit, a researcher using the highest R^2 as criterion for the "best" model will always choose the most complex, inclusive model. A researcher could continue to add as many predictors as there are observations, at which point the model has no degrees of freedom (*df*). A model with no *df* explains 100% of the sample variance (i.e., $R^2 = 1$), but it has no utility because the predicted values are not free to differ from the observed values. Such a model explains nothing—it is descriptive at best.

Given the same approximate level of predictive accuracy, the principle of parsimony prefers the simplest available model. A common expression of this principle is Occam's Razor, named for Sir William of Occam, who stated that if two explanations are practically equivalent (i.e., they make nearly equal predictions), the simplest explanation should be preferred. In a statistical modeling context, "simplest" indicates the model requiring the fewest independent entities (e.g., predictor variables, functional form of effect on outcome) to make predictions of the same accuracy. An adjusted R^2 has been formulated for general linear models in the spirit of this principle, "punishing" the estimated goodness-of-fit by taking the number of parameters into

[1] The distinction between *predictors* and *covariates* is purely substantive, not statistical. They play the same role mathematically, but the effects of covariates on an outcome are of little to no substantive interest to an applied researcher. Covariates are included to control for nuisance effects, to generate more accurate predicted values for distinct subpopulations, or to increase power to detect effects of interest by reducing the residual variance, but the effects of interest involve predictors (also called *independent* or *quasi-independent* variables, depending on whether they are manipulated by design). Throughout this paper , I refer to both predictors and covariates as predictors.

account so that adding parameters would be preferable only when they improve model fit beyond what would be expected from sampling fluctuation (Maxwell & Delaney, 2004).

It is useful to distinguish between two types of accuracy—(a) in estimation and (b) in prediction—which necessitates distinguishing between two types of quantity that appear in mathematical and statistical models: variables and parameters. The quantities *X* and *Y* in (2) are predictor and outcome variables, respectively; they are vectors of individuals' scores on some measureable phenomena. The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ in (2) are estimates of population parameters β_0 and β_1 that describe the relationship between the variables. The terms *prediction* and *estimation* refer to scores (i.e., $\hat{Y}_i$ conditional on $\hat{X}_i$) and parameters, respectively, and model fit could refer to the accuracy of predictions or to the accuracy of the form of the model (i.e., What variables are included as predictors, and thus what parameters describe the effects of those predictors?). The error term e_i is a variable that represents how each case's observed value differs from that case's predicted value. Although individual residuals can be calculated from the model results, e_i is an unobserved (i.e., *latent*) variable, and its variance is the portion of the total variance in *Y* that is not explained by the predictor(s). Thus, the residual term also represents any and all true effects on *Y* of potential predictors not included in the analysis model.

The intercept-only model (1) will typically yield less accurate predictions than (2), so the fit of (2) will be superior to the fit of (1) with respect to predicted values. However, both models might be accurate with respect to the parameters, assuming the normality assumption holds and the *X*–*Y* relationship is linear. That is, neither model should be expected to include all predictors that cause individual differences in *Y*, but because e_i represents any and all such potential predictors, the model in (2) is correct in the sense that it provides the best linear unbiased predictions (BLUPs) of *Y* conditional on observed values of *X*, and the model in (1) is correct in

the sense that it provides the best linear unbiased estimate (BLUE) of the population mean (i.e., $\hat{\beta}_0$ is the unconditional sample mean $\bar{Y}$).

With respect to the free parameters included in an analysis model, R^2 does not measure of model fit; it is only a measure of fit with respect to accuracy of predicted values. Diagnostics for general linear models are not calculated directly from parameters, but the omission of important effects can be detected using plots of residuals against predictors (or potential predictors). There are many other measures developed to indicate fit (or misfit) of a general linear model with respect to predictions, such as the predicted sum of squares (a leave-one-out method). In other modeling contexts (e.g., a generalized linear model such as binary logistic regression), goodness of fit might be defined in terms of observed and expected counts in unconditional and conditional contingency tables, respectively, for categorical data, rather than in terms of explained and unexplained variance in a continuous variable. The common aspect of all such methods typically involves evaluating predicted values with respect to observed values, but describing all such methods is beyond the scope of this review. The remainder of the review will focus on model goodness-of-fit in the context of SEM.

Defining Model Fit in SEM

Models are mathematical representations of the population[2] processes that give rise to observable, real-world phenomena. Human behavior can be influenced by a variety of sources, so the true population process for any particular social phenomenon might be infinitely complex. By necessity, an analysis model is merely an approximation of the true population (MacCallum, 2003). Structural models can include several predictors and outcomes, representing more

[2] The term population informally refers to a group of people with some common characteristic(s) of interest. In statistics, a population is a process that gives rise to observable data. Even the entire "population" (in the informal sense) is merely a sample of all cases that could presently be observed, but not all possible cases that could ever be observed. Throughout this review, I use the term population in the statistical sense: a data-generating process.

complex relationships than a general linear model can. In fact, the general linear model can be seen as a special case of SEM, one in which there is only one outcome variable with any number of predictors. SEM is more general because several outcomes can be included in the model, each of which has a linear prediction equation associated with it—a *sub*model—and the parameters of all submodels are estimated simultaneously. A variable can even take on the role of a predictor and an outcome in the same model, representing a chain of causation—this is often called a *mediation* model. SEMs can also include latent variables, which are estimated by modeling their effects of observed variables used to measure them. The flexibility and complexity of SEM make it an attractive modeling framework for social and behavioral scientists.

In the context of SEM, model fit is typically defined in terms of the summary statistics (i.e., means, variances, and covariances). Specifically, when fitting a specified model to observed data, the estimated parameters yield predicted values of the variances and covariances among the variables (Brown, 2006), and a well-fitting model is one whose model-implied (i.e., predicted) covariance matrix closely resembles the observed covariance matrix. To explain why SEM model fit concerns discrepancies of predicted summary statistics rather than individual predicted scores, I must briefly discuss how SEMs have traditionally been estimated.

Covariance structure analysis. Many psychological constructs (e.g., depression, intelligence) cannot be measured directly because they cannot be perceived with the senses, and are thus commonly referred to as latent constructs. Instead, latent constructs must be measured indirectly. Observable behaviors can indicate someone's level on a latent variable—for example, higher political conservatism could be expected to correspond with observing (a) higher indications of such an orientation on a questionnaire or (b) more frequently casting votes for conservative candidates in elections. Psychological researchers have historically used scales to

measure latent constructs because each scale item is designed to indicate a respondent's level on a latent variable. Scales can measure attitudes in social psychological research; frequency or duration of symptoms can measure psychological disorders in a clinical context; or test performance can measure competency in an educational setting. In each of these examples, the indicators are assumed to be related due to a common cause—the latent construct they are designed to measure.

The common factor model can be applied to data that conform to this assumption. In its simplest form (a single-factor model), observable indicators x are treated as outcomes of an unobserved common factor F (i.e., a latent construct):

$$x_{pi} = \lambda_p \times F_i + \varepsilon_{pi} \tag{3}$$

where i is an index for N observations ($i = 1, 2, \ldots N$), p is an index for P observable indicators ($p = 1, 2, \ldots P$), λ_p is the regression weight relating indicator x_p to factor F, and ε_p is the residual term, representing the uniqueness of indicator x_p—unique in that it is unrelated to the common factor F or any other indicators. The model in (3) assumes that F and all x_p are centered at their means, but this assumption can be relaxed by adding an intercept term (τ) for each x:

$$x_{pi} = \tau_p + \lambda_p \times F_i + \varepsilon_{pi} \tag{4}$$

Because F is not directly observed, it is not possible to directly estimate the regression weights (λ), intercepts (τ), or variance of residuals (ε). However, regression coefficients can be estimated without access to individual observations (x, y), if certain summary statistics are available—namely the mean vector ($\boldsymbol{M}$) and covariance matrix ($\boldsymbol{S}$) of x and y. In the linear model (2), regression coefficients are estimated as:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{5}$$

The variance of F in (3) cannot be estimated, nor can its covariance with each indicator, because

F is unobserved. But the covariance between each indicator and *F* can be estimated by imposing certain constraints (see Brown, 2006) and reproducing the observed covariances among indicators as a criterion. The effect of *F* on each *x* in (3) or (4) can be iteratively estimated (e.g., using MLE), as long as the model can be identified by fixing the variance of *F* to be equal to 1 (fixed-factor approach) or to be equal to the common variance of one indicator x_p (marker-variable approach).

The common factor model is therefore typically fit to data as an analysis of the covariance structure among a set of observed variables. More complex exploratory and confirmatory factor analyses (EFA and CFA) also operate on the assumption that covariances among manifest variables can be explained by a number of common factors (fewer than the number of observed variables) that have a linear effect on the indicators. For example, a two-factor model for person *i*'s p^{th} manifest variable x_{pi} would be represented as

$$x_{pi} = \tau_p + \lambda_{p1}F_{1i} + \lambda_{p2}F_{2i} + \varepsilon_{pi} \quad (6)$$

This model expresses the observed $\boldsymbol{S}$ as a function of (as many or fewer) parameters—the matrix of regression weights (Λ), the covariance matrix of latent variables (Φ), and the covariance matrix of indicator residuals (Θ):

$$\boldsymbol{S} \sim \hat{\boldsymbol{\Sigma}} = \Lambda\Phi\Lambda^{t} + \Theta \quad (7)$$

and the observed $\boldsymbol{M}$ as a function of Λ, the latent means ($\boldsymbol{\alpha}$), and the indicator intercepts ($\boldsymbol{\tau}$):

$$\boldsymbol{M} \sim \hat{\boldsymbol{\mu}} = \boldsymbol{\tau} + \Lambda\boldsymbol{\alpha} \quad (8)$$

These factor analysis models are special cases of SEM, in which the latent variables are freely correlated without any directed paths (i.e., no regressions among latent variables). If there are latent regressions, represented in the β matrix, the more general covariance structure is

$$\hat{\Sigma} = \Lambda\,(I^{Q} - \beta^{-1})\,\Phi\,(I^{Q} - \beta^{-1})^{t}\,\Lambda^{t} + \Theta \quad (9)$$

where I^Q is an identity matrix of dimension Q = the number of latent factors.

Model fit in covariance structure analysis. Because SEMs are traditionally fit as analyses of covariance structure, models must be evaluated in terms of covariance structure.[3] There are no predicted values for the indicators (outcome variables) because individuals' values on the latent factors (predictor variables) are not observed. Instead, a covariance structure analysis (CSA) results in model-implied predictions $\hat{\Sigma}$ for values in the population covariance matrix Σ among the indicator variables. Because Σ is unavailable for direct comparison with $\hat{\Sigma}$, model fit is evaluated using discrepancies between $\hat{\Sigma}$ and the observed covariance matrix $\boldsymbol{S}$. There are numerous ways to quantify these discrepancies, and several fit indices have been formulated to detect or describe different aspects of model misfit. Exhaustive reviews of numerous fit measures may be found in Hu and Bentler (1998) and West, Taylor, and Wu (2012), but I discuss only a few popular ones here.

Residuals-based fit measures. Just as residuals can be calculated between observed and predicted individual scores in a general linear model, residuals in CSA are calculated by subtracting elements in $\hat{\Sigma}$ from elements in $\boldsymbol{S}$ (also, elements in $\hat{\boldsymbol{\mu}}$ from elements in $\boldsymbol{M}$, if the model includes a mean structure). Residuals can be inspected on an individual basis to discover which relationships among observed variables are not adequately reproduced by the model. A summary measure of the residuals can also be used. The square-root of the average of squared residuals is the root mean-squared residual (RMR), which provides an average magnitude of residuals in the original (co)variance metric. More commonly, a standardized measure is calculated (SRMR) by scaling the residuals by their respective standard deviations. A weakness of using a single-number summary of residuals is that misspecifying the relationships between a

[3] This constraint is not necessary in a Bayesian context, as later sections will discuss.

small number of variables might go unnoticed in a model with a large number of observed variables if, on average, the residuals are small.

χ^2 fit statistic. Parameters are estimated iteratively, with the criterion of minimizing a discrepancy function *F* between $\hat{\Sigma}$ and $\boldsymbol{S}$. Discrepancy functions are thus a function of residuals (i.e., differences between observed sample moments and model-implied population moments). In general, the elements of $\hat{\Sigma}$ and $\boldsymbol{S}$ can be stacked into individual vectors $\hat{\boldsymbol{\sigma}}$ and $\boldsymbol{s}$, respectively, and each discrepancy is squared and summed, after being scaled by a weight matrix $\mathbf{W}$ (Browne, 1984):

$$F_{\text{general}} = (\boldsymbol{s} - \hat{\boldsymbol{\sigma}})^{\mathrm{T}}\mathbf{W}^{-1}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}) \tag{10}$$

In unweighted least squares (ULS) estimation, the weight matrix is an identity matrix, so it is merely the sum of squared discrepancies:

$$F_{\text{ULS}} = (\boldsymbol{s} - \hat{\boldsymbol{\sigma}})^{\mathrm{T}}(\boldsymbol{s} - \hat{\boldsymbol{\sigma}}) = \frac{1}{2}tr\left[\left(\boldsymbol{S} - \hat{\Sigma}\right)^2\right] \tag{11}$$

Weighted least squares (WLS) estimation has several special cases. In generalized least squares (GLS) estimation, the weight matrix is a function of $\boldsymbol{S}$, and assuming multivariate normality the equation can be simplified to

$$F_{\text{GLS}} = \frac{1}{2}tr\left[\left(\mathbf{I}^{\mathrm{Q}} - \boldsymbol{S}^{-1}\hat{\Sigma}\right)^2\right] \tag{12}$$

Asymptotically distribution-free (ADF) estimation involves calculating the weight matrix from the excess kurtosis among the indicators, allowing the normality assumption to be relaxed in asymptotically large sample sizes (e.g., $N > 1000$ or 5000). The most popular discrepancy function among applied researchers—due in no small part to it being the default estimator in most software—is the maximum likelihood (ML) estimator:

$$F_{\text{ML}} = \log\left|\hat{\Sigma}\right| - \log|\boldsymbol{S}| + tr\left(\boldsymbol{S}\hat{\Sigma}^{-1}\right) - P \tag{13}$$

where P is the number of observed variables. F_{ML} also assumes multivariate normality of indicators. If a mean structure is included in the model, F_{ML} is amended with another term for discrepancies in the mean vector:

$$F_{\mathrm{ML}} = \log\left|\hat{\Sigma}\right| - \log|\boldsymbol{S}| + tr\left(\boldsymbol{S}\hat{\Sigma}^{-1}\right) - P + (\boldsymbol{M} - \boldsymbol{\mu})^{\mathrm{T}}\hat{\Sigma}^{-1}(\boldsymbol{M} - \boldsymbol{\mu}) \tag{14}$$

Due to its popularity, I will refer solely to F_{ML} throughout this review unless otherwise stated. Regardless of which discrepancy function is used, a test statistic is calculated as the product of the discrepancy function and the sample size[4]:

$$T = N \times F_{\mathrm{ML}} \tag{15}$$

If the normality assumption is met, N is large enough, and the model is correctly specified, T approximately follows a central χ^2 distribution with *df* equal to the number of observed sample moments (means, variances, and covariances) minus the number of estimated parameters in the model. The deviance can also be used to calculate the ML χ^2 statistic. The deviance = $-2 \times \log\left(p(Y|\boldsymbol{\theta})\right)$, where $p(Y \mid \boldsymbol{\theta})$ is the likelihood of observing the observed data (Y), conditional on the vector of model parameters ($\boldsymbol{\theta}$). The deviance is distributed as a χ^2 random variable, so the χ^2 statistic for a model is the difference between the deviance of that target model and the deviance of the saturated model.

If the variables are continuous but nonnormal, an adjusted χ^2 statistic (and *SE*s) can be calculated using excess kurtosis of indicators. If the hypothesized model does not precisely match the population model, then T is approximately distributed as a noncentral χ^2 random variable, with the same *df* but also a noncentrality parameter λ that depends on the magnitude of

[4] Early software such as LISREL (Jöreskog & Sörbom, 2006) and EQS (Bentler, 2006) used $N - 1$ instead of N because without a mean structure, CSA likelihood follows from a Wishart distribution. More recently developed software such as Mplus (Muthén & Muthén, 2012) and lavaan (Rosseel, 2012) include a mean structure by default, and so their likelihood functions follow from a normal distribution and use N as the multiplier (Widamin & Thompson, 2003).

discrepancy between hypothesized and true models.

The χ^2 statistic provides a test of exact fit—that is, a test of the null hypothesis that there is no difference whatsoever between the model-implied and observed sample moments (H_0: $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S}$), which is a proxy for the untestable null hypothesis that the hypothesized target model is identical to the true population model (H_0: $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$). Similar to other test statistics, larger *N* increases its power to detect smaller inconsistencies with H_0. Because hypothesized models are, by necessity, mere approximations of reality, a test of exact fit has limited utility. Large *N* yields enough power to detect even small model–data discrepancies, so small that they are of no practical importance (in the sense that predicted values are close enough to observed values that they would be useful in an applied setting). It is this limitation of the χ^2 statistic that motivated several methodologists to develop alternative indices of fit, a few of which I discuss next.

χ^2-based fit indices. The χ^2 statistic provides a test of statistical significance of the observed model–data discrepancy. Like other statistical tests (e.g., independent-samples *t*), interpretation of a rejected H_0 is facilitated by a measure of effect size (e.g., Cohen's *d*). Practical fit indices were developed for the same purpose when evaluating the practical significance of model–data discrepancy. Other than the aforementioned residuals-based fit indices, most fit indices are calculated as a function of the χ^2 statistic.

The only fit measure with a known distribution is the root mean-squared error of approximation (RMSEA), which is based on the noncentrality parameter $\hat{\lambda}$, estimated as the difference between χ^2 and its expected value (*df*):

$$\text{RMSEA} = \sqrt{\max\left(0, \frac{\hat{\lambda}}{df(N)}\right)} = \sqrt{\max\left(0, \frac{\chi^2 - df}{df(N)}\right)} \qquad (16)$$

RMSEA is thus a measure of how much misfit there is, on average, per *df*, in the metric of the

discrepancy function (i.e., with the influence of *N* removed). Confidence intervals can be constructed for RMSEA using the upper and lower limits of the noncentral χ^2 with noncentrality parameter $\hat{\lambda}$ (Curran, Bollen, Chen, Paxton, & Kirby, 2003), which can then be used to test hypotheses of close fit rather than exact fit (MacCallum, Browne, & Cai, 2006). A limitation of this approach is that the value of RMSEA does not have a clear interpretation (Browne & Cudeck, 1992), so setting a null-hypothesized value of RMSEA for a test of close fit is arbitrary.

RMSEA is a measure of absolute misfit, in the sense that the model is judged in isolation (without respect to another model) and higher numbers indicate worse fit. Another index based on the noncentrality parameter is McDonald's noncentrality index:

$$\mathrm{Mc} = e^{-\frac{1}{2}\left(\frac{\hat{\lambda}}{N}\right)} = e^{-\frac{1}{2}\left(\frac{\chi^2 - df}{N}\right)} \tag{17}$$

The interpretation is no more straight-forward than for RMSEA, but Mc is a measure of goodness of fit, in that higher values (theoretical upper bound of 1) indicate better fit (West et al., 2012).

The comparative fit index (CFI; Bentler, 1990) also utilizes the estimated noncentrality parameter, but it belongs to a class of indices called comparative or incremental fit indices, which quantify model fit by comparing the fit of the target model (χ^2_{T}) to the fit of a baseline model (χ^2_{B}):

$$\mathrm{CFI} = 1 - \frac{\max\left(\hat{\lambda}_{\mathrm{T}}, 0\right)}{\max\left(\hat{\lambda}_{\mathrm{B}}, 0\right)} = 1 - \frac{\max\left(\chi^2_{\mathrm{T}} - df_{\mathrm{T}}, 0\right)}{\max\left(\chi^2_{\mathrm{B}} - df_{\mathrm{B}}, 0\right)} \tag{18}$$

This is in contrast to indices such as SRMR, RMSEA, and Mc, which quantify absolute (mis)fit of an isolated model. The Tucker–Lewis index (TLI, also called NNFI; Bentler & Bonnett, 1980) is another popular incremental fit index, originally developed to help identify the appropriate number of factors in EFA. Its calculation is similar, but instead of the noncentrality

parameter (i.e., the difference between the χ^2 and *df*), it is calculated using the χ^2-to-*df* ratio.

Incremental fit indices are based on the idea that there is a continuum between the worst-fitting model (represented by the baseline model, in which variables are typically not allowed to correlate) and the best-fitting model (represented by the saturated model, in which all observed associations are freely estimated). Target models lie somewhere between these two extremes, and incremental fit indices indicate where along the continuum the target model is located—values closer to 0 indicate the target model is closer to the poor-fitting baseline model, and values closer to 1 indicate the target model is closer to the perfect-fitting saturated model. This allows nonnested target models to compared, so long as they are both nested within the same saturated model, and the same baseline model can be specified to be nested within both competing target models (Bentler & Bonnett, 1980; Widamin & Thompson, 2003).

The goodness-of-fit index (GFI) is an absolute fit index whose interpretation is similar to R^2 in general linear models—the proximity between observed sample moments and model-implied predictions of those moments. Values closer to 1 indicate closer proximity and thus better fit. Like R^2 (Maxwell & Delaney, 2004), the GFI is upwardly biased in finite samples (West et al., 2012), which led Maiti and Mukherjee (1990) to revise its calculation (GFI*), commonly referred to as gamma hat (Hu & Bentler, 1998, 1999):

$$\text{Gamma Hat} = \frac{P}{P+2\left(\frac{\hat{\lambda}}{N}\right)} = \frac{P}{P+2\left(\frac{\chi^2-df}{N}\right)} \tag{19}$$

where *P* is the number of observed variables.

These absolute and incremental fit indices are among the most commonly used because of their lack of sensitivity to sample size and their sensitivity to misfit in different types of models (Fan & Sivo, 2007, 2009). Another class of fit indices is called information criteria

because they are based on information theory, and they are used solely for model comparison. They are formulated to take model complexity into account, providing a basis on which to choose models that balance good fit with parsimony. Gelman, Hwang, and Vehtari (2013) reviewed several information criteria, the most popular of which are Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC).

Information criteria follow a common template:

$$\text{IC} = F_{\text{ML}} + Z \qquad \text{or} \qquad \text{IC} = \chi^2 + Z \tag{20}$$

where Z is a term that punishes fit (i.e., adds to the measure of misfit). Lower values of an information criterion are thus preferred, and because of (15), the rank order of the models is unchanged whether the χ^2 statistic or F_{ML} is used. The difference between information criteria lies in the calculation of Z. AIC punishes the addition of parameters: $Z = 2 \times k$, where k is the number of free parameters in the model. BIC punishes the addition of parameters more severely with increasing sample size: $Z = k \times \log(N)$. A frequently noted weakness of information criteria is that although they provide a criterion to choose among competing models (the lowest value indicates the preferred model), there is no indication of the practical difference between models. This weakness, however, is not unique to information criteria, as the metric of many fit indices is rarely well defined.

Sources of misfit. The global fit measures described above quantify global model fit (i.e., how well the model as a whole fits the data). Other tools are available to identify local sources of misfit, such as a predicted correlation between variables x and y that is much lower or higher than the observed correlation. It is important to note that local discrepancies such as this could be due to mere sampling fluctuation, in which case the model might be modified to fit a fluke in the data that would not be generalizable to future samples from the same population

(MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992). This is why it is important to consider model modification as an exploratory process and to verify any modified model using independent data—this could be accomplished by randomly splitting the original sample into training and validation samples if the original sample size were large enough (Browne & Cudeck, 1992). Errors due to mere sampling variability are referred to as *sampling error* or *estimation discrepancy* (MacCallum, 2003), and they cause discrepancies between observed and model-implied covariance matrices ($\boldsymbol{S}$ and $\hat{\Sigma}$) because no individual sample covariance matrix (to which the model is fit) will be identical to the covariance matrix of the population (Σ) from which it was drawn, even if the model were perfectly specified (i.e., no difference between the target and population models).

However, local discrepancies might also indicate true model misspecifications (e.g., omitted variables, or omitted parameters relating the variables included in the model). When the model is misspecified, discrepancies occur because the target model differs from the population model. In other words, even if $\boldsymbol{S} = \Sigma$ (i.e., no sampling error), fitting the target model to Σ would not yield an identical model-implied covariance matrix $\hat{\Sigma}$. This source of error is referred to as *model error* or *approximation discrepancy* (MacCallum, 2003). In practice, it is impossible to distinguish or separate the effects of sampling and model error. Tools for model modification (discussed in the Model Modification section) are used on the assumption that they will detect model errors, but this must be confirmed on independent data.

Application of Fit Measures in Traditional SEM

Model fit measures can be applied in numerous scenarios. I will focus on three general categories: evaluation, modification, and comparison. Evaluation refers to judging the global fit of a single SEM, without reference to any competing model. If a model is judged to fit the data

inadequately, researchers may look for sources of misspecification with the goal of modifying their original model. Modification is a method of constructing a competing model post hoc, but researchers may have specified two or more competing models a priori. Model comparison refers to choosing the most appropriate among competing SEMs, using model fit as at least one criterion—other criteria, perhaps, being generalizability and theoretical plausibility.

Model evaluation. There are many special cases of SEM: path analyses estimate regressions among observed variables; factor analyses relate observed indicators to latent constructs (i.e., a measurement model); and general SEMs include aspects of path analysis and factor analysis (i.e., a measurement model for latent constructs, accompanied by regressions among latent constructs). But any published SEM must include an evaluation of global fit, even if it has been modified or compared to other models. Because SEMs are statistical models constructed to represent theories of the relationships among variables, the global fit of a model quantifies the correspondence between a researcher's theory (the target model) and reality (the true population model, an instance of which is represented in the observed data).

Most fit measures discussed in the previous section can be used to evaluate models in an global sense. The χ^2 statistic provides a statistical test of the null hypothesis that the target model perfectly explains the sample data. Because SEM requires a large *N* to ensure convergence and precision of estimates (Bollen, 1989), this test is often powerful enough to detect even negligible discrepancies between the observed sample moments and predicted moments implied by the parameter estimates. This is not to say the χ^2 statistic is not useful, but it only provides information about whether the model fits the data perfectly, not the magnitude of discrepancy or whether the discrepancy is of any practical consequence.

Global model fit is therefore evaluated by supplementing the χ^2 statistic with one or more

practical fit indices. The intent is similar to the American Psychological Association's (2010) recent addition to publication requirements, suggesting that null-hypothesis significance tests be supplemented with effect sizes and confidence intervals. For example, GFI* (gamma hat) can be used to indicate the degree to which model-implied predictions of sample moments correspond with observed sample moments, in a proportion metric. The RMSEA can be used to estimate the amount of discrepancy between true and hypothesized models per *df*. The SRMR can be used to indicate the average amount of discrepancy between observed and model-implied correlations. The CFI and TLI can be used to indicate the degree to which the model fits better than a baseline model that assumes every variable is an independent factor. Some of these measures (or functions of them) can also be used for model comparison, discussed in the Model Comparison section.

Hu and Bentler (1998, 1999) proposed a two-index strategy for evaluating model fit. Their simulations suggested that SRMR was more sensitive to misspecification in the structural model, whereas RMSEA, CFI, TLI, and Gamma Hat were more sensitive to misspecification in the measurement model. However, Fan and Sivo (2005) demonstrated that this was an artifact of their simulation—Hu and Bentler's measurement-model misspecifications had smaller effect sizes (noncentrality parameters) than their structural-model misspecifications. When those effect sizes were held constant, SRMR showed no differential sensitivity, negating the justification for a two-index strategy. Fan and Sivo (2009) did reveal that certain indices were more sensitive than others to misspecification in the mean structure (namely, Gamma Hat and Mc), but they were also so sensitive to model size that useful cutoffs would be difficult or impossible to propose—a finding they also found applies to covariance structure misspecification (Fan & Sivo, 2007). Deciding whether a model fits well to observed data in any absolute sense is therefore

difficult or impossible. But as Marsh, Hau, and Wen (2004) stated, fit indices were never intended to be used for hypothesis testing.

Model modification. If the global fit of a model is judged to be insufficient, then a researcher can either (a) reconsider the underlying theory to formulate a new model of the phenomena of interest or (b) attempt to identify the reason why the target model does not fit well and modify it in an ad hoc fashion to address the source of misspecification. The former consumes time and effort that the researcher has already spent formulating the original target model, and like any other human being, many researchers might not be easily convinced by the evidence (i.e., data) that their theories (i.e., models) are incorrect, at least not entirely.

This is perhaps why the latter alternative is more common practice, but once researchers use clues in the data to modify a hypothesized model, they no longer operate in a confirmatory framework, but an exploratory framework. There is nothing wrong with doing so, if this fact is openly reported along with the results. Exploratory research is useful for generating hypotheses, which can then be confirmed or disconfirmed using future, independently sampled data.

Models can be modified in a build-up or tear-down fashion. In a build-up approach, a restricted model is fit initially, in which only theoretically necessary parameters are freely estimated. If the initial model is judged to fit the data poorly, additional free parameters are added in a sequential fashion until acceptable fit is achieved. A tear-down approach begins with as unrestricted a model as one can identify, and proceeds to fix parameters sequentially. The tear-down approach is perhaps more commonly used in a model comparison framework (e.g., tests of measurement invariance), when a series of competing nested models are identified a priori and fit sequentially to identify whether each set of constraints is plausible. A data-driven build-up approach is a form of model-modification that MacCallum (1986) referred to as a

specification search.

Tools for model modification. Because of their availability in most SEM software, the most popular tools for model modification are modification indices (MI), expected parameter change (EPC), and residuals. Residuals are discrepancies between each element in $\boldsymbol{S}$ and its corresponding element in $\hat{\Sigma}$. Whereas RMR indicates the average discrepancy between covariance elements (and SRMR indicates the average discrepancy between correlation elements), the full matrix of residuals can reveal which specific observed relationships are not adequately characterized by the model. Large residuals occur most frequently for pairs of variables that are not directly related in the model.

For example, in a two-factor CFA, the indicators of the first factor are typically only related to indicators of the second factor indirectly (i.e., via the factor correlation). But if the two factors represent disorders with some similar symptoms, then the correlation among those symptoms (indicators) would be higher than could be explained merely by the correlation between the disorders (factors) that are the cause of those symptoms. Thus, the standardized residual would be large, indicating a local source of misfit that a researcher might conclude is an indication of misspecification. Because there is a theoretical explanation for why the residual is large, the researcher would be justified in reformulating the model in some way. If the symptom description is nearly identical for both disorders, the researcher might include only one of the indicators (or average of the two) and allow it to be an indicator of both factors. If the similar symptoms are not nearly identical, the simpler solution might be to let both remain as indicators of their separate disorders, but freely estimate their residual correlation (i.e., postulate that they are related in some way additional to what can be accounted for by the correlation between their respective disorders).

MI and EPC do not refer directly to discrepancies between $\boldsymbol{S}$ and $\hat{\Sigma}$, but to a function of them. To identify the model, certain (in fact, most) structural parameters are fixed to specific values so that other model parameters (i.e., elements of Λ, Φ, Θ, and B) can be freely estimated (identification rules can be found in Brown, 2006). Parameters may be fixed to identify constructs (e.g., fixing the mean and scale of a latent construct to 0 and 1, respectively) or because theory leads researchers to hypothesize certain direct effects to be negligible (i.e., each indicator measures only one construct in a CFA, rather than all constructs being allowed to affect all indicators, as in EFA).

Recall the example of measuring mental disorders—when the residual covariance between two similar symptoms is freed in the modified model, there is one less *df* that was given up to estimate that parameter. When a parameter is freed, the χ^2 statistic will always decrease, indicating better fit to the data. The difference between χ^2 statistics from nested models fit to the same data is also distributed as a χ^2 random variable, with *df* equal to the difference in *df* due to freeing the parameter(s). This provides a significance test of whether the amount of decrease is significant. The χ^2 difference test ($\Delta\chi^2$) is discussed in greater detail in the Model Comparison section, but an introduction is relevant here because a MI for a fixed parameter is an estimate of the amount that the χ^2 statistic would change if that particular parameter were freely estimated, holding all other parameters constant (Sörbom, 1989). Likewise, the EPC is an estimate of how much the parameter itself would change if it were estimated instead of fixed at a certain value (Saris, Satorra, & Sorbom, 1987).

The use of MI is straight-forward. All major SEM software packages provide a MI for each fixed parameter of the model. In a typical specification search (MacCallum, 1986), the largest MI is identified, and if it meets some criterion for significance (e.g., greater than 10, or

significant at the $\alpha = .001$ level), then the model is specified with that parameter freely estimated. The new model is fit to the data, and if model fit is still deemed inadequate, the process is repeated by freeing the next highest MI, until the global model fit is acceptable. This is a very exploratory process that tends to over-fit models to nuances in the data, making the models less generalizable to future observations.

MacCallum (1986; MacCallum et al., 1992) found that specification searches using MI only tend to lead researchers to a true model when the model they started with is already close to the true model and it is being fit to a sample of $N > 300$. A fixed parameter's MI is calculated on the assumption that all other fixed and estimated parameters will remain at their current values (i.e., the model is otherwise correctly specified), which is unrealistic for two reasons: (a) a misspecified parameter may cause other parameter estimates to be biased and (b) there may be more than one misspecified parameter. Because MIs are themselves only estimates of the expected change in the χ^2 statistic, they are subject to sampling variability. Therefore, the order in which parameters are freed fluctuates from sample to sample, sometimes resulting in parameters being freed that should remain fixed. For this reason, MacCallum (1986; MacCallum et al., 1992) suggested that the best method would be to identify a priori competing models and compare them directly, rather than making post hoc changes to improve fit of a single hypothesized model.

To improve the consistency of model modifications, Saris et al. (1987) proposed incorporating EPC when deciding whether to free a parameter with a large MI. They demonstrated that MIs are more sensitive to some model parameters than others, so the MI might be large for a parameter whose value might change very little if freed (i.e., small EPC), whereas a parameter whose EPC is large might have a small MI. They suggested that a parameter should

be freed if both the EPC and MI are large, but should not be free when EPC is small, regardless of whether the MI is large. If the EPC is large and the MI is small, however, then it is unclear whether to free the parameter because its large EPC might be due to sampling fluctuation.

Kaplan (1990) suggested extending Saris et al.'s proposal to consider reasons for sensitivity of large MIs, including missing data, violated distributional assumptions, and power considerations. Saris, Satorra, and van der Veld (2009) incorporated Kaplan's (1990) suggestion into a revision of their original (Saris et al., 1987) MI–EPC method. Saris et al. (2009) incorporated a power analysis for the MI test, so that researchers can (a) evaluate the magnitude of an observed MI in light of its power to detect misspecification in that parameter, and subsequently (b) decide whether to consult the EPC for additional evidence. If an MI test has low power but the observed MI is significant, the parameter can be confidently freed. If instead an MI test has high power but the observed MI is nonsignificant, then there is no justification to free that parameter. When the observed MI is high but the test has high power, the test alone is inconclusive because it might merely be sensitive to that parameter, in which case only if the EPC is also large should that parameter be freed. When the test has low power and the observed MI is low, then the test in inconclusive—there is no evidence that a parameter should be freed, but that might be due to low power to detect misspecification. The sampling variability of EPC is too high to provide information about whether a parameter should be freed in this case.

Even Saris et al.'s (2009) more integrative approach fails to overcome the main limitations of MI and EPC: they are estimates (hence, subject to sampling variability) based on the assumption that the rest of the model is correctly specified. Their utility thus appears dubious at best. MacCallum et al.'s (1992) revealed that MI-directed model modifications rarely lead to models that resemble the true model, and that modifications made to an incorrect model

vary wildly from sample to sample, and more recent research (e.g., Whitaker, 2012) provide no evidence that their main limitations are overcome by using both MI and EPC. MacCallum's (1986) long-standing advice—to formulate competing models a priori, comparing them rather than modifying an original model in a data-driven, post hoc manner—still appears pinnacle.

Model comparison. Model comparison is moot when modifying an initial model because data is used to provide post hoc clues about what significant changes could be made to create a new model from the initial model. Specifying a priori models is a more robust approach because theoretical uncertainty is taken into account ahead of time, rather than "fishing" for better results after the target model has been fit to data, making it impossible to infer whether the model is being adjusted merely to fit nuances of a particular sample. When competing specifications are specified a priori, researchers can be more confident in their results (MacCallum et al., 1992).

To specify competing models to compare to a target model, a researcher should anticipate how their model might be insufficient. This might entail identifying indicators that could measure more than one construct in the SEM, and specifying competing models that allow correlated residuals or cross-loadings to take that into account; identifying specific effects that might be so negligible they could be fixed to zero (or vice versa); or reversing the role of predictor and outcome among a pair of constructs. These models would represent different (sets of) hypotheses derived from the same theory, but competing models could also be specified to represent distinct competing theories (e.g., common-factor model vs. network perspective of mental disorders; Cramer, Waldorp, van der Maas, & Borsboom, 2010).

Tools for model comparison can be roughly divided into two categories, depending on whether the competing models must be nested to use them. Model A is nested within Model B when Model A has all of its free parameters in common with Model B, but Model B freely

estimates at least one additional parameter that Model A does not. That is, the entire set of Model A's parameters are a subset of Model B's parameters. More generally, Model A is nested within Model B if Model B can precisely reproduce any model-implied moments that Model A can (Bentler & Satorra, 2010). Nonnested models may have parameters in common as well, but both models would estimate at least one parameter that the other model does not. The following sections discuss tools for nested and nonnested model comparison, with examples to illustrate their use.

Nested model comparisons. Nested models can be compared using a test statistic. The test statistic in (15) for an individual model is distributed as a χ^2 random variable with *df* equal to the number of observed sample moments minus the number of free parameters. This statistic tests the null hypothesis that the target model perfectly explains the sample data, so the only source of discrepancy is sampling error (Browne & Cudeck, 1992). The statistic in (15) can also be calculated as the $-2 \times$ log(likelihood) of the target model minus the $-2 \times$ log(likelihood) of the saturated model, in which all sample moments are freely estimated, resulting in perfect fit with zero *df*. Thus, the χ^2 fit statistic for an individual model is equivalent to a $\Delta\chi^2$ statistic comparing the fit of the target model to the perfect fit of the saturated model. This statistic is distributed as a χ^2 random variable because any overidentified model is nested within any saturated model.

A $\Delta\chi^2$ statistic can be computed for any other pair of nested models, as well. It is the difference between the χ^2 of the more restricted model (i.e., worse-fitting because it estimates fewer parameters, having more *df*) and the χ^2 of the less restricted model. Likewise, Δdf for the $\Delta\chi^2$ statistic equals the difference between *df* for the more restricted and less restricted models, or equivalently, the number of additional parameters estimated in the less restricted model. The

null hypothesis for the $\Delta\chi^2$ test is that the nested models are equivalent because the additional parameter(s) can be constrained to a fixed value, typically zero (e.g., means, regressions, correlations) or one (e.g., variances). Like the χ^2 test of perfect absolute model fit, the $\Delta\chi^2$ statistic is overly sensitive to negligible discrepancies when the sample size is large, so H_0 might be rejected even when the constraints are approximately tenable.

Some model-comparison procedures have been formulated specifically to test sets of SEM constraints in a nested sequence, such as the four-step procedure (Mulaik & Millsap, 2000) and tests of measurement invariance (Cheung & Rensvold, 2002). The four-step procedure specifies a sequence of nested models, where Model 1 is nested in Model 2, which is nested within Model 3, which is nested within Model 4. These four nested models are specified to test certain hypothesized constraints in the target model, which is Model 2 (Mulaik & Millsap, 2000). For example, a target SEM might regress an outcome on three predictors, one of which also mediates the relationship between the outcome and the two other predictors. Model 2 would be nested within Model 3, which is specified as a CFA model (i.e., all correlations are freely estimated among the constructs in Model 2). Model 3 would be nested within Model 4, which is specified as an EFA model (i.e., all indicators load on all factors). Acceptable global fit of the EFA model confirms the number of hypothesized factors is correct; a nonsignificant $\Delta\chi^2$ test between Models 1 and 2 indicates the measurement model is correctly specified; and a nonsignificant $\Delta\chi^2$ test between Models 2 and 3 indicates the hypothesized structure among latent variables is tenable. Step 4 involves specifying an even more restricted Model(s) 4 than the target Model 3 by using the $\Delta\chi^2$ statistic to test whether parameters hypothesized to be substantial in Model 3 could in fact be constrained to zero in Model(s) 4.

To address the dependence on sample size of $\Delta\chi^2$ statistics, Cheung and Rensvold (2002)

investigated the behavior of changes in alternative indices of fit, such as CFI, RMSEA, and SRMR. They used Monte Carlo methods to estimate the sampling distribution of 20 fit indices and their changes between nested models in the context of testing whether measurement parameters (factor loadings, indicator intercepts, and residual indicator variances) are invariant across groups or measurement occasions. They proposed cutoff criteria for certain indices with small Type I error rates: ΔCFI < 0.01, ΔGamma-Hat < 0.001, and ΔMc < 0.02. Taking power into account, Meade, Johnson, and Braddy (2008) proposed a stricter ΔCFI < 0.002 criterion, and noted that ΔMc was inconsistent across different types of models. Chen (2007) proposed using multiple indices (e.g., ΔCFI < 0.005 in conjunction with ΔRMSEA > 0.01 or ΔSRMR > 0.025), but these rules varied across sample sizes.

Tests of measurement invariance involve a nested sequence of models named according to the constraints they test. Configural or "form" invariance represents the hypothesis that the pattern of fixed and freely estimated measurement parameters is identical across groups and measurement occasions, and global model fit is used as criterion (e.g., the χ^2 statistic or an index of fit such as CFI). Metric or "weak" invariance represents the additional hypothesis that the factor loadings are equivalent across groups and occasions; scalar or "strong" invariance represents the additional hypothesis that indicator intercepts are equivalent across groups and occasions; and "strict" invariance represents the additional hypothesis that residual variances (and therefore total indicator variances) are equivalent across groups and occasions. These more restricted subsequent models are evaluated using the $\Delta\chi^2$ statistic or change in a fit index (e.g., ΔCFI), as described above.

Sometimes model comparison leads to model modification—if the H_0 of weak invariance is rejected, then at least one of the factor loadings differs across groups or occasions. Similar to

post hoc tests in ANOVA, individual follow-up constraints could be specified to test which parameters are invariant. As long as partial weak invariance can be established, the scales of latent variables can be compared between groups and occasions (Muthén & Asparouhov, 2013). Likewise, if the H_0 of strong invariance is rejected, it would be necessary to establish partial strong invariance to compare latent means (Muthén & Asparouhov, 2013).

Nonnested model comparisons. A model that estimates more parameters is expected to fit data better than a nested model that estimates fewer parameters, necessitating a $\Delta\chi^2$ test to ascertain whether fit improves substantially due to the additional parameter(s). When models are not nested, Model A's parameters are not merely a subset of Model B's parameters, nor vice versa. It is possible that the models are so different that their parameters do not have the same interpretation. Even if the models are similar, it is possible for nonnested models to have the same *df* but not be equivalent. In such cases, calculating $\Delta\chi^2$ would not yield a quantity that is distributed as a χ^2 random variable. Thus, applying a $\Delta\chi^2$ test is not always appropriate for model comparison, or at least not as straightforward.

Levy and Hancock (2007) provided a framework for using a set of $\Delta\chi^2$ tests to compare nonnested models. In rare cases, competing models are so different that they do not share parameters with a common interpretation (e.g., network vs. latent variable models; Cramer et al., 2010). But in many cases, competing models are similar enough that common parameters between the two can be identified (e.g., CFA with cross-loadings vs. CFA with correlated errors). For example, a researcher might presuppose that the target model is insufficient, motivating the a priori specification of alternative models with additional parameters that are hypothesized to remedy the target model's potential deficiency. The target model would then be a restricted model that is nested within the less restricted alternative models, each of which

specifies different additional free parameters; thus, the competing models would not be nested within each other, except that the target model is nested within all alternative models.

In such a case, the nonnested alternative models could be compared indirectly via their respective comparisons with their common restricted model (Levy & Hancock, 2007). After calculating $\Delta\chi^2$ between each alternative model and the common model, the alternative models are considered distinguishable if only one $\Delta\chi^2$ is significant, in which case (a) the model without a significant $\Delta\chi^2$ would be indistinguishable from the restricted target model, and (b) the model with a significant $\Delta\chi^2$ can be assumed to fit better than the competing alternative model because it fits significantly better than the target model. If neither model's $\Delta\chi^2$ is significant, then neither alternative model can be distinguished from the restricted model, and the restricted model is to be preferred because it fits as well as the alternatives but with fewer parameters. If both alternative models have significant $\Delta\chi^2$, then they both fit better than the common model, but the competing models cannot be further distinguished using $\Delta\chi^2$, so they must be compared using other criteria.

Bentler and Bonett (1980) enumerated ways to compare models using incremental fit indices such as CFI and TLI. To compare nested or nonnested models, it is necessary to identify (a) a saturated model in which all competing models are nested and (b) a poorly fitting null model that is nested within all competing models. Typically, the saturated model is specified by freely estimating means, variances, and covariances among all observed variables as though they were all distinct factors in a CFA, leaving no degrees of freedom and $\chi^2 = 0$. The default specification of a null model in most software (e.g., EQS, LISREL, M*plus*, lavaan) is typically an independence model, which (like the saturated model) considers each observed variable a distinct factor, but independent of (i.e., uncorrelated with) all other variables. Widamin and

Thompson (2003) illustrated many common research scenarios in which the independence model is insufficient as a null model because it is not nested within all competing alternatives—these include invariance tests across multiple groups or occasions, as well as latent growth curve models in which at least one model hypothesizes homoscedasticity of residuals.

Once an appropriate null model is identified, which is as unrestricted as possible yet nested within all competing models (Widamin & Thompson, 2003), a continuum from poor fit to perfect fit is established by the null and saturated models, respectively. All competing models can be evaluated by locating their incremental fit indices on that continuum (Bentler & Bonett, 1980, p. 600). The model with the highest index is to be preferred because its fit is closest to the fit we would expect for the correct model, if it were known. Incremental fit indices could even be used in a cross-validation context, by comparing the fit of the same model to different data sets—when the sample size is unequal, χ^2 values could not be compared, but CFIs could (Bentler & Bonett, 1980, p. 600).

Information criteria defined in (20) are designed specifically for model comparison, and it is not necessary for competing models to be nested. Comparing χ^2 values between competing models would indicate which model fits the data better in an global sense. If the competing models estimate the same number of parameters, they also have the same *df*, which is the expected value of χ^2, so lower χ^2 values would be preferred. But if the competing models have different *df*, χ^2 values cannot be directly compared. Information criteria supplement the χ^2 value—or discrepancy function, which has the same rank order, shown in (15)—with an adjustment for the number of parameters estimated in the model. The more parameters are estimated, the more the estimated fit is decremented, so additional parameters must be of substantial importance to justify their inclusion.

Among a set of competing models that are fit to the same data, the fitted model with the lowest information criterion is to be preferred (Gelman et al., 2013). Information criteria do not follow theoretical distributions, so it is impossible to interpret a difference (e.g., ΔAIC or ΔBIC) as an effect size or to calculate an associated probability of observing the difference under H_0 that the models balance fit and parsimony equivalently well. Lower values are merely interpreted as demonstrating a more efficient tradeoff of fit and parsimony.

Because the punishment term in (20) is defined differently for each information criterion, different criteria behave differently in practice (Vrieze, 2012). AIC uses a constant multiple of the number of estimated parameters, whereas BIC weights the number of parameters by the log of the sample size. Thus, BIC adjusts for additional parameters more harshly in larger samples. When the true population model is one of the competing models under consideration, BIC tends to select the true model (Bollen, Harden, Ray, & Zavisca, 2014; Vrieze, 2012), but researchers should not assume the true model is a contender (MacCallum, 2003). Even as sample size increases, the sampling variability of BIC is so erratic that there is no single model that will be preferred asymptotically (Preacher & Merkle, 2012). AIC does not consistently choose the true model even when it is under consideration, but it does tend to choose the model that minimizes discrepancies between observed and predicted values (Vrieze, 2012).

Information criteria have also been criticized because their adjustment for parsimony only takes into account the number of free parameters (Preacher, 2006). Though adding parameters to a model increases its ability to fit well to a range of data patterns, the functional form of a model also affects how well a model fits data patterns. Two models with the same degrees of freedom but different functional forms (e.g., a simplex model vs. a factor model) may have different *fitting propensity*, which is what Preacher (2006) termed the ability of a model to

fit random data, regardless of whether it is the true data-generating model. Models are less parsimonious if they are more likely to fit data wholly unassociated with it.

Preacher (2006) reviewed rarely used fit indices that would take into account functional form as well as the number of parameters. One such index is the stochastic information criterion (SIC) of the same form in (20), but the punishment term for parsimony is a function of the Fisher information matrix of model parameters. The more redundancy among parameter estimates indicated by the information matrix, the greater the decrement to model fit. Other indices are the uniform index of fit (UIF) and normalized maximum likelihood (NML), both of which are calculated with Monte Carlo methods. Thousands of random data patterns are simulated, to which competing models are fit, and the observed discrepancies of the competing models are compared with respect to their distributions of discrepancies. Using this method, all aspects of competing models' ability to fit any data pattern is implicitly taken into account, so these indices adjust for parsimony in a potentially more comprehensive way, but little research has been conducted to investigate their behavior.

Bayesian SEM

Before discussing the evaluation, modification, and comparison of Bayesian models, it is necessary to explain how Bayesian estimation of SEM parameters differs from CSA. This explanation will include a conceptual introduction to Bayesian statistical inference and its contrast with frequentist inference.

Bayesian statistical inference. In the traditional frequentist paradigm, inference about a population parameter typically involves null hypothesis significance testing (Gelman & Stern, 2006, Gelman & Shalizi, 2013a). Complementary null and alternative hypotheses are specified, which together account for the entire parameter space, and the tenability of the null hypothesis

(H_0) is judged according to the likelihood of observing the sample data on the premise that H_0 is true. There are many ways to calculate the likelihood that is used to judge the statistical significance of the data. Typically, a point estimate is calculated for the parameter to be tested (e.g., the corresponding sample statistic or ML estimate) and transformed into an inferential statistic with a known sampling distribution under H_0, if certain assumptions hold. Alternatively, an interval estimate is calculated, and H_0 is judged to be plausible if the interval contains it.

In the frequentist paradigm, parameters are seen as fixed constants, whereas data are variable. The term "frequentist" came about because an inference is drawn about a parameter with reference to how frequently the observed data could be expected to occur under a certain premise about the unknown parameter. This can lead to awkward interpretations of results (Iversen, 1984). For example, the *p* value—$p(Y \mid \theta_0)$—is the probability (p) of the observed data (Y), or the frequency with which it should occur, on the condition that the unknown parameter (θ) is consistent with H_0. Likewise, 95% confidence intervals do not indicate that the interval estimate is 95% likely to contain θ; rather, the method of calculating the interval will successfully capture θ in 95% of samples drawn from the same population.

Bayesian statistical inference utilizes the same information, but it does not stop with the calculation of a likelihood $p(Y \mid \theta)$. In the Bayesian paradigm, information about θ comes not only from the data, but also from a researcher's collection of prior experience, expert judgment, and theoretical expectation (Iversen, 1984). This "prior" information is translated into a statistical summary called the prior probability distribution—$p(\theta)$—and its role in Bayesian inference is to represent what the researcher believes about the population parameter (θ) before observing any evidence from data (Y). The unknown parameter is thus considered to be variable rather than fixed, and the observed data is treated as fixed. This does not mean that Bayesians

disregard sampling variability. Bayesians merely treat known quantities (observed data) as constant to make probabilistic statements about unknown quantities (parameters), whereas frequentists make inferences about parameters indirectly and unintuitively via probabilistic statements about the observed data, conditional on a fixed but unknown H_0 parameter.

Bayesian inference culminates in a posterior probability distribution—$p(\theta \mid Y)$—which represents what the researcher concludes about θ after observing data. The posterior distribution is the entirety of a Bayesian inference (Iversen, 1984), and it is calculated using both prior belief and evidence from observed data. In fact, posterior probability of a parameter is proportional to the product of the likelihood of the data and the prior probability of the parameter:

$$p(\theta|Y) \propto p(Y|\theta) \times p(\theta) \tag{21}$$

Conceptually, researchers can begin with prior beliefs about phenomenon of interest, collect data to obtain more information about that phenomenon, and allow the evidence to update their beliefs or change their minds entirely. Thus, the likelihood is literally the weight of the evidence that changes the prior distribution into the posterior distribution, and larger sample sizes translate into greater weights of evidence that can completely overwhelm any prior belief.

Bayesians need not consider a prior distribution to be a formal representation of actual beliefs, but rather a summary of assumptions about the relative likelihoods of possible values for θ (Gelman & Shalizi, 2013a). Gelman and Shalizi (2013a, 2013b) consider priors to be model assumptions like any other, such as the distribution of errors or the function form of the relationship between predictors and outcome. In practice, priors need not be informative at all; they can be uniform distributions with disparate upper and lower limits, indicating that the true parameter could be almost any possible value, all of which seem equally likely to the researcher. Uninformative priors such as these place all of the weight of estimation on the shoulders of data,

giving priors an even smaller influence than they usually do in samples of substantial size.

Thus, the link between frequentist and Bayesian statistical inference can be made by viewing frequentist inference as a special case of Bayesian inference. In the frequentist paradigm, inference is made using only the likelihood of the data under H_0. In the Bayesian paradigm, the use of uninformative priors yields posterior distributions that are identical in form to the likelihood, so Bayesian inferences would be identical to frequentist results. For example, the mode of the posterior distribution would be the ML estimate, and the standard deviation of the posterior distribution would be its *SE*.

Even in this special case, however, the interpretation of results under the two paradigms would differ. Rather than calculate the probability of the data under H_0 (i.e., the frequentist *p* value), the posterior distribution allows a researcher to infer for example, the probability that θ is greater (or less) than a null-hypothesized value θ_0. Rather than interpreting frequentist interval estimates as the probability of the method of estimation to capture the unknown θ—making no probabilistic statement about whether a particular interval estimate did so—a Bayesian 95% credible interval indicates much more intuitively that the true θ is 95% likely to be within the upper and lower bounds.

Estimating Bayesian models. Bayesian methods for estimating model parameters involve simulation—namely, Markov Chain Monte Carlo (MCMC) methods, which operate iteratively using algorithms such as Gibbs sampling (see Gelman et al., 2014, chapter 11). The current state of a Markov chain depends only on the previous state. Sampling algorithms draw parameters sequentially at each current state by updating draws from the previous state. Once the current state is updated, the next iteration begins. If the model is appropriately identified, Markov chains eventually converge on a stable distribution, which is the joint posterior

distribution of model parameters. All subsequent iterations in the Markov chain can be treated as random draws from the posterior, and a large sample of them should adequately represent the posterior distribution, allowing a researcher to summarize the posterior using the simulated values.

In this framework, all unknown quantities can be drawn from the joint posterior, not just the model parameters. For example, missing data for an observation can be imputed by drawing values from the posterior, conditional on observed variables for that observation. Prediction intervals for hypothetical future observations can be simulated by drawing their predicted values from the posterior. Latent variables can be drawn from the posterior as well, so SEMs are not limited to being estimated in a CSA framework, giving BSEM numerous advantages over traditional CSA. For example, CSA cannot directly estimate interactions among latent variables, although a full-information method called latent moderated structural equations (LMS; Klein & Moosbrugger, 2000) is currently implemented only in M*plus*. However, interactions among latent variables are easily handled in BSEM because product terms can be calculated when latent variables are drawn from the posterior.

Bayesian Model Fit

Traditional SEMs are evaluated with a set of tools uniquely developed for CSA, not for evaluating other types of model (e.g., general linear models[5]). BSEM, however, is not limited to being specified as an analysis of covariance structure, so tools for evaluating BSEMs are the same tools that were developed for evaluating Bayesian models in general. Unless it is explicitly

[5] Although one can specify a general linear model using SEM software in order to obtain SEM fit measures, such models are typically saturated in terms of covariance structure. There is a single outcome (or correlated set of outcomes in a multivariate model) that is related to all predictors. This means that even if the linear model has several residual *df*, the SEM specification will have perfect fit with $df = 0$ because linear models base total *df* on sample size, whereas SEMs base *df* on the number of observed sample moments. Thus, covariance structure fit measures for general linear models would not be informative.

stated otherwise, when I refer below to a method for evaluating a BSEM, the reader can assume the same method can apply to other types of Bayesian model. I will focus on the same three categories that were the focus of evaluating CSA models: evaluating global fit of an isolated model, modification of a model to improve fit, and comparison of competing models.

Model evaluation. The frequentist *p* value allows researchers to test whether their data are consistent with H_0, which in the case of the χ^2 test statistic in (15) is the hypothesis that the target model corresponds perfectly to the true population model. Frequentist statistics treat parameters as a fixed quantity—be it a scalar, vector, matrix, or an array of scalars, vectors, or matrices (as in multiple-group SEM)—so the *p* value is calculated holding the H_0 quantity fixed. In BSEM, the parameters vary and are characterized by a posterior distribution, so the *p* value would also vary across the posterior. There are numerous ways to calculate a single *p* value that takes the posterior distribution into account (Levy, 2011), the most popular of which involves posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996).

The motivation behind PPMC is identical to traditional hypothesis testing: to test whether the observed data are consistent with H_0 (i.e., consistent with the target model); however, unlike traditional hypothesis tests of null hypotheses, PPMC tests H_0 using simulation methods, capitalizing on the MCMC process of sampling model parameters from the joint posterior distribution. A simulated data set (Y_{rep}) of the same size as the observed data (Y_{obs}) is generated for each "sample" of parameters from the posterior ($\boldsymbol{\theta}^i$, the vector of parameters at iteration *i* in the Markov chain). Whether the observed data are consistent with H_0 can then be tested by checking whether Y_{obs} resemble data generated under H_0 (i.e., Y_{rep}).

Resemblance between Y_{obs} and Y_{rep} can be defined in any number of ways (Gelman et al., 1996; Levy, 2011). At each iteration in the Markov chain, the test statistic in (15) can be

calculated for both Y_{obs} and Y_{rep}, as could SRMR or any of the fit indices in (16)–(19). To compare the fit of the model to both Y_{obs} and Y_{rep}, a Bernoulli random variable is assigned a value of 1 if the model fits better to Y_{obs} and 0 if the model fits better to Y_{rep}. This Bernoulli random variable has an expected value of 50% when the model fits Y_{obs} well because both Y_{obs} and Y_{rep} are consistent with H_0. However, the less adequate the model is at capturing aspects of the data, the lower its expected value becomes, due to the fact that Y_{rep} remains consistent with H_0. High PPP values may indicating the model is overfitting the data.

The observed proportion of MCMC iterations that yield better fit for Y_{obs} is called the posterior predictive *p* value (PPP; Gelman et al., 1996), which is an estimate of the probability ($\hat{\pi}$) that the data (Y_{obs}) are consistent with H_0. Thus, low values provide evidence that H_0 is untenable as an explanation for the data. When used to make a binary decision about whether to reject H_0—as in traditional null-hypothesis significance testing (Gelman & Stern, 2006)—PPP tends to have Type I error rates lower than α levels set by the researcher. Bayarri and Berger (2000) proposed conditional and partial PPP values, which yield nominal Type I error rates but are more computationally intensive, less flexible than PPP, and apply only in the context of binary-decision-making null-hypothesis significance tests.

Gelman and Shalizi (2013a, 2013b; Gelman et al., 2014) do not advocate using PPP for traditional hypothesis testing, but rather as a diagnostic tool to identify whether and how a model can be improved (see also Kruschke, 2013; Morey, Romeijn, & Rouder, 2013). I discuss model modification in the following section, but I note here that the developers of PPP (Gelman et al., 1996) did not intend for it to be used as a test statistic. Muthén and Asparouhov (2012) assert that PPP should be treated as an index of practical fit, similar to those defined in (16)–(19), although they nevertheless suggest that "using [PPP] values of .10, .05, or .01 appears

reasonable" (p. 315).

There has been little development of methods to evaluate isolated Bayesian models in terms of data–model fit. Johnson (2004) proposed a Bayesian χ^2 goodness-of-fit statistic, whose properties are quite similar to PPP (i.e., evaluation of the fit statistic across the posterior). The PPP is the most developed method of evaluating global fit, probably because of its flexibility. Any discrepancy measure can be used to compare the fit of two models, although software might only provide PPP based on the χ^2 statistic (this is the case in M*plus*; Muthén & Asparouhov, 2012; Muthén & Muthén, 2012). For instance, Levy (2011) evaluated the posterior predictive distribution of SRMR along with the χ^2 statistic. The appropriateness of specific aspects of the model could be tested, rather than the model as a whole. In fact, prior predictive model checking predates PPMC (Gelman et al., 1996; Levy, 2011), and it can be used to evaluate the appropriateness of the chosen priors (i.e., whether the prior distribution generates data that resemble the observed data).

Model modification. If a low PPP based on the overall model–data discrepancy (quantified by χ^2) indicates global misfit, then further investigation is needed to discover why the model is inadequate. Local sources of misfit might include inappropriate function form of relationships (e.g., linear vs. curvilinear slopes, additive vs. interactive effects among multiple predictors) or the omission of an important relationship. Methods for identifying local misspecification appear to be less developed than methods for identifying global misspecification, although PPMC is flexible enough to be employed for identifying either global or local misspecification.

Gelman and Shalizi (2013a) advocate plotting the raw data along with the line of best fit implied by the model, a diagnostic that applies equally well in frequentist and Bayesian

frameworks. Kruschke (2013) and Morey et al. (2013) indicate the PPMC can be used to generate alternative models that can then be compared to the original target model (I discuss model comparison methods in the following section). For example, Kruschke used a plotting method in accordance with PPMC to uncover a possible curvilinear effect. But this type of method would be impossible in any BSEM in which the predictors are unobserved latent variables. Checking residuals (i.e., $\boldsymbol{S} - \hat{\Sigma}$) as in traditional SEM would be possible, but quite difficult because there is a posterior distribution of parameters, and thus a posterior distribution of $\hat{\Sigma}$. The posterior distribution of a summary measure such as SRMR would be simpler to investigate, which is one quantity that Levy (2011) investigated with PPMC.

Fox and Glas (2005) previously proposed Bayesian analogs to MIs, but these were much more restricted in that they were statistical hypothesis tests, specifically developed for IRT models, and the two MIs they proposed were each specific to testing IRT parameters. Categorical factor analysis parameters can be transformed to IRT parameters, so Fox and Glas' proposal may yet be applicable to BSEMs in general.

Muthén and Asparouhov (2012, pp. 316–317) proposed a method for identifying local sources of misfit that is analogous to the use of MI in traditional SEM. Frequentist estimators such as MLE require several parameters to be fixed in order to identify the model. For example, setting scale of a construct requires fixing either the factor variance or one of the factor loadings to one, and unless there is theoretical justification for expecting nonzero values, residual correlations and cross-loadings are typically fixed to zero (Bollen, 1989). In a Bayesian context, this would be equivalent to specifying a completely informative prior, such as a normal distribution with $\mu = 0$ and $\sigma = 0$. Bayesian estimation thus allows a less restricted, but still informative, normal prior with $\mu = 0$, but with $\sigma = 0.10$ (which would indicate 95% of parameters

fall within the bounds ± 0.20). If the data differ sufficiently from this prior, the posterior will indicate a very low probability that the parameter is zero. In hypothesis testing language, if the 95% credible interval does not include zero, the H_0 that the parameter is zero may be rejected.

This is a very new proposal that has yet to be tested. The degree to which the likelihood of the data can overwhelm the specified prior (or vice versa) will affect the power of this method to detect local sources of misspecification. It is also unclear whether a nontarget parameter with a small-variance prior would behave any more reliably than a traditional MI in leading a researcher to specify a model that more closely resembles the true population. There are, however, at least two advantages of the Bayesian analog. First, the BSEM analog of MI is the parameter estimate, so in a Bayesian framework, the MI and the EPC would not be distinguished as they are in SEM. The more reliable but more complicated proposals to use both MI and EPC would thus not be necessary in BSEM. Second, MI is calculated assuming all other parameters remain fixed, so parameters can only be freed one-at-a-time, and this sequence of model modifications leads to overfitted models that rarely resemble the true model (MacCallum et al., 1992). In contrast, the BSEM analog is estimated jointly with all other parameters in the model, including all nontarget parameters that have small-variance priors. Thus, multiple potential modifications can be identified in a single step.

Model comparison. As in traditional SEM (MacCallum, 1986; MacCallum et al., 1992), Bayesian methodologists tend to advise specifying competing models a priori (e.g., Gelman et al., 1996). Gelman et al. (1996) proposed PPMC specifically to judge model fit in the absence of any alternative models, when it would still be necessary to check whether the model should be rejected—or at least to decide whether to modify the model or to specify alternatives. They developed PPMC to address this necessity because up until then, the best developed tool (dating

back to the 1960s; Kass & Raftery, 1995) for evaluating Bayesian models was only intended for explicit model comparison via posterior odds. This section is devoted to two broad classes of tools for comparing BSEMs: Bayes factors and information criteria.

Bayes factors. The posterior odds of a pair of models can be defined as the ratio of their posterior probabilities, each of which is the product of their respective likelihoods and priors.

$$\frac{p(\boldsymbol{\theta}_1|Y)}{p(\boldsymbol{\theta}_2|Y)} = \frac{p(Y|\boldsymbol{\theta}_1)}{p(Y|\boldsymbol{\theta}_2)} \times \frac{p(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_2)} \tag{22}$$

When the prior probability distributions are equivalent, the posterior odds ratio reduces to the ratio of the likelihoods, called the Bayes factor (Kass & Raftery, 1995).

$$\text{BF} = \frac{p(Y|\boldsymbol{\theta}_1)}{p(Y|\boldsymbol{\theta}_2)} \qquad \text{thus} \qquad \frac{p(\boldsymbol{\theta}_1|Y)}{p(\boldsymbol{\theta}_2|Y)} = \text{BF} \times \frac{p(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_2)} \tag{23}$$

Conceptually, the Bayes factor can be thought of as the quantity that changes the ratio of prior probabilities (representing a researcher's prior belief about how likely two models are to be true) into the ratio of posterior probabilities (representing the updated belief after seeing evidence provided by data). Because the models are fit to the same data, the likelihood ratio is a direct comparison of how much more likely the data are to have arisen from a population described by the model in the numerator than from the model in the denominator.

The Bayes factor is the most well developed and studied tool for Bayesian model evaluation, although it is not without limitations. Notably, it might not be possible for the joint prior distribution for all model parameters to be equal in both models, so it might only be possible to calculate Bayes factors that are conditional on individual parameters rather than Bayes factors for the model as a whole. Many applied researchers use uninformative priors that are improper (i.e., not conjugate) to let the data carry all the influence on the posterior, in which case Bayes factors are undefined (Gelman et al., 2014, p. 183). Even in situations when a Bayes

factor can be calculated, it is not included in standard output of any software package, so it would not be a straightforward task for researchers to calculate it. Bayes factors are thus less appealing than information criteria, which are frequently included in standard software.

In addition to the practical problems, Gelman et al. (2014) illustrate conceptual problems with Bayes factors; namely, Bayes factors are appropriate when the models being compared involve discrete parameters. However, most research situations involve continuous parameters. Even if the hypotheses being compared are fixed values in a continuous distribution (e.g., H_0: treatment effect is 0 vs. H_1: treatment effect is 1) and the prior odds are 1, they will be sensitive to aspects of the prior distribution, which is an undesirable characteristic. When working with continuous parameters, Gelman and Meng (1998) proposed path sampling as a method for approximating the Bayes factor by estimating the posterior across the range of the continuous parameter (e.g., in increments between 0 and 1, rather than only at the fixed hypothetical values). Like other methods of calculating the Bayes factor, path sampling is conceptually complex, difficult to program, and unavailable as a standard feature of any software package. Song and Lee (2012) provide examples of how to program path sampling in OpenBUGS software (Lunn, Spiegelhalter, Thomas, & Best, 2009), which would also work in JAGS (Plummer, 2013).

Information criteria. More recently, Bayesian versions of information criteria defined in (20) have been proposed. Although BIC stands for "Bayesian" information criterion, Gelman et al. (2013, 2014) stress that the name is misleading because it is not at all a Bayesian measure. BIC was originally developed as an approximation to the Bayes factor, calculated by excluding several additive (or multiplicative) terms that asymptotically approach zero (or one), making them unnecessary assuming sample size is close to infinity. Bollen, Harden, Ray, and Zavisca (2014) recently investigated the behavior of BIC and several alternative formulations that drop

fewer terms, making them better approximations of the Bayes factor. Their conclusions were similar to Vrieze (2012): BIC selects the correct model when it is among the candidate models, but chooses the simplest model otherwise.

Gelman et al. (2013, 2014) noted that because the goal of BIC is not to estimate a model's predictive accuracy on new data, it belongs to a different class of information criteria than the ones I discuss below. Bearing in mind that researchers should never expect to be able to specify a completely "true" model of a real population process in the social sciences (MacCallum, 2003), the utility of BIC is limited to situations in which researchers are simply looking for the simplest model in a set of competing models, not necessarily the one the "works" best by providing the most accurate predictions.

AIC is designed to estimate a model's out-of-sample prediction error (Vehtari & Ojanen, 2012), which it seems to do successfully in practice, at least on average (Vrieze, 2012). However, the calculation of AIC is problematic for hierarchical models, even in a frequentist framework. The punishment term in (20) for AIC is twice the number of parameters in the model, but in multilevel models each observation has an associated random effect of their Level-2 (or higher) unit. If the intraclass correlation coefficient (ICC) is zero (indicating no between-cluster differences), then all variability occurs at Level 1, so the number of parameters is the same as it would be defined in a single-level model. If ICC is one, then all variability occurs between clusters, so the number of parameters is increased by the number of clusters. Most situations are somewhere in between the two extremes of ICC, but it is unclear how to choose a single value for the adjustment in AIC.

This problem persists in Bayesian models, even if the model is not multilevel. When priors are completely uninformative (i.e., flat, uniform over the entire sampling space), Bayesian

estimates are equivalent to estimates derived using ordinary least squares or MLE, so the effective number of parameters is the same as would be defined under those frequentist methods. However, when informative priors are specified, the effective number of parameters in the model is decreased proportionate to the weight of information in the prior (i.e., weight that is removed from the shoulders of the data; Gelman et al., 2013, 2014).

Spiegelhalter, Best, Carlin, and van der Linde (2002) proposed a generalization of AIC called the deviance information criterion (DIC). Recall that deviance is calculated as $-2 \times$ log(likelihood) in frequentist estimation methods such as MLE, and it is distributed as a χ^2 random variable. Using Bayesian estimation, the likelihood of the data can be calculated at each iteration of the Markov chain because the parameters on which the probability of the data is conditioned are treated as variables. Thus, there is a posterior distribution of deviance statistics. The average of that distribution ($\overline{D}$) is used as the χ^2 component in (20), and the punishment term is the effective number of parameters (pD), defined as the difference between $\overline{D}$ and the deviance calculated at the posterior mean: $\mathrm{pD} = \overline{D} - D(\bar{\theta})$.

AIC is a special case of DIC, and they are asymptotically equivalent when using flat priors in a nonhierarchical model. A further generalization of AIC (and thus of DIC), called the Watanabe–Akaike (or alternatively, the "widely applicable") information criterion (WAIC; Watanabe, 2010), calculates the effective number of parameters in a more fully Bayesian manner (Gelman et al., 2013, 2014; Vehtari & Ojanen, 2012). It is new and relatively unstudied, but its behavior is promising as it is asymptotically equal to Bayesian cross-validation (Vehtari & Ojanen, 2012).

The equations below illustrate how the calculations differ. DIC calculates pD as twice the difference between (a) the log-likelihood of the entire data set evaluated at a point estimate

for the entire joint posterior (i.e., the mean vector for the joint posterior distribution) and (b) the average of the posterior distribution of log-likelihoods for the entire data set.

$$\mathrm{pD}_{\mathrm{DIC}} = 2\left\{\log\left[p\left(Y \mid \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\theta}^{i}\right)\right] - \frac{1}{M}\sum_{i=1}^{M}\log\left[p\left(Y \mid \boldsymbol{\theta}^{i}\right)\right]\right\} \quad (24)$$

The index *i* iterates over the range of *M* sampled vectors $\boldsymbol{\theta}^{i}$ from the joint posterior distribution in the MCMC process. Gelman et al. (2013) do not consider DIC to be fully Bayesian because it does not utilize the entire posterior to calculate the pointwise discrepancies for each observation. In contrast, WAIC calculates pD as a similar difference, but separately for each observation in the data set. That is, pD calculates the difference between (a) the log of the average of an individual's posterior distribution of likelihoods and (b) the average of the posterior distribution of that individual observation's log-likelihood—and twice the sum of those differences across all *N* individuals is pD.

$$\mathrm{pD}_{\mathrm{WAIC}} = 2\sum_{n=1}^{N}\left\{\log\left[\frac{1}{M}\sum_{i=1}^{M}p\left(Y_n \mid \boldsymbol{\theta}^{i}\right)\right] - \frac{1}{M}\sum_{i=1}^{M}\log\left[p\left(Y_n \mid \boldsymbol{\theta}^{i}\right)\right]\right\} \quad (25)$$

WAIC thus averages individual predictive discrepancies across the entire posterior, rather than the discrepancy of the entire sample conditional on a point estimate (i.e., average of the posterior). For this reason, Gelman et al. (2014, p. 173) describe WAIC as "a more fully Bayesian approach for estimating the out-of-sample expectation."

The concept of nested models is not as clear cut in the Bayesian framework as it is in the frequentist framework, since the parameters involved in calculating the likelihood are not the only parameters in the model. Priors also affect the posterior, and the more informative they are, the more less the effective number of Bayesian parameters resembles the number of parameters in a frequentist estimator (e.g., MLE). Regardless, nesting is of little consequence because tools for Bayesian model comparison do not rely on nesting—Bayes factors and information criteria

can both be used to compare nonnested models. However, information criteria are advantageous because they are easier to calculate, lack the noted limitations of Bayes factors, and are more readily available as standard output in software packages that provide Bayesian estimators: OpenBUGS (Lunn et al., 2009) and JAGS (Plummer, 2013) provide DIC, and M*plus* (Muthén & Muthén, 2012) provides both DIC and BIC. WAIC is not yet automatically calculated by any software, although Andrew Gelman (2013) has stated that it might be possible to implement it in a future version of the Bayesian software project, Stan (Stan Development Team, 2014).

Summary of Bayesian Model-Comparison Tools

Evaluation of traditional SEMs remains an area of active research in numerous contexts, such as establishing appropriate cutoff values for fit indices with no known sampling distribution (Curran et al., 2003; Fan & Sivo, 2005, 2007, 2009; Hu & Bentler, 1998, 1999; Marsh et al., 2004), invariance testing (Cheung & Rensvold, 2002; Chen, 2007; Meade et al., 2008), model selection (Bollen et al., 2014; Preacher, 2006; Preacher & Merkle, 2012; Vrieze, 2012), and model modification (Kaplan, 1990; MacCallum, 1986; MacCallum et al., 1992; Saris et al., 1987, 2009; Sörbom, 1989; Whitaker, 2012). It is therefore no surprise that the same topics in the context of BSEM are far from settled.

Bayes factors are an area of active research. Gelman et al. (2013) described the evaluation of hypotheses about continuous parameters as problematic, but van de Schoot, Hoijtink, Hallquist, and Boelen (2012) recently proposed a method to test hypotheses of inequality constraints against their complements using Bayes factors. Although it is not automated in any software, they provide instructions for the user to run the Bayesian model in the popular SEM software M*plus*, save the appropriate information, and calculate the Bayes factor manually. Morey and Rouder (2011) also recently extended the calculation of a Bayes

factor, but they generalized it to test interval hypotheses rather than point hypotheses.

Recent research on invariance testing in a Bayesian framework include novel approaches that do not require fitting the series of nested models as described in Cheung and Rensvold (2002). For example, Verhagen and Fox (2013) proposed specifying an IRT measurement model as a multilevel IRT model, in which responses to indicators of the same construct are considered repeated measures nested within individuals as well as within groups (across which the researcher wishes to test invariance of item parameters). This generalizes easily to other latent variable models, such that factor loading estimates would be random across groups, and items would be considered invariant if the random effect had variance close to zero (i.e., not significantly different from zero). Muthén and Asparouhov (2013) propose a two-stage approach in which differences among factor loadings between groups are first estimated using small-variance priors centered at zero, after which any group loading differences that are flagged as significantly different from zero are freed, while all others are constrained to equality. Both of these proposals seem promising, but are new and need to be validated using Monte Carlo simulations of various scenarios.

Gelman et al. (2014, pp. 172–173) provide simpler calculations of DIC and WAIC from the variance of the posterior distribution of log-likelihoods instead of differences between means:

$$\mathrm{pD}_{\mathrm{DIC}} = 2 \times \mathrm{Var}(\log[p(Y \mid \boldsymbol{\theta})]) \tag{26}$$

and

$$\mathrm{pD}_{\mathrm{WAIC}} = \sum_{n=1}^{N} \mathrm{Var}(\log[p(Y_n \mid \boldsymbol{\theta})]) \tag{27}$$

The calculations yield asymptotically equivalent results, but for WAIC, summing posterior variances across individual log-likelihoods results in greater computational stability than using differences. For DIC, however, (26) is less numerically stable, although it always results in a

positive estimate of pD, unlike (24). Further research is necessary to establish how discrepant these computational methods are when fitting latent variable models to finite samples. I will refer to DIC_1 and $WAIC_1$ when using the mean-deviations method in (24) and (25) to calculate pD, and DIC_2 and $WAIC_2$ when using the variance method in (26) and (27) to calculate pD.

Regardless of which computation of DIC and WAIC is preferable, they are perhaps among the most fruitful of areas for future research, if for no other reason than they are simpler to compute than Bayes factors, are available in standard software (DIC, at least), and require fewer assumptions or restrictions on model specification than Bayes factors. The use of small-variance priors to identify local sources of model misfit also seems particularly promising because of its straightforward application and interpretation. Because this method is expected to be sensitive to prior specification, it is necessary to investigate its power in different types of models, different levels of misspecification, different sample sizes, and different levels of prior information. Because PPP is also easily computed, and it is provided in the popular BSEM software M*plus* (using the χ^2 statistic as criterion), future research into its finite sampling behavior is also warranted.

Because small-variance priors are so easily implemented and information criteria are so easily computed, they are likely to be adopted quickly by applied researchers interested in using Bayesian methods. The goal of this dissertation is to provide information about the finite sampling behavior of these tools, so that practical guidance can be given for how to apply these methods to real data. Information criteria are the focus of Study 1, and small-variance priors are the focus of Study 2.

PART II: Assessing Bayesian Tools for Selecting an Optimal Measurement Model

In Study 1 I investigate DIC, which is readily available in popular Bayesian modeling

software (BUGS and JAGS) and in SEM software that provides Bayesian estimation options (Amos and M*plus*). I also investigate the more recently proposed WAIC, as well as *SE* estimates meant to characterize the sampling variability of WAIC.

Monte Carlo Design for Study 1

Table 1 summarizes the manipulated variables and their levels. I imposed invariance constraints across either two groups (in a single-factor, multiple-group model) or two occasions (in a two-factor, single-group model) on factor loadings and item intercepts for a single latent factor with four standard normal indicators. Fewer parameters were estimated in multiple-group models than in longitudinal models, which include four residual correlations between the same items measured over time. However, the longitudinal models must reproduce 16 additional observed covariances between items across time that the multiple-group models do not, so this design reveals the effect of model type on DIC and WAIC variability and model preference.

Figure 1 depicts the data-generating model with fixed and manipulated population characteristics. In the data-generating model, the latent factor in each group or occasion was $\sim N(\mu = 0, \sigma = 1)$. In the longitudinal model, the factor correlation between Times 1 and 2 was set to 0.5. In conditions without DIF, factor loadings for Items 1–4 range from moderate to high (similar to Hu & Bentler, 1998, 1999; Kim & Yoon, 2011; Stark, Chernyshenko, & Drasgow 2006): $\lambda^T = [0.65, 0.70, 0.75, 0.80]$. In every condition, error variances were set to $1 - \lambda^2$, so that the total variance of each indicator remained $\sigma^2 = 1$ (i.e., $\theta^T = [0.5775, 0.51, 0.4375, 0.36]$ in the no-DIF conditions). Item intercepts in the no-DIF conditions had a mean of zero: $\tau^T = [0.2, 0.3, 0.0, -0.5]$. Because the latent means were zero, the intercepts were also the indicator means.

Effect size (i.e., magnitude of DIF) was manipulated incrementally to investigate how model preferences are affected as the competing models become less appropriate. Five levels of

uniform DIF were manipulated simultaneously with five levels of nonuniform DIF. Differences in the Item-3 intercepts ($\Delta\tau_3$) varied from 0 to −0.8 in increments of 0.2. Because the total item variances are one, these are standardized mean-differences (i.e., Cohen's *d*), so the magnitude of DIF ranged from small ($d = 0.2$) to large ($d = 0.8$) according to Cohen's (1988) criterion. An effect size for nonuniform DIF is not as straightforward because it represents differences in regression slopes (in our case, correlations, because factor and indicator variances are one), so I consulted past simulation studies for guidance. Differences in the Item-4 factor loadings ($\Delta\lambda_4$) varied from 0 to −0.4 in increments of 0.1, which is the same range used by Meade et al. (2008), although they varied DIF in increments of 0.02. Similarly, Kim and Yoon (2011) defined small and large DIF as $\Delta\lambda = -0.2$ and −0.4, respectively, and Stark et al. (2006) defined small and large DIF as $\Delta\lambda = -0.15$ and $\Delta\lambda = -0.4$, respectively.

Table 1

Manipulated Variables in Monte Carlo Design for Studies 1 and 2

Study	Variable Name	Description	Levels
1 & 2	Type	Type of invariance under investigation	2 groups or 2 occasions
1 & 2	*N*	Total sample size ($n_g = N / 2$)	200, 300, 400, 600, or 800
1 & 2	DIF	Magnitude of DIF ($\Delta\lambda_4$ and $\Delta\tau_3$): focal group (or second occasion) is lower than the reference group (or first occasion)	$\Delta\lambda_4 = 0.0$, 0.1, 0.2, 0.3, or 0.4 $\Delta\tau_3 = 0.0$, 0.2, 0.4, 0.6, or 0.8
1	Parsimony	Whether the model has parsimony error. Correlated errors are added for a pair of variables (e.g., a testlet), which are excluded from the analysis model	$\rho_{21} = 0.0$ or 0.2
2	Prior	*SD* of normal prior for $\Delta\lambda$ and $\Delta\tau$, constrained near zero: $\sim N(\mu = 0, \sigma = ?)$	$\sigma = 0.05$ or 0.10

Note. Study 1 focuses on model selection. Study 2 focuses on DIF detection.

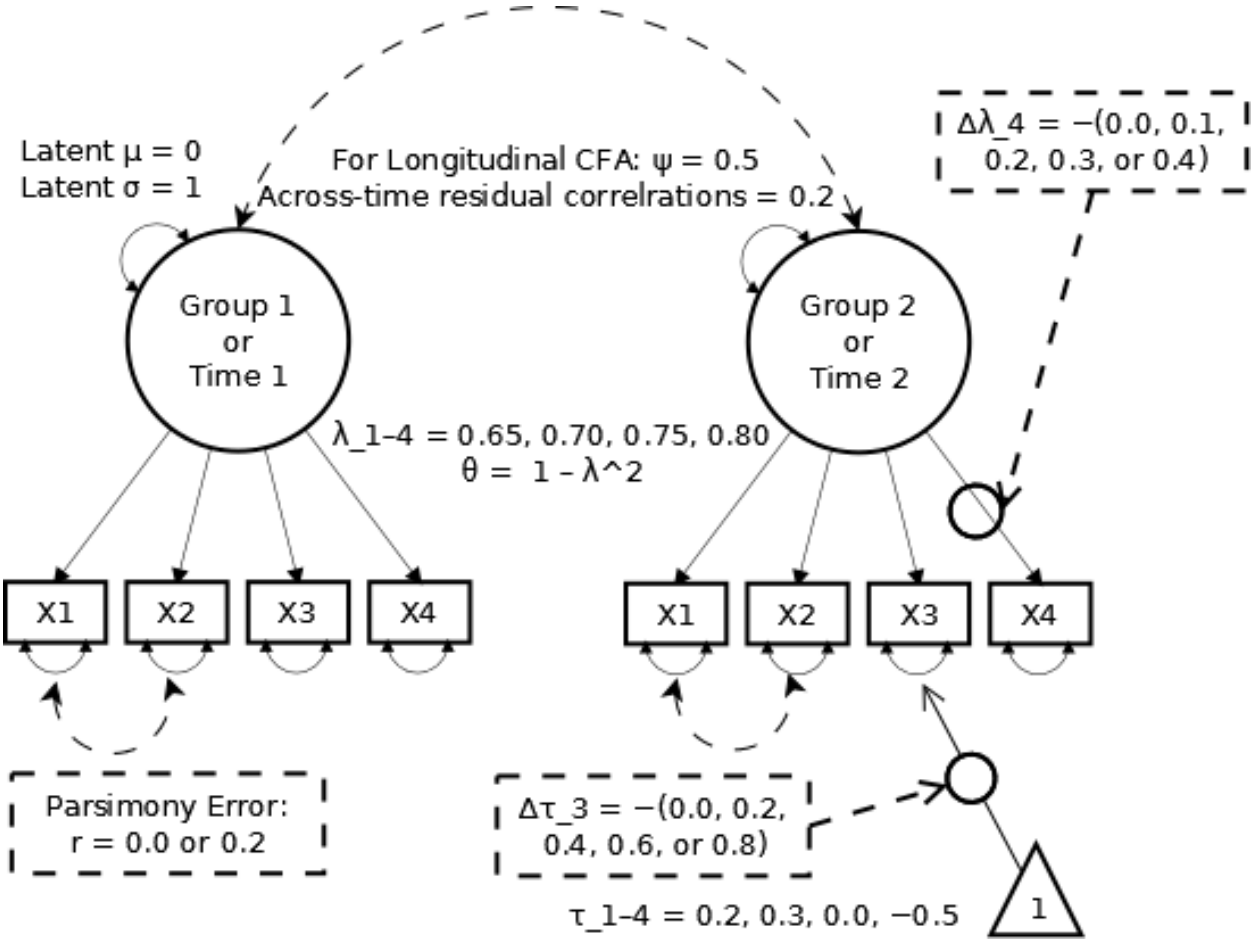

Figure 1. Population model(s) for data generation in Study 1. Solid lines represent population characteristics that are constant across all models, whereas dashed lines represent varying conditions described in the textboxes. The population model for Study 2 excludes parsimony error (i.e., there are no unmodeled residual correlations between Items 1 and 2 in the population) but is otherwise equivalent.

Because DIF is not the only possible source of model misspecification, an unmodeled residual correlation (ρ_{21} = 0 or 0.2) between the first two items was manipulated as a source of parsimony error (also referred to as model error or approximation discrepancy; MacCallum, 2003). A minor error correlation such as this might be quite common in practice, reflecting a testlet or negatively worded items. Omitting these parameters from the analysis model when it

exists in population would mean that when using ML, the analysis model would not be able to perfectly reproduce the population covariance matrix and mean vector; thus, the expected value for the χ^2 test statistic would be greater than the model's *df* (the expected value under the H_0 of perfect fit). The AIC asymptotically chooses among competing models the one that minimizes out-of-sample predictive errors (Vrieze, 2012). DIC and WAIC are Bayesian generalizations of AIC (Gelman et al., 2013; Vehtari & Ojanen, 2012), so it is of interest to see how they perform when the true population model is not among the competing models, as well as how variability of model preferences is affected when parsimony error is added to sampling error.

Similar sample size (*N*) conditions from past research on testing invariance in CFA (e.g., Meade & Bauer, 2007) were chosen to investigate how variability in model preferences changes as more information is provided by the observed data. The total sample size has five levels, in a similar range of small to large sample sizes seen in past research: N = 200, 300, 400, 600, and 800. For multiple-group models, these are divided into n_g = 100, 150, 200, 300, and 400 per group. A two-group situation with equal group sizes mimics common situations, such as when invariance is tested between sexes or experimental groups, and a two-occasion model would be used to test invariance in pre- and postintervention conditions. As *N* increases, sampling variability of parameter estimates decreases, as does the sampling variability of some fit indices—even ones whose means are not sensitive to *N*, such as CFI and RMSEA. But the variability of information criteria increases with *N* (the number of individual likelihoods), whether the information criteria are calculated using χ^2 (whose expected value increases proportionally with *N*) or using the log-likelihood directly. So it is difficult to anticipate how increasing *N* will affect variability in model selection using DIC or WAIC.

This will be a 2 (multiple-group or longitudinal model) × 5 (N = 200, 300, 400, 600, or

800) × 5 (magnitude of DIF; see Table 1) × 2 (presence or absence of parsimony error) factorial design. To reduce the sampling variability between conditions, I simulated data from the longitudinal model and analyzed it using the longitudinal model as well as the multiple-group model by treating the observations at separate occasions as separate groups, ignoring items' residual covariances across time. For each replication, I simulated data for the largest *N* and drew subsets from that sampling frame for other sample size conditions.

Procedure. I generated 500 samples of multivariate normal data from each population's model-implied mean vector and covariance matrix using the `mvrnorm` function in the R package *rockchalk* (Johnson, 2015). Bayesian models were fit to data with the Bayesian modeling software Stan (Stan Development Team, 2014), using the R package *rstan* (version 2.5). I monitored convergence by checking Gelman and Rubin's (1992) potential scale-reduction factor ($\hat{R}$) after each run. Starting with 500 burn-ins, if $\hat{R}$ for any model parameter exceeded 1.10, I ran the model again, doubling the iterations until either convergence was reached or the number of iterations exceeded 100,000. I saved the posterior *M*, *SD*, and 95% credible limits for each model parameter, which were estimated using 1000 iterations from each of three chains (regardless of how many iterations were needed for the burn-in phase), yielding 3000 draws from the target posterior distribution for stable 95% credible intervals. All estimated parameters had noninformative or weakly informative priors (see Appendix), making results similar to MLE. Models were fit with MLE using *lavaan* (Rosseel, 2012), to compare WAIC and DIC to AIC.

For each replication, I fit a sequence of three models commonly used to test measurement equivalence (e.g., Cheung & Rensvold, 2002). A configural (or "form") invariance model specifies the same number of factors and pattern of freely estimated parameters for each group or occasion, none of which are constrained to equality across groups or time. The model was

identified by fixing the factor variance to one and the factor mean to zero for both groups (or occasions). A metric (or "weak") invariance model establishes a common scale of measurement by constraining factor loadings to equality across groups or time. Because the latent scale of measurement is identified in the first group by fixing the factor variance to one, the factor variance for the second group (or occasion) is freely estimated. A scalar (or "strong") invariance model establishes a common scale and location by constraining factor loadings and item intercepts to equality across groups or time. The latent scale and location are identified in the first group by fixing the factor mean and variance to zero and one, so the factor mean and variance for the second group (or occasion) are freely estimated. The factor correlation and each item's residual correlation were also estimated in the longitudinal conditions.

When DIF is nonexistent, the scalar invariance model is the true model, but even when DIF is small (or when parsimony error is present), the scalar invariance model might be the optimal measurement model if it does not result in a large amount of misfit. When DIF is large, the configural and metric invariance models are overparameterized and the scalar invariance model is underparameterized. I also fit a fourth model in which all loadings and intercepts were constrained except for the fourth loading and the third and fourth intercept, which is what would be fit if the correct DIF parameters were identified (the likelihood of correct DIF detection is investigated in Study 2). In the presence of DIF and parsimony error, the fourth model is the true model, but it is overparameterized when DIF = 0. I included this model to see to what degree DIC and WAIC can distinguish between the fit of the optimal (but imperfect) measurement model and the fit of the true model when DIF is present, as well as the true scalar invariance model from the slightly overparameterized partial invariance model when DIF is absent.

Results and Discussion

Out of all 200,000 fitted models (4 models × 500 replications × 100 conditions), only 15 did not converge on a stable posterior distribution that yielded $\hat{R} < 1.1$ for all model parameters, although only seven of these had any $\hat{R} > 1.2$. Nonconvergence occurred almost exclusively in conditions of the smallest sample size ($N = 200$) and largest DIF ($\Delta\tau_3 = -0.8$ and $\Delta\lambda_4 = -0.4$) when fitting the most constrained model (scalar invariance). Although the posterior-mean estimates of the parameters were in an acceptable range, the between-chain variability was so great that it resulted in very large outliers of DIC_2. Therefore, these 15 observations were ignored when calculating measures of variability and relative efficiency of information criteria, which are strongly affected by these outliers. However, they were included when recording model preferences, which are made based on rankings rather than magnitude of information criteria. Only DIC_2 was noticeably affected by these nonconverged models, and as shown in sections below, the greater variability of DIC_2 and of its model preferences makes it the least preferable information criterion regardless.

Variability of information criteria. Because information criteria are calculated from the sum of individual log-likelihoods, their magnitude is linearly related to *N*. Analyses of variance (ANOVAs) for each information criterion indicated that more than $\eta^2 = 99\%$ of their variability is explained by *N*, so to assess the influence of other Monte Carlo factors, separate ANOVAs were run for each information criterion at each level of *N*, treating DIF (five levels), parsimony error (present or absent), invariance model (four levels), and model type (multiple-group or longitudinal) as independent variables. I used Type I *SS* to calculate η^2 because the removal of so few nonconvergent observations resulted in no practical difference between Types I and III *SS*. Cohen (1988) provided criteria for interpreting the size of η^2 (negligible < 1% < small < 6% < moderate < 14% < large). Nonnegligible effect sizes are reported in Table 2.

Table 2

Effect Sizes (η^2) of Monte Carlo Factors on Information Criteria

		Total Sample Size				
Information Criterion	Monte Carlo Factor	200	300	400	600	800
$WAIC_1$	DIF	42.9%	50.7%	54.2%	58.6%	61.5%
	Invariance Model	5.1%	6.1%	6.5%	7.2%	7.6%
	Parsimony	5.3%	6.2%	6.7%	7.2%	7.4%
	Type (multiple-group / -time)	2.2%	2.6%	2.7%	3.0%	3.3%
	DIF × Invariance Model	4.7%	5.5%	5.8%	6.3%	6.6%
$WAIC_2$	DIF	42.5%	50.2%	53.7%	58.0%	60.9%
	Invariance Model	5.0%	6.1%	6.5%	7.1%	7.4%
	Parsimony	5.2%	6.0%	6.5%	7.0%	7.2%
	Type (multiple-group / -time)	2.7%	3.2%	3.5%	3.8%	4.2%
	DIF × Invariance Model	4.6%	5.4%	5.7%	6.1%	6.4%
DIC_1	DIF	27.4%	31.6%	33.3%	35.5%	37.0%
	Invariance Model	3.7%	4.2%	4.4%	4.7%	4.9%
	Parsimony	3.2%	3.6%	3.9%	4.1%	4.2%
	Type (multiple-group / -time)	33.8%	36.6%	38.0%	39.4%	40.2%
	DIF × Invariance Model	3.1%	3.6%	3.7%	4.0%	4.1%
DIC_2	DIF	13.6%	21.6%	26.4%	29.8%	30.9%
	Invariance Model	2.2%	1.3%	1.1%	1.1%	1.3%
	Parsimony	4.7%	6.2%	7.0%	7.0%	7.2%
	Type (multiple-group / -time)	11.5%	25.4%	32.9%	39.7%	41.9%
	DIF × Invariance Model	1.2%	1.5%	1.6%	1.8%	1.9%
AIC	DIF	30.4%	36.2%	39.4%	42.8%	45.3%
	Invariance Model	3.3%	4.3%	4.8%	5.4%	5.9%
	Parsimony	3.5%	4.1%	4.5%	4.9%	5.1%
	Type (multiple-group / -time)	12.0%	15.1%	16.9%	19.1%	20.6%
	DIF × Invariance Model	3.7%	4.3%	4.7%	5.1%	5.4%

Note. Only effects with $\eta^2 > 1\%$ are shown.

All factors (DIF, parsimony error, invariance model, and model type) had substantial main effects on each information criterion, as did the interaction between DIF and invariance model. No other interactions had substantial effects on any information criteria. Although overall sampling variance increases with *N*, so does the proportion that is explained. For WAIC, the only large effect size was the main effect of DIF, which explained between 42–62% of variability in WAIC. All other effects on WAIC were small to medium, which is illustrated by the similar asymptotic behavior of $WAIC_1$ among panels in Figure 2. Means are shown only for the most asymptotic condition (N = 800); plots at other Ns look very similar, with uniformly lower means. Because it shows almost identical behavior, no Figure is provided for $WAIC_2$.

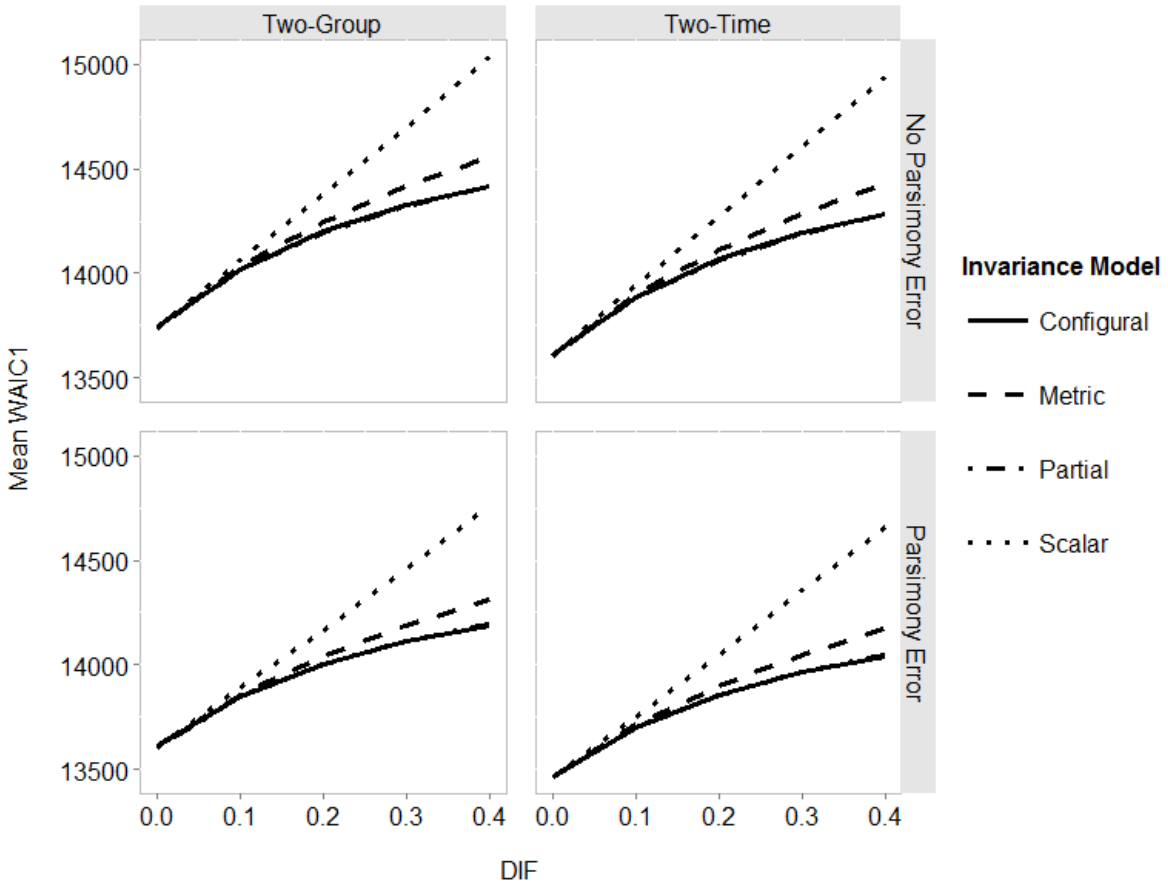

Figure 2. Mean $WAIC_1$ across conditions when N = 800. Lines for the Partial Invariance model (true when DIF > 0) are barely visible, obscured by lines for the Configural Invariance model.

The substantial interaction is illustrated by the steeper slopes for the scalar invariance model as DIF increases. The small-to-medium effect of parsimony error is reflected by the small change in intercepts between the top and bottom panels, and the even smaller change in intercepts between left and right panels shows how small the effect of model type is. This is ideal behavior for an information criterion because only model misspecification (in the form of DIF, parsimony error, and model constraints) seems to have noticeable effects on their expected behavior. It is surprising that average values of WAIC showed so little discrepancy between models when there is little or no DIF, but this is common to DIC and AIC as well, and model preferences reported in the next section show clear preferences for the most parsimonious model.

The mean behavior of AIC seems less ideal because although it is mostly affected by DIF (η^2 = 30–45%), it is also largely affected model type (η^2 = 12–21%). This might imply that even holding other characteristics of the data (*N*) and model (level of misspecification) constant, model preferences might be affected merely by whether the invalid constraints are made in a multiple-group or longitudinal context. This unideal behavior is even more apparent in DIC_1, which is even more affected by model type (η^2 = 34–40%) than by DIF (η^2 = 27–37%). Whereas Figure 3 shows that AIC tends to be somewhat higher for multiple-group models, Figure 4 shows that DIC_1 tends to be noticeably lower for multiple-group models. Other than this effect of model type, the other effects (parsimony error, invariance model and its interaction with DIF) remain qualitatively similar to the mean behavior of WAIC.

DIC_2, on the other hand, has behaved more erratically. Like DIC_1, it is more influenced by model type (η^2 = 11–42%) than by DIF (η^2 = 13–31%), except when *N* = 200. But Figure 5 shows that there are smaller differences in mean DIC_2 across invariance models, implying less discrepancy among models, and the intersecting lines imply less consistent preferences.

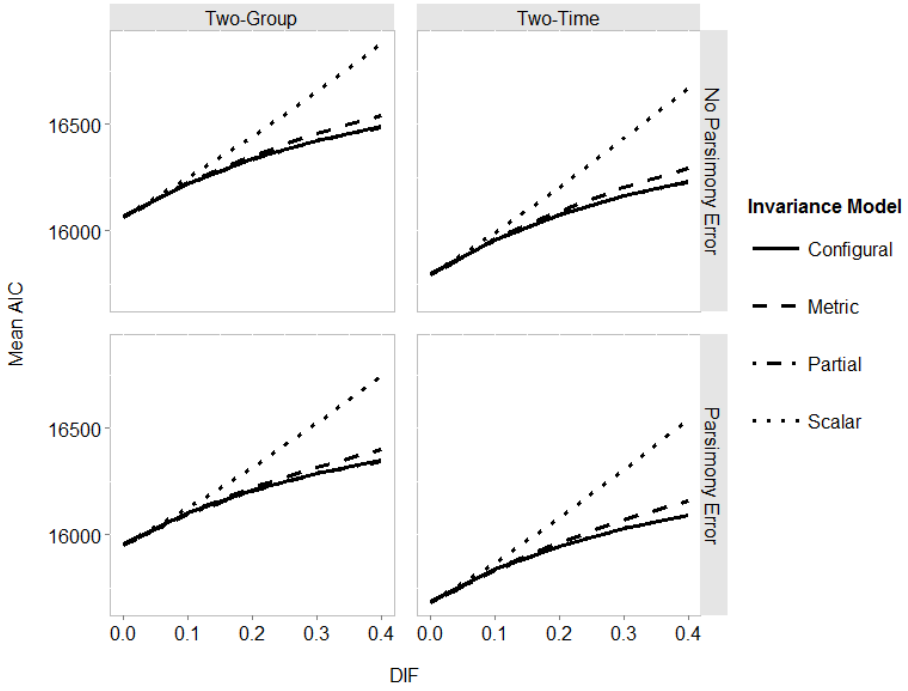

Figure 3. Mean AIC across conditions, when N = 800.

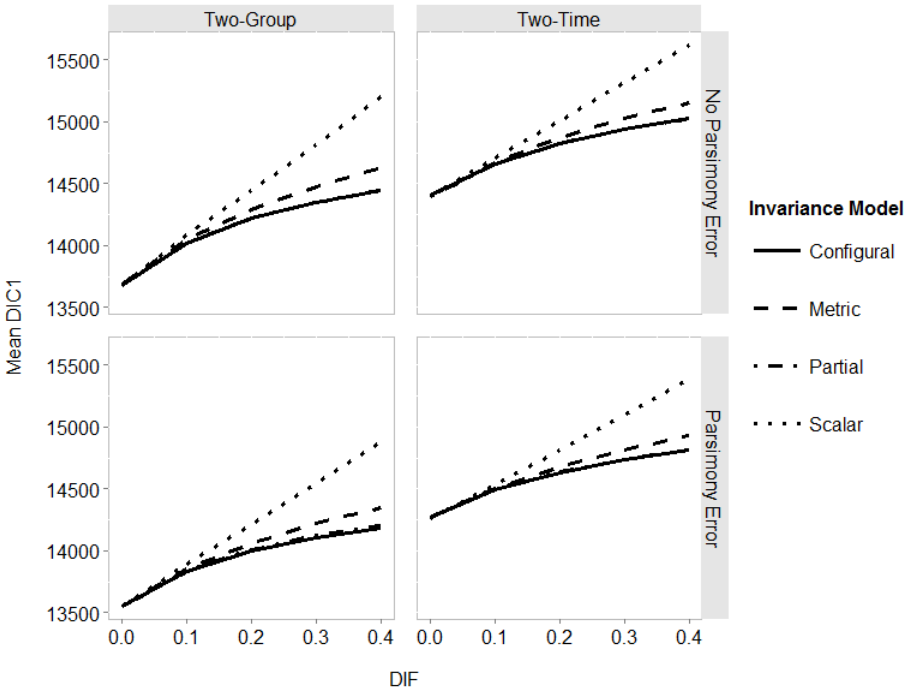

Figure 4. Mean DIC_1 across conditions, when N = 800.

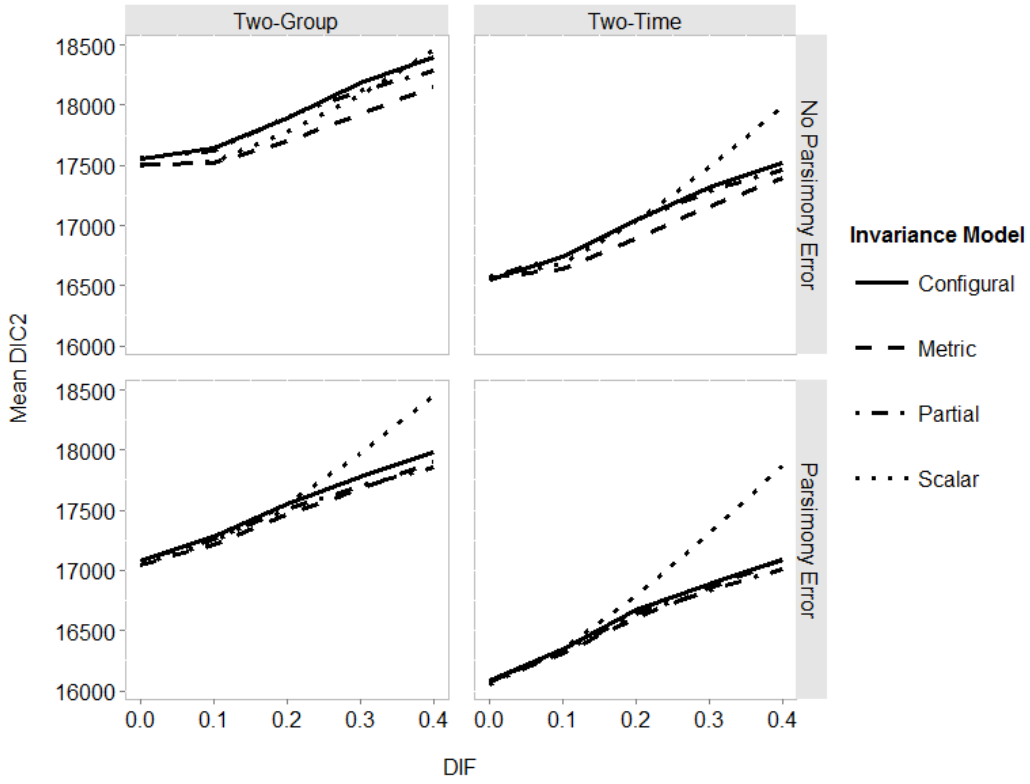

Figure 5. Mean DIC_2 across conditions, when N = 800.

Whereas mean behavior of information criteria were affected by all Monte Carlo factors, parsimony error and invariance models had no noticeable effect on the variance of most information criteria. Figure 6 shows that when N = 800, model type also had negligible influence on the *SD* of WAIC and AIC, but the *SD* of DIC_1 was consistently much larger for multiple-group than longitudinal models. As with the means in Figures 2–5, plots at smaller *N*s look very similar. Overall, AIC had the lowest *SD*, which is not surprising because the punishment term is a constant (twice the number of parameters), whereas for DIC and WAIC the fit and punishment terms are both random variables. Figure 6 excludes DIC_2 in order to more clearly see differences between the other four information criteria. Figure 7 includes DIC_2, which consistently has much greater variance than other information criteria. Furthermore, DIC_2 is more variable for the least restrictive configural invariance model, especially for larger DIF.

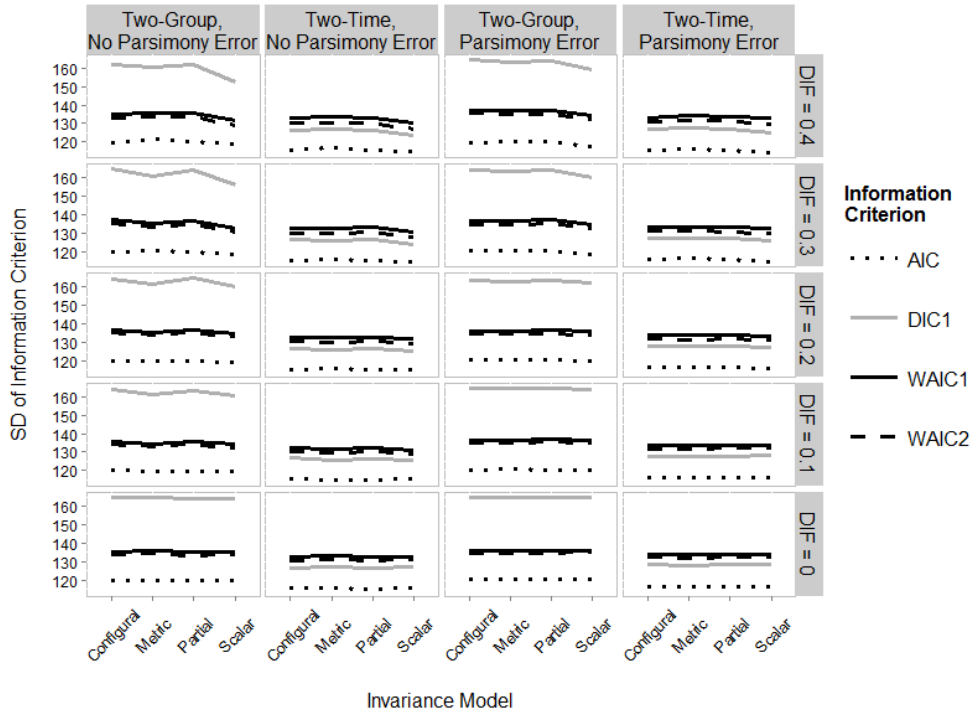

Figure 6. Standard deviations of four information criteria across conditions, when N = 800.

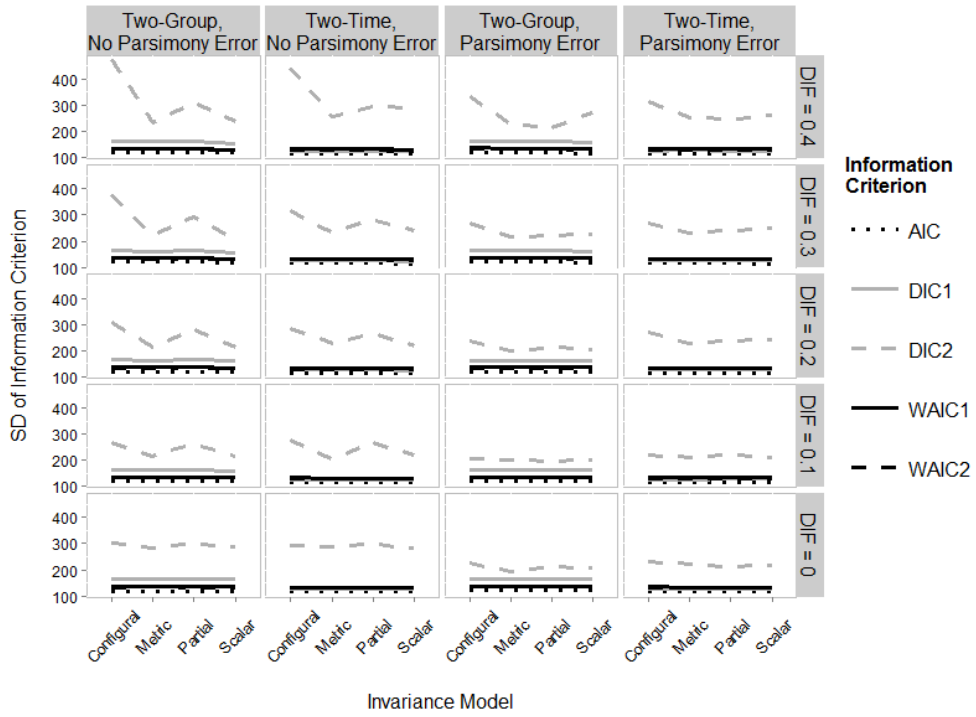

Figure 7. Standard deviations of all five information criteria across conditions, when N = 800.

The high variability of DIC_2 makes it less preferable than DIC_1, which is further illustrated by relative efficiency of DIC_1 to DIC_2 (i.e., the ratio of DIC_2 variance to DIC_1 variance). Figure 8 shows that at smaller sample sizes and larger DIF, the sampling variance of DIC_2 can be as much as 100 times the sampling variance of DIC_1, but even at larger sample sizes and smaller DIF, the variance of DIC_2 is often at least twice that of DIC_1.

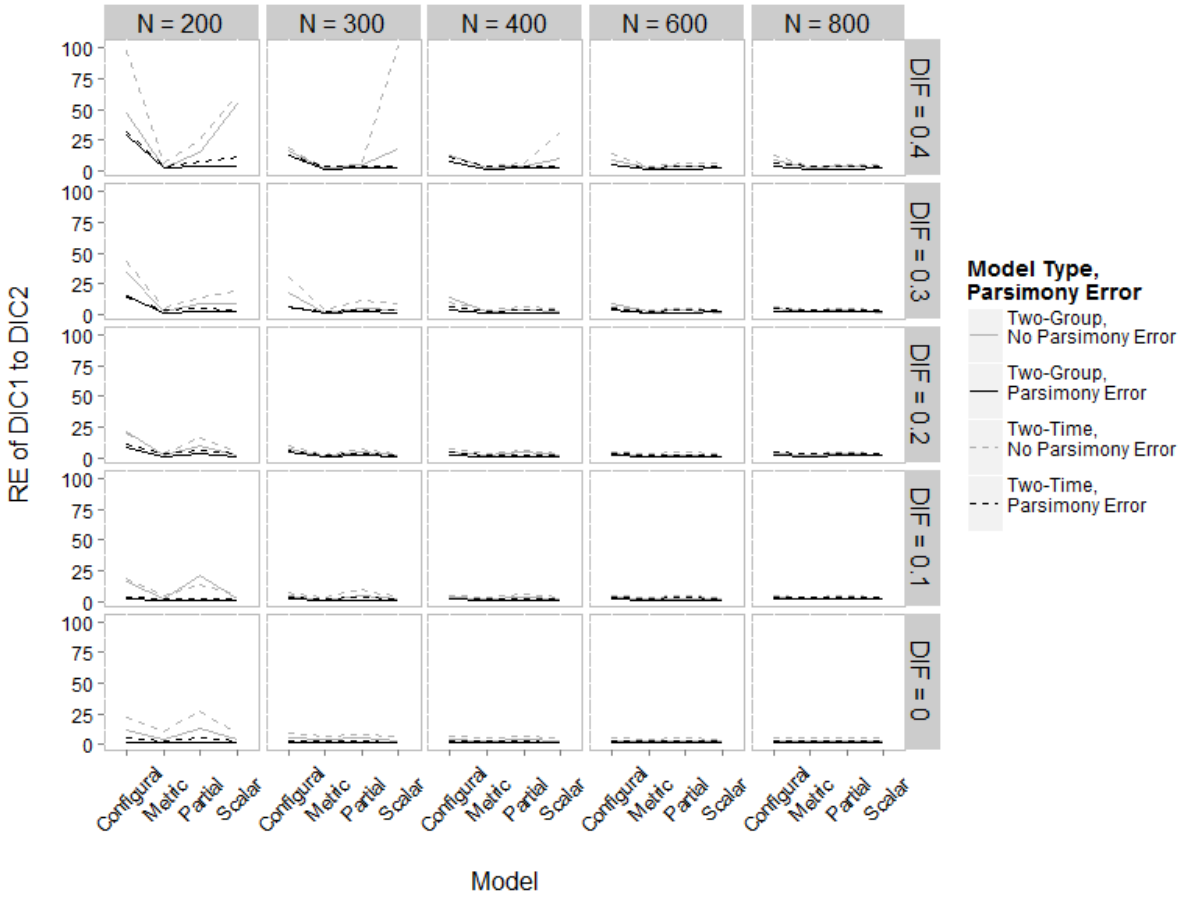

Figure 8. Relative efficiency of DIC_1 to DIC_2.

The greater variability of DIC_2 is expected and consistent with past research (Gelman et al., 2013). By contrast, the posterior-variance computation of pD for WAIC results in less sampling variability because the posterior variance is calculated separately for each observation,

then summed across observations, which creates stability (Gelman et al., 2013). Figure 9 shows that this gain in precision is small, but consistent. $WAIC_1$ has consistently greater sampling variance than $WAIC_2$, but it is always between 2–5%. The gain in precision for $WAIC_2$ appears greater for longitudinal models than for multiple-group models, but this slight difference may not generalize to other types of models. Model misspecification (large DIF, parsimony error) leads to slightly greater discrepancies in precision, particularly for the scalar invariance model, which has the most invalid constraints.

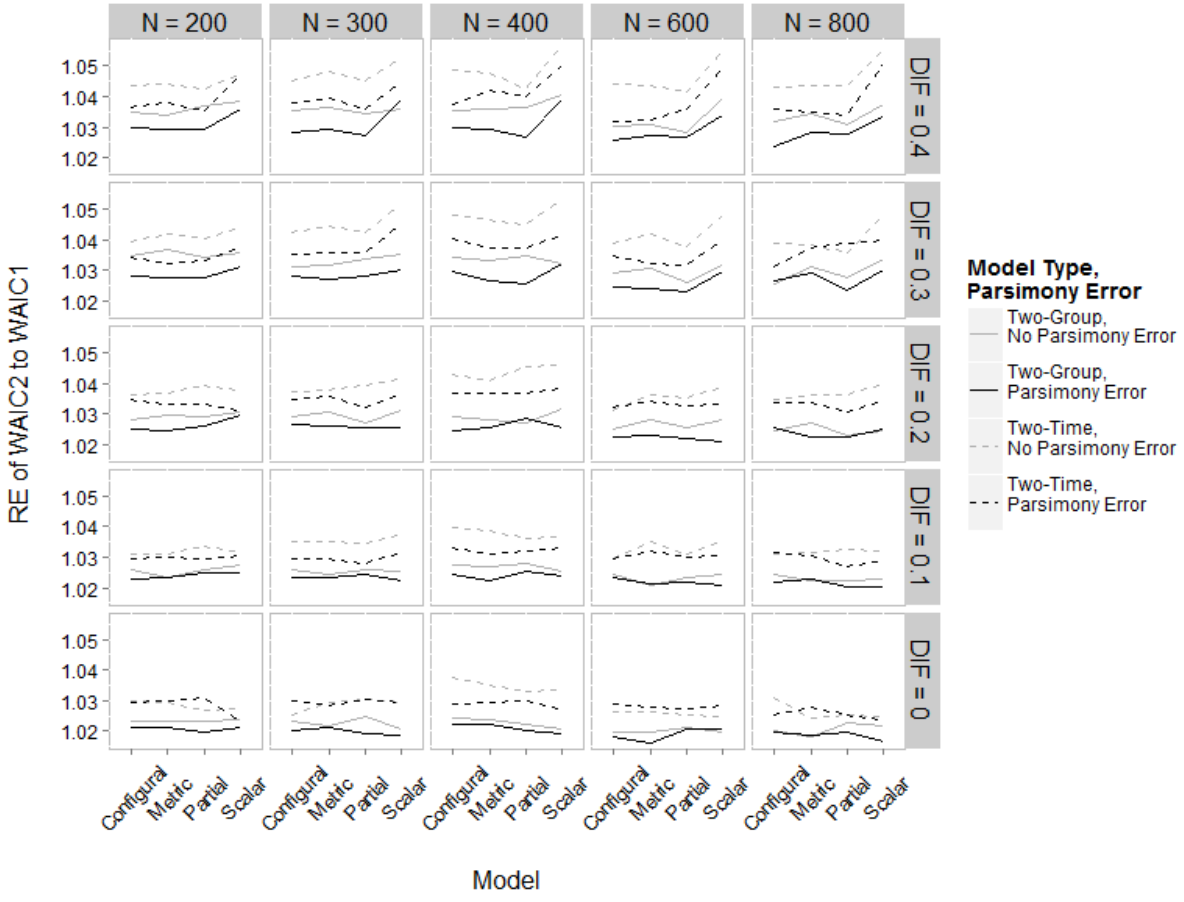

Figure 9. Relative efficiency of $WAIC_2$ to $WAIC_1$.

As expected, the more efficient computation of WAIC is $WAIC_2$, and the more efficient computation of DIC is DIC_1. Figure 10 compares the efficiency of these two information criteria. For multiple-group models, DIC_1 has more than 40% greater sampling variability than $WAIC_2$, making $WAIC_2$ the preferred information criterion. For longitudinal models, DIC_1 is slightly more efficient, but $WAIC_2$ is still 92.9-95.4% as efficient as DIC_1. Equivalently, the reciprocal relative efficiency indicates that for longitudinal models, $WAIC_2$ has only 4.8–7.6% greater sampling variance than DIC_1, so their model preferences may have nearly equal consistency across samples. This is investigated in the section on model rankings.

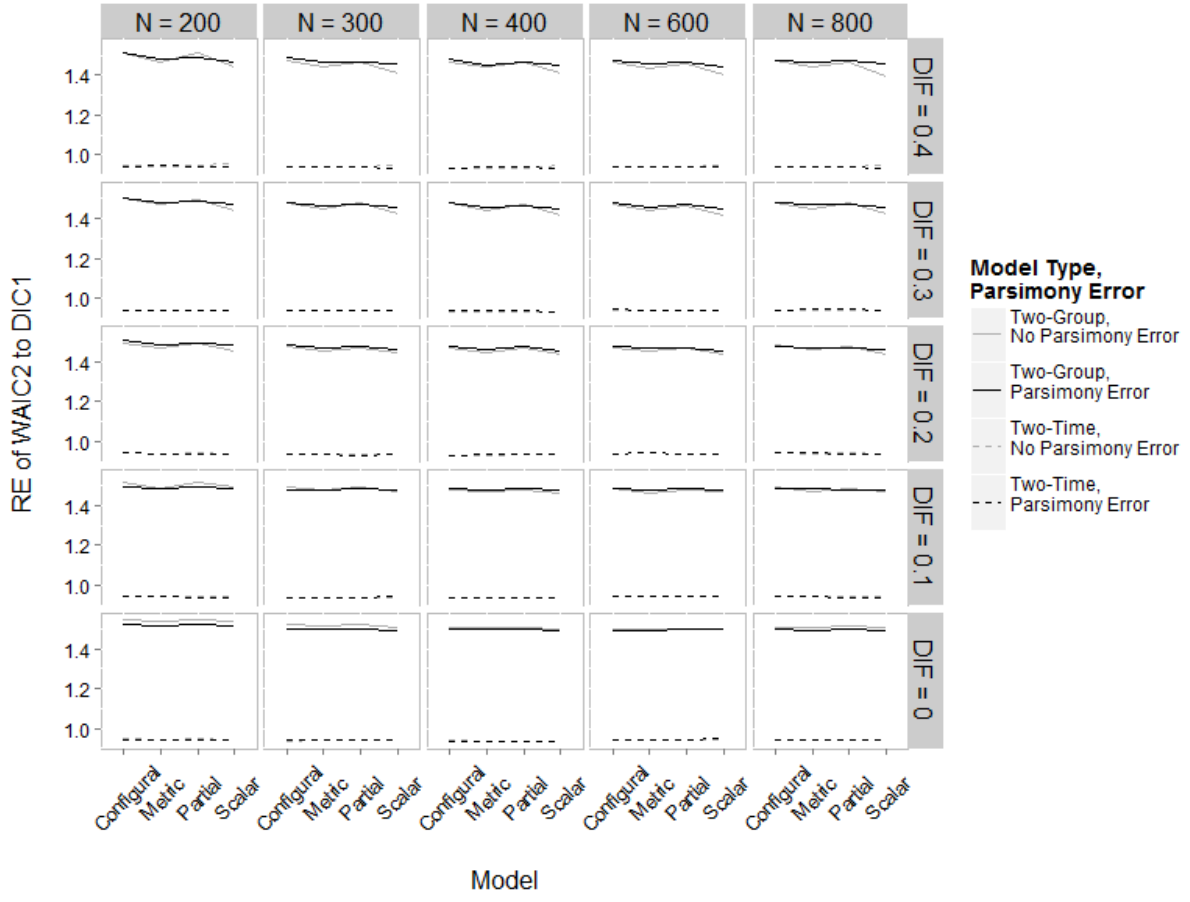

Figure 10. Relative efficiency of $WAIC_2$ to DIC_1.

To evaluate the newly proposed *SE* estimates for WAIC (Vehtari & Gelman, 2014), I calculated relative *SE* bias within each condition to illustrate how the observed variability of WAIC compares to its average estimated *SE*:

$$\text{Relative } SE \text{ bias} = \frac{average\ SE - observed\ SD}{observed\ SD}. \quad (28)$$

Figure 11 reveals $WAIC_2$ *SE*s to be 20–28% smaller than observed sampling *SD*s. Similar patterns were observed for $WAIC_1$ (not depicted), but bias for $WAIC_1$ was consistently about 2% more negative than bias for $WAIC_2$. Longitudinal models appear to have less bias than multiple-group models, but this difference is slight and may not generalize to other types of models. Bias appears to decrease as DIF increases, and when DIF is large, bias is somewhat less extreme for the most constrained model (scalar invariance).

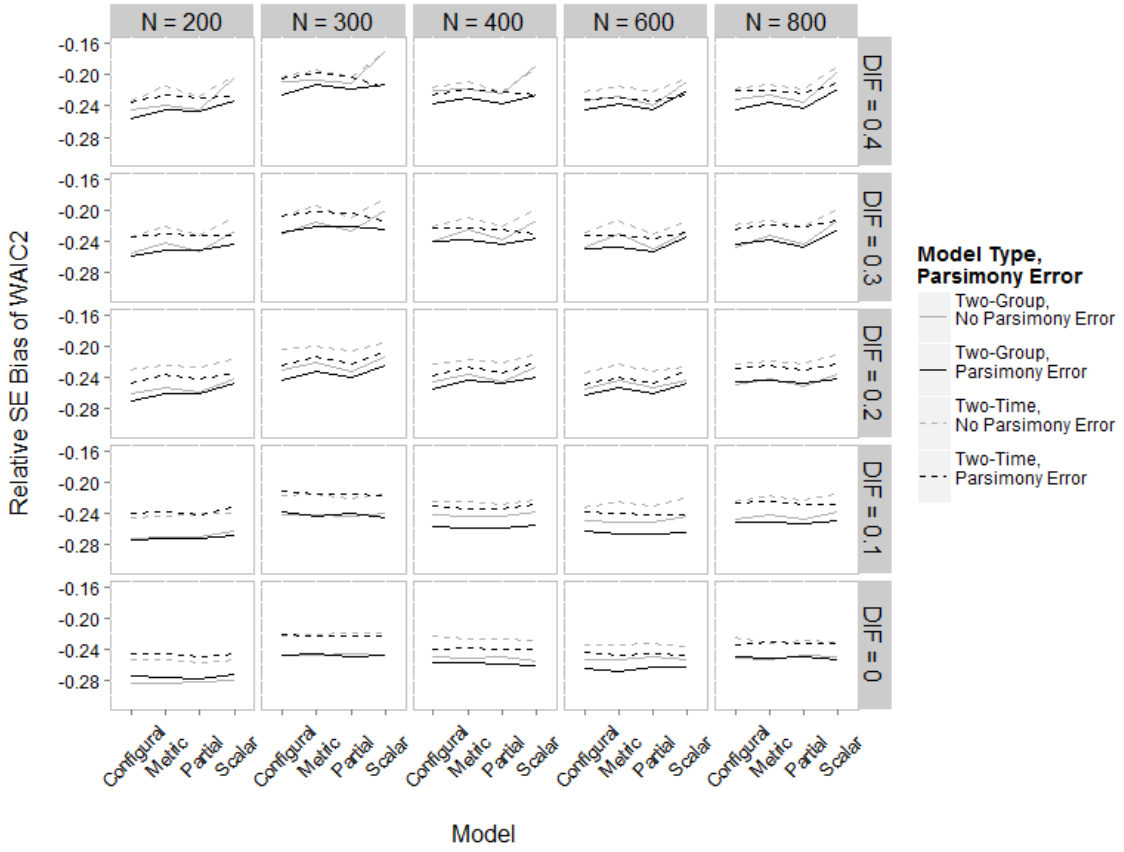

Figure 11. Relative *SE* bias for $WAIC_2$.

Impact of model misspecification. Before reporting model rankings and preferences using each information criterion, it is important to understand the impact of making invalid parameter constraints on model fit and on estimates of the latent mean and variance in the second group or occasion[6]. To reveal the degree to which model misspecification impacted expected values (free of sampling error) of the latent mean and variance estimates in the second group or occasion, I fit the scalar invariance model to each of 10 population covariance matrices and mean vectors (five levels of DIF and 2 levels of parsimony error). I specified the largest sample size condition (N = 800 in the longitudinal model, or 400 per group in the multiple-group model) for the purposes of calculating descriptive fit indices (CFI and RMSEA).

Figure 12 shows that as DIF increases, bias of the second factor mean becomes more negative. DIF on the *x* axis can be interpreted as Cohen's *d*, and because the first factor variance = 1, values on the *y* axis can be interpreted as Glass' Δ—a variation on Cohen's *d* calculated using the variance of a reference group rather than a pooled variance. Bias is negligible when DIF is medium or less (i.e., $| \Delta\tau_3 | < 0.5$), but bias becomes substantial (but still small) as DIF becomes large (i.e., $| \Delta\tau_3 | = 0.8$). Bias is slightly more extreme in multiple-group than longitudinal models, and for both models the bias is less extreme when there is parsimony error.

Figure 13 shows that as DIF increases, bias of the second factor variance becomes more negative. Bias on the *y* axis can be interpreted as change proportional to the true variance of one. When factor loadings have no more than medium DIF (i.e., $| \Delta\lambda_4 | < 0.2$), the second factor variance is only 10–15% lower than the true variance. When DIF is large, negative bias is almost 25% when there is no parsimony error, but less than 20% when there is parsimony error. There is no noticeable difference in bias between longitudinal and multiple-group models.

[6] Latent mean and variance parameters remain fixed to zero and one for the first group or occasion.

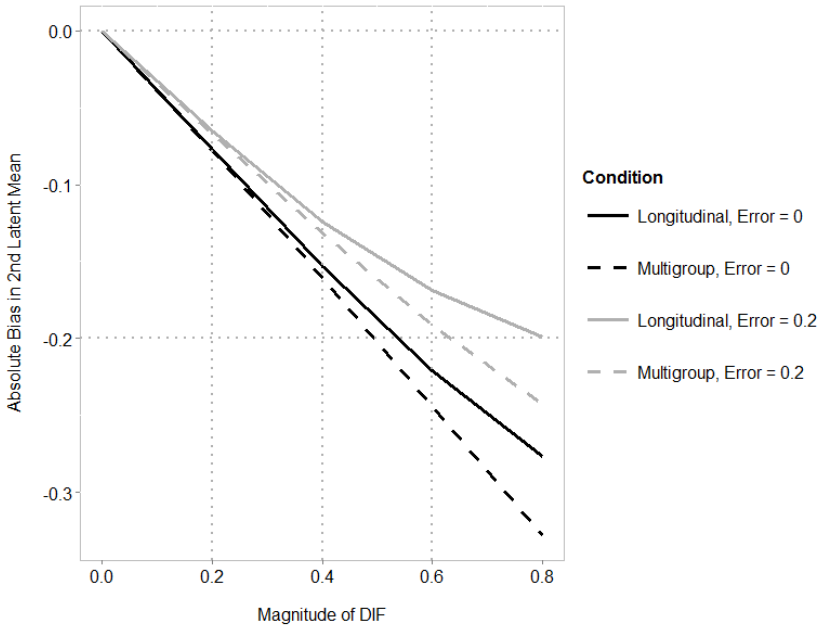

Figure 12. Effect of DIF, model type, and parsimony error on latent-mean bias. As DIF values vary from $\Delta\tau = 0$ to -0.8 by 0.2 on the *x* axis, DIF values for $\Delta\lambda$ simultaneously vary from 0 to -0.4 by 0.1.

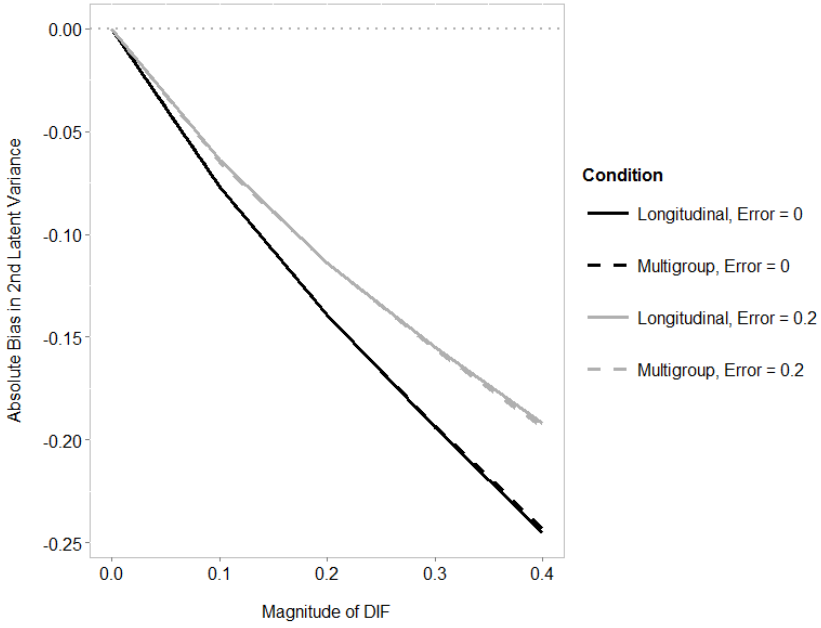

Figure 13. Effect of DIF, model type, and parsimony error on latent-variance bias. As DIF varies from $\Delta\lambda = 0$ to -0.4 on the *x* axis, DIF values for $\Delta\tau$ simultaneously vary from 0 to -0.8 by 0.2.

Figures of average bias calculated using parameter estimates from each sample size condition show the same patterns as Figures 12 and 13. Figures 14 and 15 indicate how model fit is affected by plotting large-sample approximations of expected values of popular practical fit indices. CFI and RMSEA both indicate unacceptably poor fit when DIF is large (i.e., $| \Delta\lambda_4 | > 0.25$ and $| \Delta\tau_3 | > 0.5$), regardless of model type or parsimony error (using Bayesian estimation, we would expect PPP to be close to zero). When DIF is absent but there is parsimony error (far-left of grey lines in Figures 14 and 15), the model still fit very well, verifying that the unmodeled residual correlations do not introduce enough misfit to justify rejecting the model altogether. Fit indices without parsimony error eventually converge to the same value as DIF increases. For CFI there is little difference between model types, but RMSEA is more sensitive to DIF-related misfit in multiple-group models than in longitudinal models.

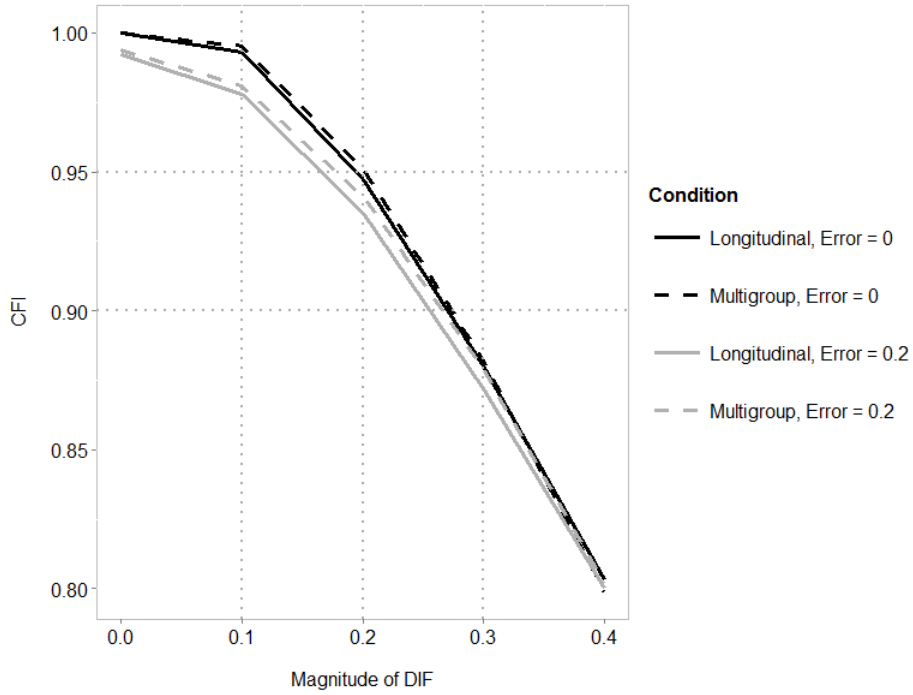

Figure 14. Effect of DIF, model type, and parsimony error on CFI. As DIF values vary from $\Delta\lambda = 0$ to -0.4 on the *x* axis, DIF values for $\Delta\tau$ simultaneously vary from 0 to -0.8 by 0.2.

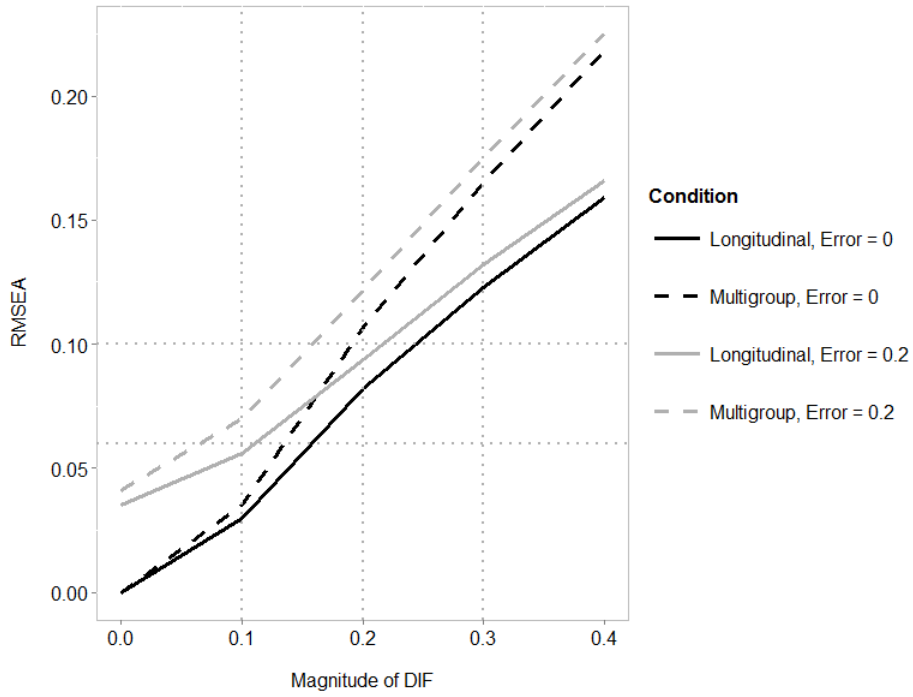

Figure 15. Effect of DIF, model type, and parsimony error on RMSEA. As DIF values vary from $\Delta\lambda = 0$ to -0.4 on the *x* axis, DIF values for $\Delta\tau$ simultaneously vary from 0 to -0.8 by 0.2.

Model rankings and preferences. Among the three models fit in sequence to each replication (configural, metric, and scalar invariance), the lowest information criterion indicates which model should be preferred as providing an optimal balance between parsimony and predictive accuracy. The scalar model is the correct model when DIF = 0, so ideally it would be the most commonly preferred model in these conditions. Fit is still good and latent parameter estimates are only minimally biased when DIF is minimal, so choosing the more constrained metric or scalar models might not lead to substantive interpretations whose invalidity is of practical consequence. Fit is only adequate when DIF is moderate, so the configural model should be expected have the lowest information criteria, indicating the invariance constraints are not tenable and thus steps should be taken to identify items with DIF (investigated in Study 2).

Model preferences are depicted in Figure 16, in which each panel compares AIC, DIC_1,

$WAIC_1$, and $WAIC_2$; only *N* and DIF vary across panels, as model preferences of these four information criteria did not vary substantially across model type or parsimony error. The two black lines depict $WAIC_1$ (solid) and $WAIC_2$ (dashed), which are very similar in all panels. Surprisingly, although the $WAIC_2$ calculation seemed preferable due to slightly less sampling variability than $WAIC_1$, the scalar model is chosen slightly less often by $WAIC_2$ when DIF is absent; however, this difference is slight and of no practical consequence. The dashed grey line depicts DIC_1, whose model preferences are very similar to WAIC in all conditions, so their differences in sampling variability appear to have no practical consequence. AIC's model preferences (solid grey line) are provided for comparison to using MLE, although other fit indices are typically used to test invariance in a frequentist framework (e.g., ΔCFI).

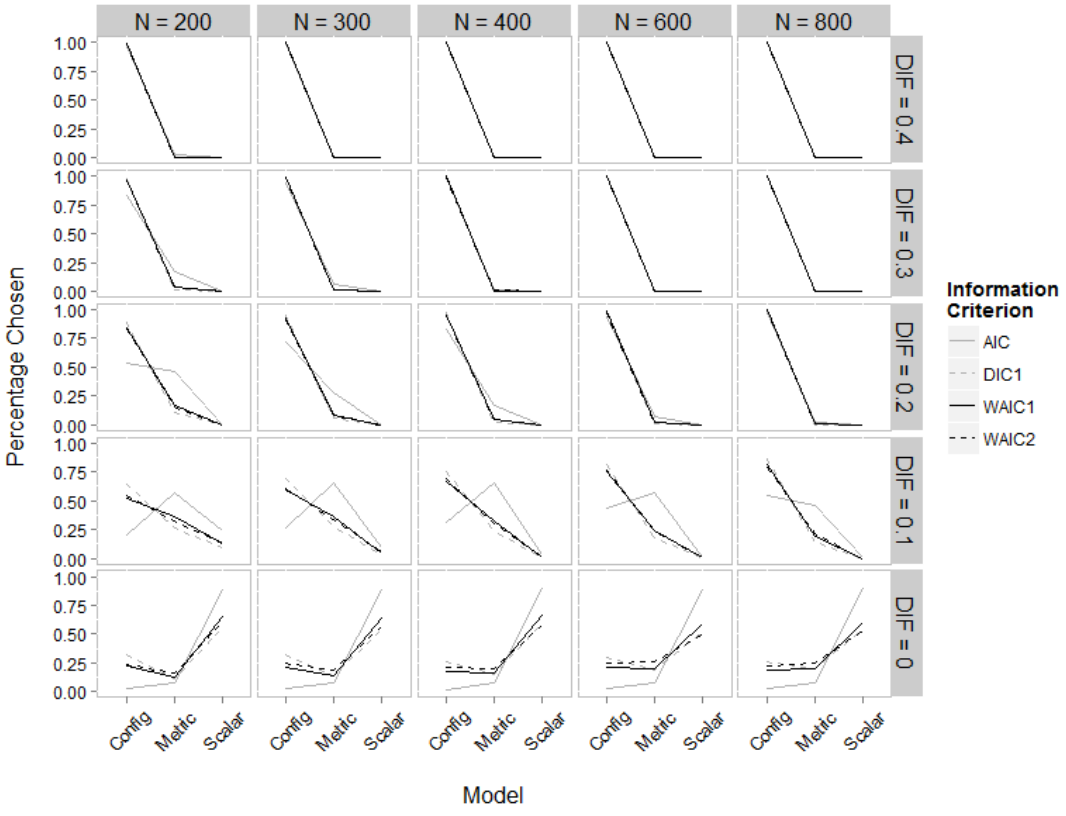

Figure 16. Model preferences based on ranked AIC, DIC_1, $WAIC_1$, and $WAIC_2$. Results are collapsed across model types (multiple-group vs. longitudinal) and presence of parsimony error.

When DIF is absent (bottom row of panels), the scalar invariance model is the most commonly preferred model, although it is only chosen in 50–65% of samples, and the overparameterized configural and metric invariance models would each be selected in up to 25% of samples. $WAIC_1$ has the highest rates of choosing the correct (scalar) model in a Bayesian context when DIF is absent, but those rates are not as high as when using AIC in MLE (89% regardless of *N*).

When DIF is minimal ($\Delta\lambda_4 = -0.1$, $\Delta\tau_3 = -0.2$), DIC_1 and WAIC already prefer the scalar model the least often, although AIC prefers it slightly more often when $N = 200$. In these conditions, WAIC and DIC_1 prefer the metric model more often than the scalar model, and the configural model is preferred most often, especially as *N* grows. AIC also prefers the configural model least, but AIC prefers the metric model most often, although the discrepancy between metric and configural decreases as *N* increases, until configural is the most preferred model when $N = 800$. Minimal DIF should not lead to grossly invalid substantive conclusions, so these variable model preferences should not be problematic.

When DIF is moderate ($\Delta\lambda_4 = -0.2$, $\Delta\tau_3 = -0.4$), DIC_1 and WAIC consistently prefer the configural model, which is correct. The scalar model was never preferred by any criteria, and the metric model was chosen by DIC_1 and WAIC in less than 10% of samples when $N > 200$. When $N = 200$, DIC_1 chose the metric model in 10.5% of samples, whereas both WAICs chose the metric model in 16% of samples. In contrast, AIC chose the metric model about as often as the configural model when $N = 200$, but chose the configural model more frequently as *N* increased. When DIF is large ($\Delta\lambda_4 \geq -0.2$, $\Delta\tau_3 \geq -0.4$), WAIC and DIC_1 almost exclusively prefer the configural model (96–100% of samples; only < 99% when $N \leq 300$), but AIC still chooses the metric model in up to 17% of samples except in the largest *N* or DIF conditions.

DIC_2 is depicted separately in Figure 17 because its behavior varies more across model type and parsimony error. When DIF is absent or minimal, all models appear nearly equally preferable at larger *N*, although the metric and scalar model appears only slightly more preferable at smaller *N*. When DIF is substantial (top three rows of panels), the metric model is typically the most frequently chosen. The configural model is the second most frequently chosen, at varying rates across *N*, DIF, model type, and parsimony error. At the largest DIF and $N \geq 400$, metric and configural models are both chosen in about half of longitudinal samples with parsimony error. In all other conditions of substantial DIF, the overly constrained metric invariance model would most frequently be chosen, potentially leading to invalid inferences about differences in latent-variable variance.

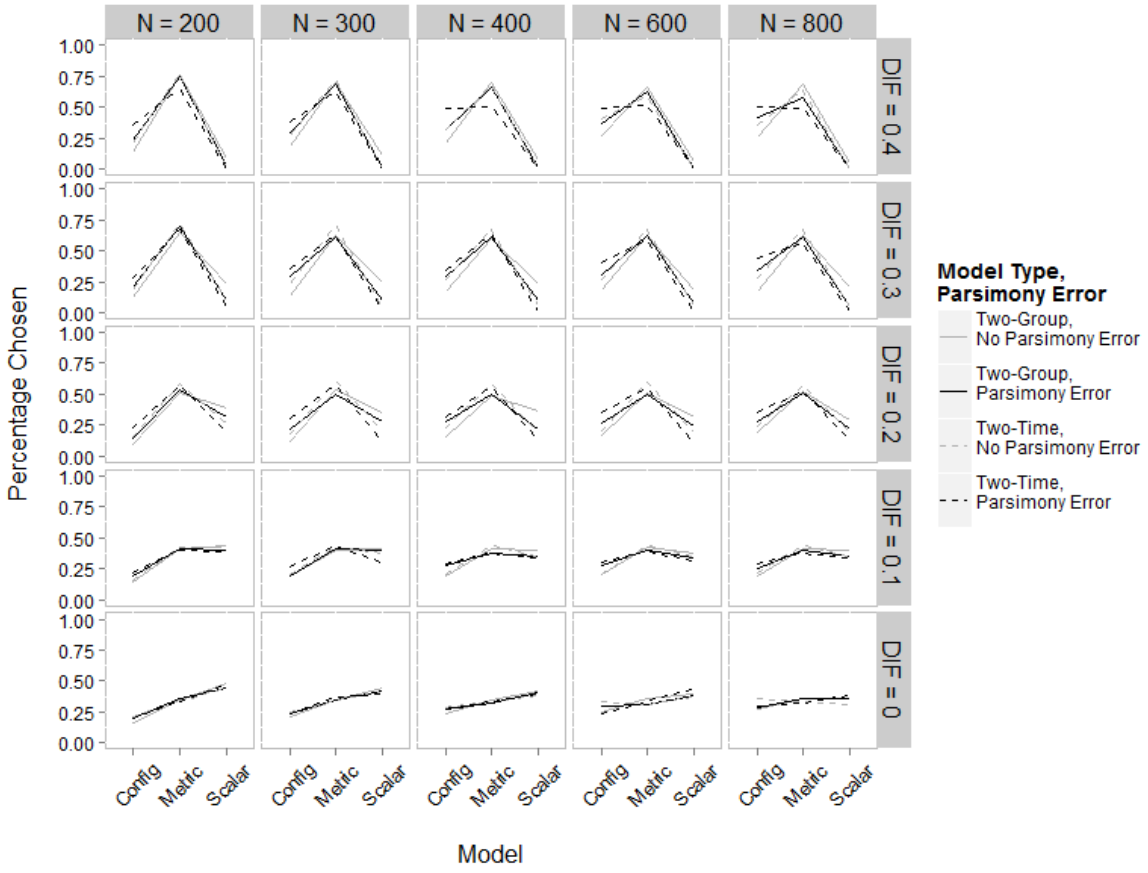

Figure 17. Model preferences based on ranked DIC_2.

Based on the model selection rates, WAIC and DIC_1 apparently lead to practically equivalent decisions, in spite of their differences in variability. Because their variability differs so little across models within a condition (i.e., holding the model, sample, and population characteristics constant), relative efficiency has little consequence on model selection behavior. When DIF is substantial, WAIC and DIC_1 more consistently choose the configural model than AIC would when using MLE, so researchers using Bayesian CFA would more often correctly choose to search for DIF among the items to establish partial invariance. Unfortunately, WAIC and DIC_1 also choose either the configural or metric model about half the time when DIF is completely absent, indicating a high rate of what would be called Type I errors in a ML context. Using MLE for comparison, AIC seems to choose the correct scalar model at a much higher rate (almost 90% across conditions), but when substantial DIF is present, it takes greater *N* and DIF for AIC to more consistently choose the appropriate configural model. Under no condition did DIC_2 show preferences for the most preferable model, so I do not recommend its use for selecting an optimal measurement model in Bayesian CFA.

Researchers may also be interested in using SE_{WAIC} to calculate a 95% CI for WAIC, which would indicate whether the most highly preferred model is substantially or "significantly" more preferable than the second most highly preferred model. Figures 18 presents the rate at which the 95% CI of the most highly preferred model's WAIC excludes the second most highly preferred model's WAIC. The rates in Figure 18 therefore represent "rejection" rates if applied researchers were to use SE_{WAIC} in a similar way to using the *SE* of a mean-difference to test whether it is significantly different from a particular value (e.g., zero) for the H_0. Similarly, Figures 19 and 20 present rates at which the 95% CIs of the first and second most highly preferred models' WAICs, respectively, contain the third most highly preferred model's WAIC.

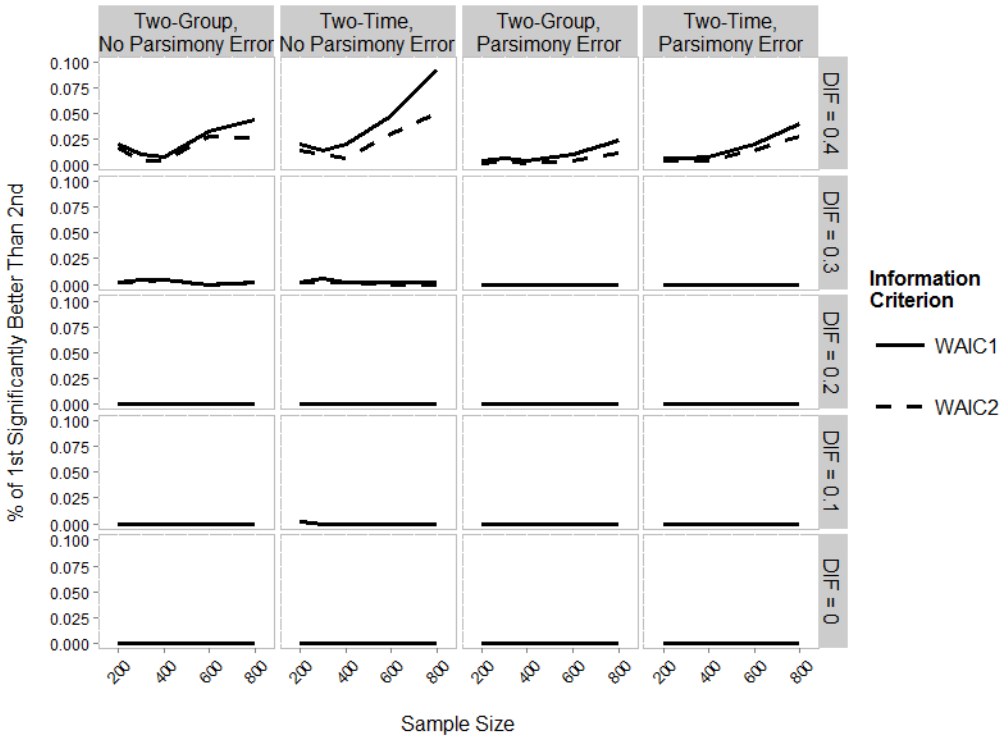

Figure 18. How often the lowest WAIC's 95% CI contains the next lowest WAIC. Note that the *y* axis is zoomed in on 0–10%.

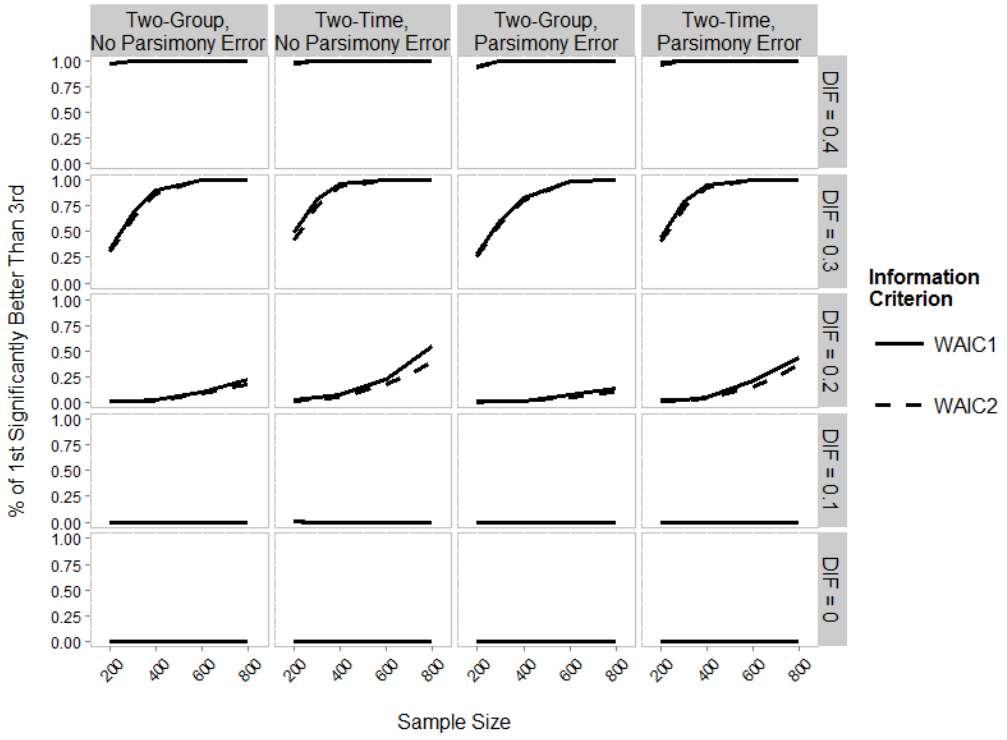

Figure 19. How often the lowest WAIC's 95% CI contains the highest WAIC.

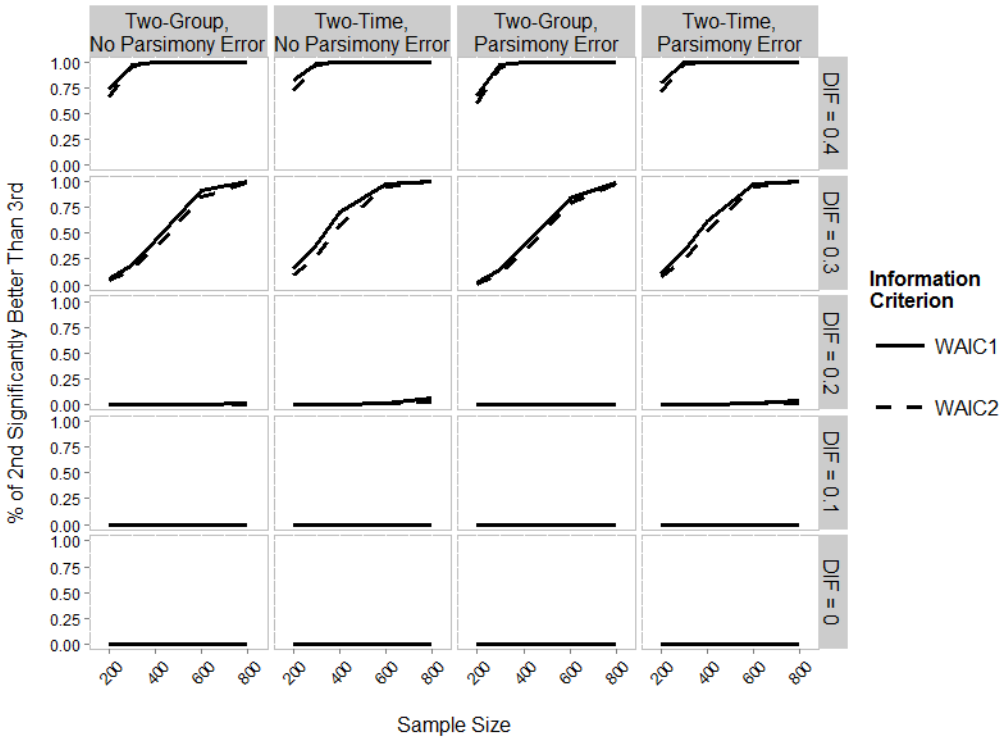

Figure 20. How often the second lowest WAIC's 95% CI contains the highest WAIC.

Given the negative bias of SE_{WAIC}, 95% CIs should favor higher "power" to detect significant parsimony-adjusted differences in fit between models. Regardless of whether that were true, this method rarely indicates substantial differences between the top two ranked models. Only when DIF is most extreme would the top models ever appear distinguishable in practice, and even in these conditions the models would be indistinguishable in less than 10% of samples. Note that the *y* axis in Figure 18 is zoomed in on the 0–10% range, whereas Figures 19 and 20 have *y* axes that span the entire 0–100% range.

The most highly preferred model is almost always distinguishable from the third ranked when DIF is high (top two rows of Figure 19), especially when *N* is large. But the first ranked model is seldom distinguishable from the third ranked model when DIF is moderate and never distinguishable when DIF is negligible or absent. The second and third ranked models depicted

in Figure 20 are mostly distinguishable under the same conditions as seen in Figure 19, although not as often as the first and third ranked are. The WAIC rankings in Figure 16 show much clearer preference for the configural model over both the metric and scalar models, even when DIF is moderate, so using SE_{WAIC} to judge model equivalence appears more conservative than desirable, at least under the conditions in this simulation. Figures 2 and 16 suggest that on average, the top ranked model under large DIF is the configural model, followed by the metric and configural models. So using SE_{WAIC}, the most restrictive scalar model would frequently be deemed less tenable than the metric or configural models (i.e., the top two ranked models), which would be appropriate. But as Figure 2 implies, the metric and configural models would typically be deemed indistinguishable because their WAICs are more similar.

The results of Study 1 imply that among configural, metric, and scalar invariance models, WAIC and DIC_1 will tend to prefer the least constrained model in the presence of substantial DIF, effectively rejecting the H_0 of measurement invariance. In practice, researchers faced with this information must then identify which indicators have DIF in order to establish at least partial measurement invariance. If a researcher identifies only the correct parameters that differ across groups or occasions, the correct partial invariance model should be the most preferred. Figure 21 shows model rankings as in Figure 16, but including the correct partial invariance model in DIF conditions. Consistent with the overlapping lines in Figure 2, WAIC and DIC_1 suggest the fit of the partial invariance model is practically indistinguishable from the configural model, especially when DIF $\geq$ 0.2; however, AIC strongly prefers only partial invariance model in all conditions when DIF is present, which is the more parsimonious model.

In the next section, I investigate the frequency with which true DIF can be detected with small-variance priors for DIF parameters.

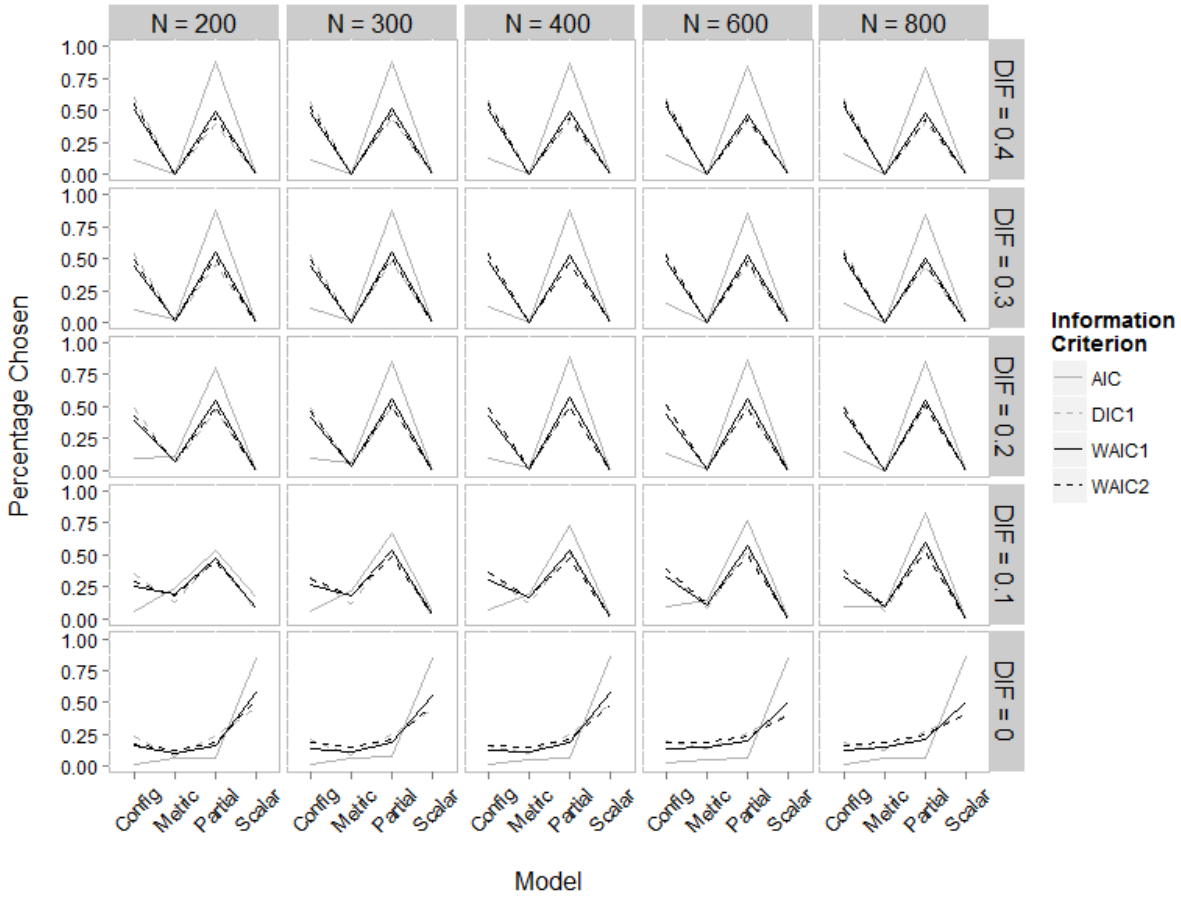

Figure 21. Model preferences (including the partial invariance model) based on ranked AIC, DIC_1, $WAIC_1$, and $WAIC_2$. Results are collapsed across model types (multiple-group vs. longitudinal) and presence of parsimony error.

PART III: Assessing Bayesian Tools for Detecting DIF

Study 1 indicates that substantial uniform and nonuniform DIF cause WAIC and DIC to prefer the least constrained model: configural invariance. In practice, this situation would lead researchers to locate the offending parameter(s). The focus of Study 2 is to evaluate Muthén and Asparouhov's (2012) method for utilizing highly informative priors to identify neglected parameters (in this context, identifying DIF parameters). An advantage of testing invariance in a Bayesian framework is that the model can be parameterized to directly address specific research

questions. So rather than estimating independent factor loadings in each group (λ_1 and λ_2), a common factor loading can be estimated (λ) along with a difference parameter ($\Delta\lambda$) in Group 2. For example, the factor loading in Group 1 is λ, and the factor loading in Group 2 is $\lambda + \Delta\lambda$, so the parameter $\Delta\lambda$ directly represents the degree of nonuniform DIF for that item.

The H_0 of invariance ($\Delta\lambda = 0$) can be tested by checking whether the 95% credible interval—or Bayesian confidence interval (BCI)—for $\Delta\lambda$ or $\Delta\tau$ includes zero. The power to detect DIF is expected to increase with *N* and the effect size (i.e., the actual magnitude of DIF between the populations). I expect Type I error rates for non-DIF items to be less than nominal (i.e., less than 5% when using 95% BCI to test H_0) because the informative prior will constrain the posterior estimates of DIF parameters to remain very close to zero. Informative priors impose the same constraint on DIF parameters that are truly nonzero, so power is also expected to be greater when a larger prior variance (i.e., a less informative prior) is used, allowing data to exert greater influence on the estimated posterior distribution.

Monte Carlo Design for Study 2

Table 1 summarizes the manipulated variables and their levels. The same population models were used to generate longitudinal or multiple-group data for Study 2 as in Study 1 (see Figure 1), with the exception of parsimony error because unmodeled residual correlations are not expected to influence estimates of measurement parameters. In Study 2, I also manipulate the magnitude of standard deviation (σ) used to specify prior distributions for DIF parameters, so this will be a 2 (multiple-group or longitudinal model) × 5 (*N* = 200, 300, 400, 600, or 800) × 5 (magnitude of DIF) × 2 (prior $\sigma = 0.05$ or 0.10) factorial design.

Normal priors with $\mu = 0$ were specified for $\Delta\lambda$ and $\Delta\tau$, with $\sigma = 0.05$ or 0.10, corresponding to approximately 95% probabilities that $\Delta\lambda$ or $\Delta\tau$ falls within ±0.10 or within

±0.20, respectively. This constraint quantifies the prior belief that DIF parameters are unlikely to exceed these limits, which could be considered negligible or small differences on a standardized scale. In practice, researchers should choose prior variances (or corresponding limits) that reflect what would be considered ignorable differences in the scale of the observed variables being modeled. I chose these priors based on practical suggestions in Muthén and Asparouhov (2012) and on Monte Carlo simulation results in Jorgensen, Garnier-Villarreal, Pornprasertmanit, and Lee (2014). Because the priors are a model assumption and do not affect data generation, I analyzed the same data by fitting each model twice (once for each level of prior σ), in addition to the same variance-reduction techniques across levels of *N* and model type that I discussed at the end of the Monte Carlo Design section of Study 1.

I also fit models using MLE as a point of comparison because Muthén and Asparouhov (2012, p. 317) suggested that the small-variance priors should provide information about model modification that is superior to the use of MIs in MLE. Their reason for this claim is that MIs assume only one parameter will be freed and that all other parameter estimates will remain fixed at their current estimates when the model is fit again, whereas Bayesian estimates of nontarget parameters are all provided in a single model. Informative priors should also minimize Type I errors, resulting in fewer spurious modifications. To compare their performance in each condition, I recorded whether the highest significant MI corresponds to the equality constraint that should be freed to reflect true DIF. Models fit using MLE do not include small-variance priors for DIF parameters, but constrain the measurement parameters to exact equality, which is analogous to specifying a prior with $\mu = 0$ and $\sigma = 0$ for the DIF parameter.

Procedure. As in Study 1, I used R to generate data (500 replications per condition), *rstan* to fit Bayesian models to data (monitoring $\hat{R}$ for convergence, saving 1000 post-burn-in

draws from each of three chains; see priors in the Appendix), and *lavaan* to fit models to data using MLE. I fit three models to the data: Model 1 to detect nonuniform DIF and Models 2f and 2b to detect uniform DIF using the forward or backward approach[7], respectively. The backward approach is expected to fail when there is substantial DIF because the analysis model incorrectly equates parameters that differ in the data-generating model—that is, testing an item for DIF requires that the equated anchor items have no DIF, and violating that assumption leads to detecting DIF where it does not exist (Woods, 2009). Using Bayesian estimation, DIF parameters are estimated for all items simultaneously, so no items are used as anchors. Furthermore, the prior constraints on DIF parameters should decrease the frequency of Type I errors. I therefore test both the forward and backward approaches in Study 2 because I hypothesize the backward approach for Bayesian CFA will not result in inflated Type I errors.

To identify nonuniform DIF, Model 1 corresponds to a configural invariance model that is almost a metric invariance model. The factor loadings were constrained nearly to equality by specifying an informative prior with $\mu = 0$ for the DIF parameters ($\Delta\lambda_{1-4}$). The first factor variance was fixed to one in order to set the scale, and the second factor variance was freely estimated. Although the loadings could differ between groups or occasions, the model should still be identified if the priors for DIF parameters adequately constrain the parameter space. Factor means were both fixed to zero, and item intercepts were free to vary across groups or time. The factor correlation and each item's residual correlation were also estimated in the longitudinal conditions. Items with nonuniform DIF were flagged if the $\Delta\lambda$ was unlikely to be

[7] In a frequentist framework, CFA and IRT methods have been compared for detecting DIF, including a comparison of the forward approach commonly use in CFA (i.e., starting with a free baseline model and adding constraints to loadings or discrimination parameters, then adding constraints to intercepts, thresholds, or difficulty parameters) to the backward approach commonly used in IRT (i.e., starting with a fully constrained baseline model, then relaxing location and scaling parameters for each item). Kim and Yoon (2011) and Stark et al. (2006) found that both approaches provide sufficient power to detect DIF, but the backward approach is prone to high Type I error rates.

zero (i.e., when the 95% BCI did not include zero). Using ML estimation, I also recorded whether the largest MI corresponded to the fourth factor loading.

To identify uniform DIF using Model 2b (the backward approach), all factor loadings were constrained to exact equality across groups or time (i.e., no $\Delta\lambda$s will be estimated), and only the first factor variance was fixed to one. This corresponds to starting with the most constrained model, then releasing constraints on measurement parameters (starting with location parameters), which is the more common approach in IRT. Item intercepts were constrained nearly to equality by specifying an informative prior with $\mu = 0$ for the DIF parameters ($\Delta\tau_{1-4}$). The first factor mean was fixed to zero in order to set the location, and the second factor mean was freely estimated. Although the intercepts could differ between groups or occasions, the model should still be identified if the priors for DIF parameters adequately constrained the parameter space. The factor correlation and each item's residual correlation were also estimated in the longitudinal conditions. Items with uniform DIF were flagged if the $\Delta\tau$ was unlikely to be zero (i.e., when the 95% BCI did not include zero). Because Model 2b also contains an incorrectly constrained loading, I recorded whether one of the two largest MIs (instead of only the highest) corresponded to the third intercept.

Using Model 2f (the forward approach) differs from Model 2b only in that the fourth factor loading was freely estimated (no prior constraints) in each group or occasions, and thus the fourth intercept was not constrained to near-equality. This corresponds to the situation where the fourth item is correctly flagged for nonuniform DIF in Model 1 (using the forward approach)—how often this is likely to happen in practice is an outcome in the current investigation. The factor correlation and each item's residual correlation were also estimated in the longitudinal conditions. Items with uniform DIF were flagged if the $\Delta\tau$ was unlikely to be zero (i.e., when

the 95% BCI did not include zero). I also recorded whether the largest MI corresponded to the third intercept.

Results and Discussion

Nonconverged models. Out of all 50,000 data sets (500 replications × 100 conditions), Models 2b and 2f—which estimated DIF in four and three intercepts, respectively—almost always converged on a stable posterior distribution that yielded $\hat{R} < 1.1$ for all model parameters. Nonconvergence occurred frequently with Model 1, which estimated DIF in four factor loadings. Nonconvergence was somewhat more problematic with larger priors, but much more problematic with larger *N*. Figure 22 shows that convergence was never lower than 21%, so there were always at least 105 observations to analyze within each condition.

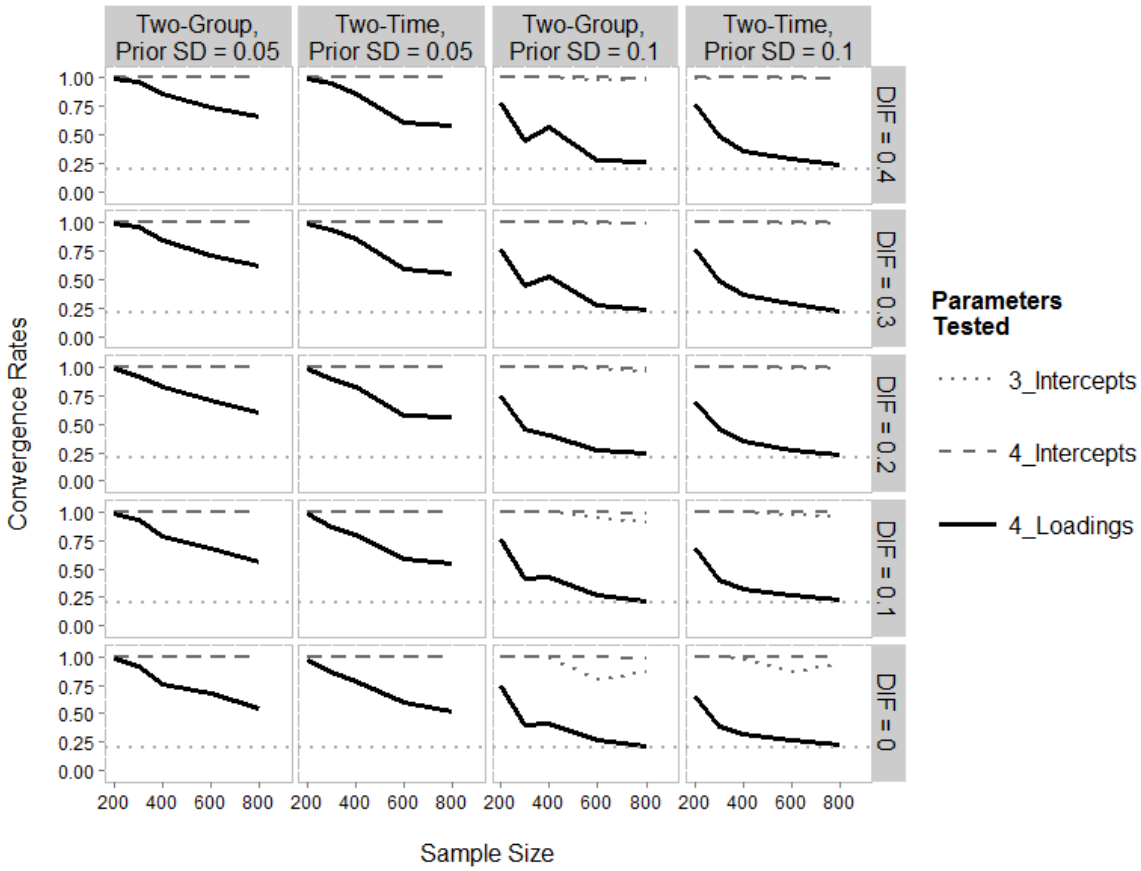

Figure 22. Convergence rates for each model across conditions.

Investigating nonconverged models from all conditions revealed the same pattern. Trace plots of the parameter estimates showed that all three chains converged on a stable estimate of the posterior, but at least one of the chains converged on a different posterior than the other(s). This indicates a multimodal posterior, so two distinct solutions could be found to reproduce the data equally well. For example, the most common solution yielded posterior means of factor loadings, error variances, and the factor *SD* that were close to the true population values. However, in the other solution, the second factor *SD* had a remarkably higher posterior mean (close to 2), the first factor loadings had much smaller posterior means (close to 0.4), and the second factor loadings were close to zero (i.e., estimated DIF was close to −0.4). Although the DIF parameters were constrained to be close to zero, the priors were not informative enough to identify the model, so sometimes a chain would settle in a different region of the posterior. This problem was exacerbated by less informative priors and by larger *N*, which overwhelmed the already insufficient prior.

Refitting the models with more constrained priors on DIF parameters might help the convergence problem, but that would decrease their "power" to detect DIF by shrinking BCIs. In the condition with worst convergence, adding weakly informative priors to factor loadings and the Group-2 factor *SD* solved the convergence problem (see details in Part IV). Because at least 100 replications in each condition converged on the target posterior, and results did not substantially change by changing the priors (i.e., mean parameter estimates and rejection rates were similar), I removed nonconverged solutions and proceeded with analysis.

Variability of parameter estimates. Because Monte Carlo sample sizes were unequal across conditions, partial-η^2 was calculated from ANOVA results using Type III *SS*. As shown in Table 3, Estimates of the second latent mean were largely influenced by the magnitude of DIF

in the population (partial-η^2 = .785), the model (partial-η^2 = .16), and moderately by the interaction between DIF and model (partial-η^2 = .08); all other factors had negligible effects (partial-η^2 < .01). Figure 23 illustrates that bias in latent means grows with DIF, and more so in Model 2f (in which the invalid constraint on λ_4 is released) than in Model 2b. The same factors affected estimates of the second latent *SD*. Figure 24 shows bias only when an invalid constraint is placed on λ_4. Model 1 places an approximate equality constraint on λ_4, whereas Model 2b places an exact equality constraint on λ_4, but both levels of constraint lead to similar bias in the latent *SD*. For these models, greater DIF leads to greater bias, but for Model 2f there is no bias at any level of DIF, which characterizes the interaction between DIF and Model.

Table 3

Effect Sizes (partial-η^2) of Monte Carlo Factors on Parameter Estimates

	Latent Parameter		DIF Parameter	
Monte Carlo Factor	*M*	*SD*	$\Delta\lambda_4$	$\Delta\tau_3$
N			19.0%	54.3%
DIF	78.5%	13.0%	6.7%	96.1%
Prior σ			42.5%	70.7%
Type (multiple-group / -time)				3.5%
Model (i.e., Model 1, 2b, or 2f) [a]	15.9%	20.5%		
N × DIF				37.3%
N × Prior σ			6.7%	6.6%
DIF × Prior σ				55.6%
DIF × Type				1.4%
DIF × Model [a]	8.2%	8.5%		
N × DIF × Prior σ				3.4%

Note. Only effects with η^2 > 1% are shown. Type III *SS* used to calculate partial-η^2.

[a] The effect of Model was only included as a factor in ANOVAs for latent *M* and *SD*.

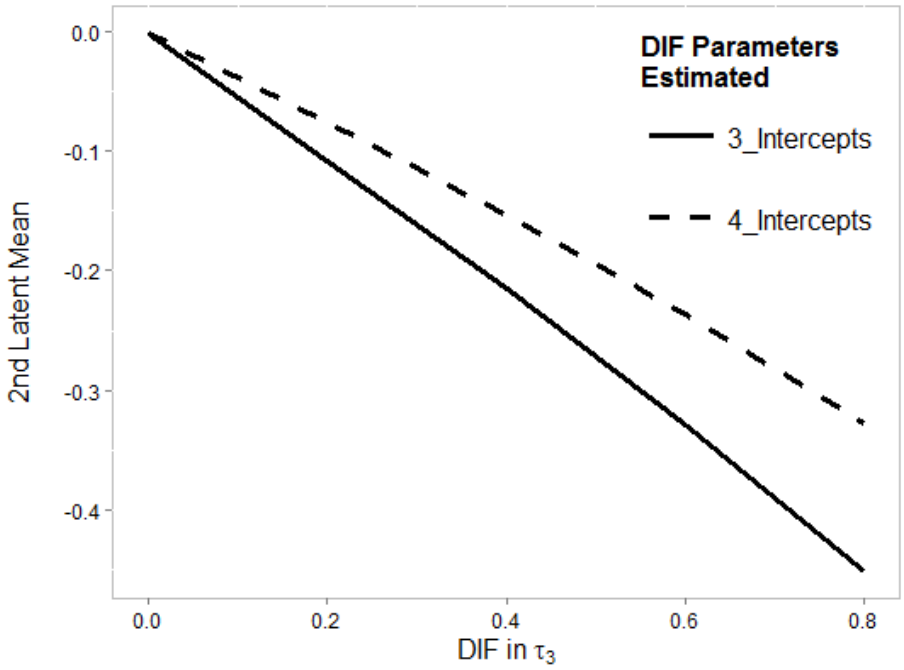

Figure 23. Bias in the second latent mean grows in magnitude as DIF increases.

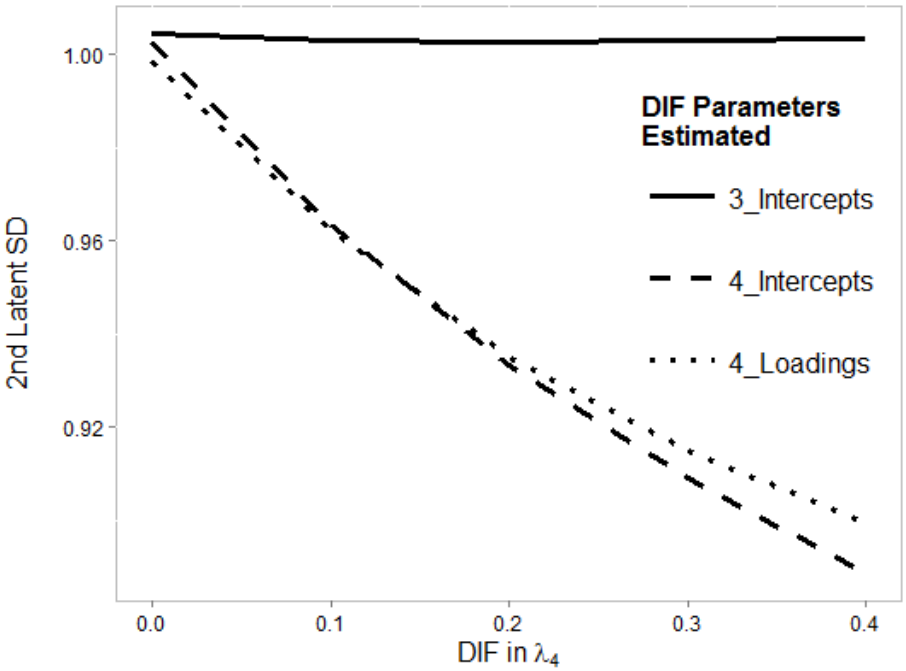

Figure 24. Average posterior mean of the second latent *SD* by DIF.

The DIF parameter in Model 1 ($\Delta\lambda_4$) is largely influenced by the magnitude of *N* and the prior σ, moderately by their interaction, and moderately by the magnitude of DIF (see Table 3). Figure 25 illustrates the nature of the interaction: $\Delta\lambda_4$ is greater in absolute value when the prior σ is less informative, but the difference between prior σ = 0.05 and 0.10 is smaller when larger *N* overwhelms the prior. Greater DIF in the population is expectedly reflected in lower (i.e., more negative) $\Delta\lambda_4$ estimates, but it is also noteworthy that the misfit due to the invalid constraint on $\Delta\lambda_4$ also manifests in higher estimates of $\Delta\lambda_{1-3}$, which are compensated by the negatively biased latent *SD* in Figure 24.

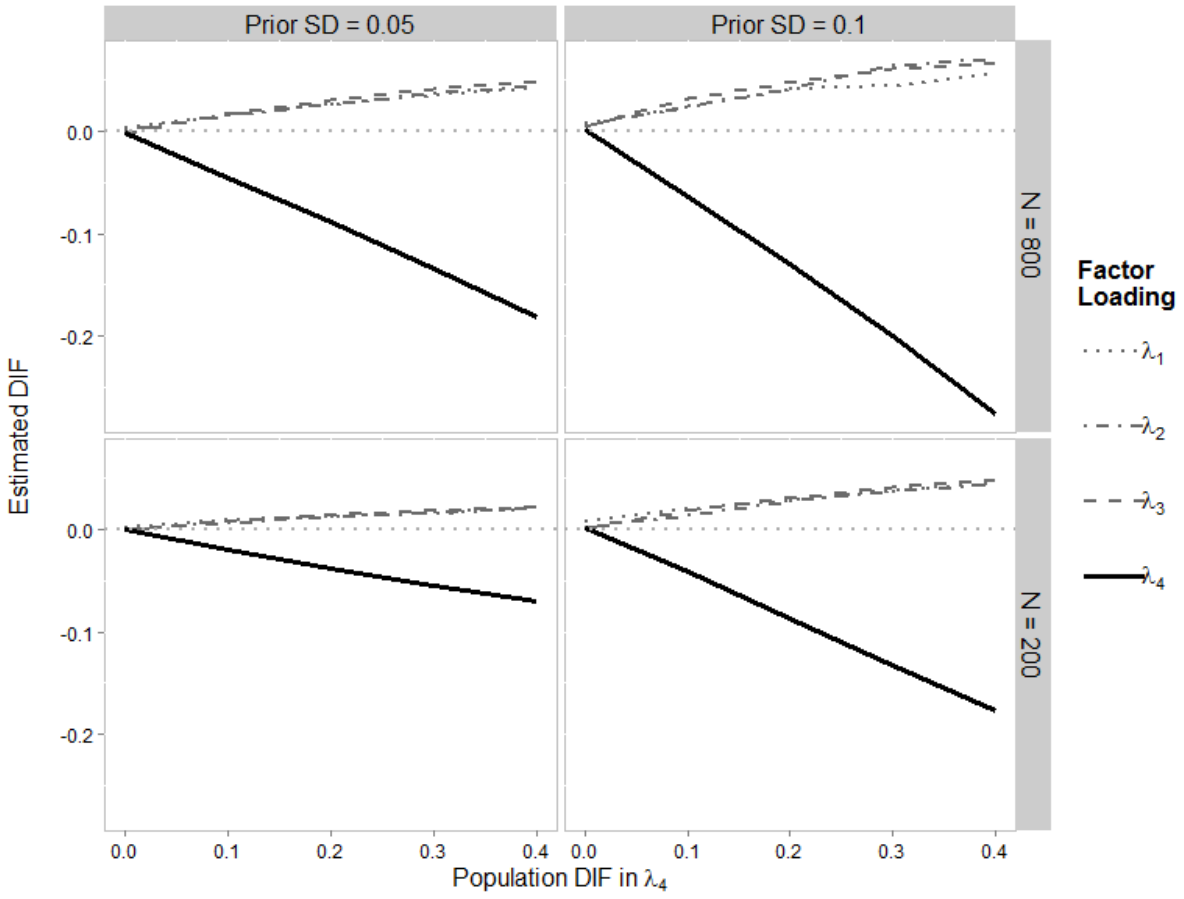

Figure 25. Average posterior mean of Δλs by DIF, prior σ, and *N*.

The DIF parameter in Model 2b ($\Delta\tau_3$) is largely influenced by the magnitude of DIF, *N*, and the prior σ, as well as interactions among these factors (see Table 3). Figure 26 looks very similar to Figure 25, but the *y*-axis has a wider range, illustrating why the effect sizes for $\Delta\tau_3$ are so much larger than for $\Delta\lambda_4$. The nature of the interactions is similar: $\Delta\tau_3$ is greater in absolute value when the prior σ is less informative, but the difference between prior σ = 0.05 and 0.10 is smaller when larger *N* overwhelms the prior. As for $\Delta\lambda_4$, greater DIF in the population is reflected in lower (i.e., more negative) $\Delta\tau_3$ estimates, and the misfit due to the invalid constraint on $\Delta\tau_3$ also manifests in higher estimates of $\Delta\tau_{1-2}$ and $\Delta\tau_4$, which are compensated by the negatively biased latent mean in Figure 23. Similar results were found for Model 2f.

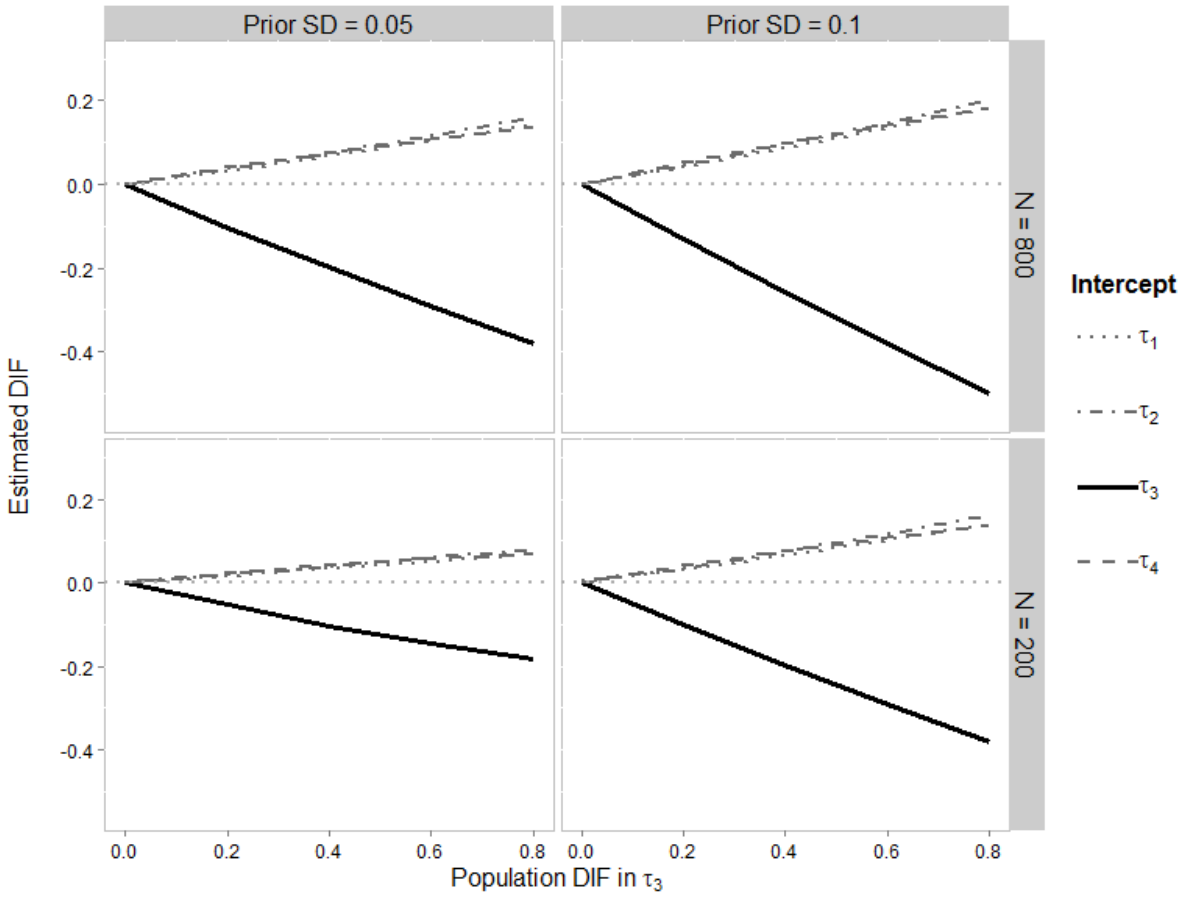

Figure 26. Average posterior mean of Δτs by DIF, prior σ, and *N*.

Rejection rates. I report rejection rates for all DIF estimates, as well as the frequency with which the ML modification indices. When DIF exists ($\Delta\lambda_4$ and $\Delta\tau_3$), this is power: $p(\text{Test}^+ \mid \text{DIF}^+)$. In the absence of DIF, this is the Type I error rate: $p(\text{Test}^+ \mid \text{DIF}^-)$. Because control of familywise Type I error rates typically leads to a reduction in power, researchers may choose to compromise by allowing inflation of Type I error rates, so long as the number of falsely rejected hypotheses is only a small proportion (e.g., 5%) of all rejected hypotheses (Maxwell & Delaney, 2004). Therefore, I also report the false discovery rate (FDR = $p(\text{DIF}^- \mid \text{Test}^+)$, which is the complement of the positive predictive value (PPV = 1 − FDR = $p(\text{DIF}^+ \mid \text{Test}^+)$. FDR is discussed further in Maxwell and Delaney (2004, pp. 230–234).

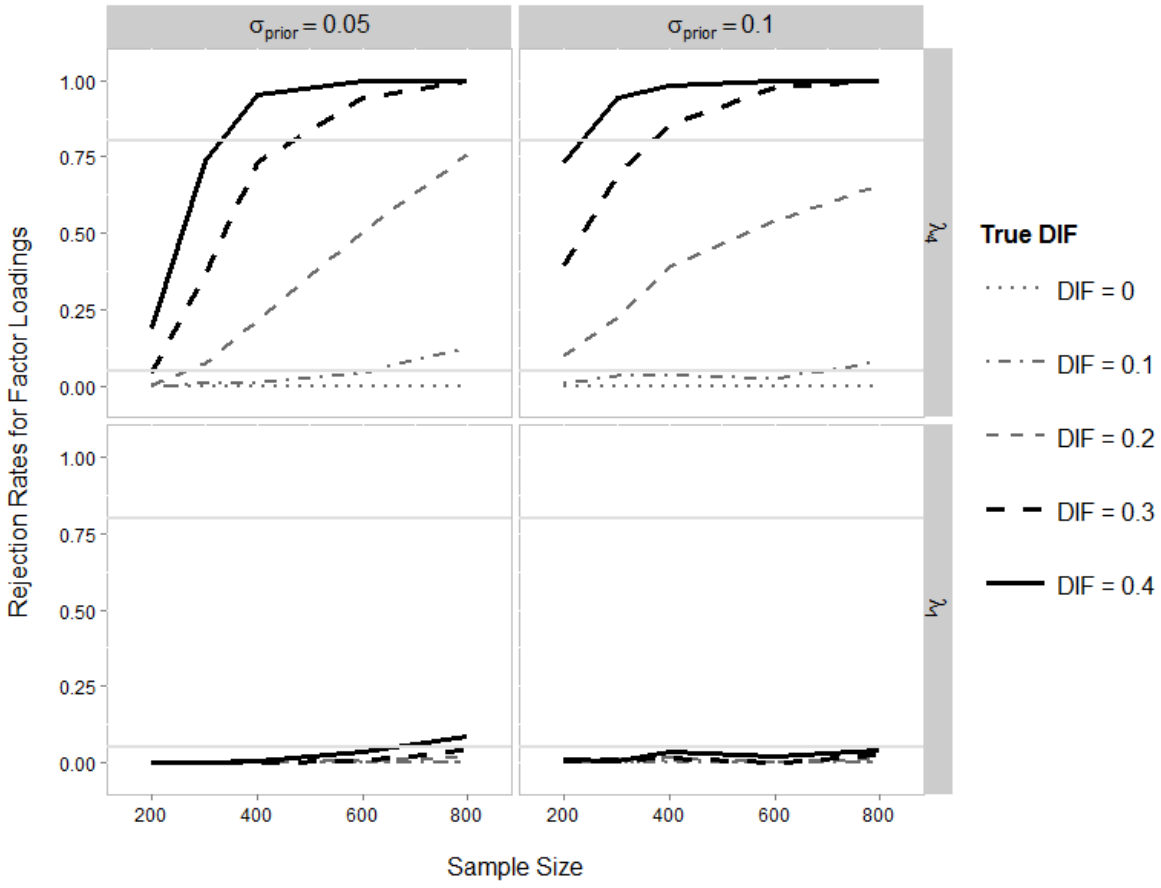

Figure 27. Rejection rates for $\Delta\lambda$s by DIF, prior σ, and *N*. The bottom panels depict only one parameter (λ_1) for which DIF = 0, but the same pattern was observed for λ_2 and λ_3.

As seen with DIF estimates, rejection rates were primarily influenced by magnitudes of population DIF, *N*, and prior σ. Because model type (multiple-group or longitudinal) had only negligible influence on rejection rates, results are collapsed across those conditions. The top panels of Figure 27 show adequate power (~80%) to detect small DIF ($\Delta\lambda_4 = 0.2$) only when $N = 800$, but moderate or large DIF is detectable when $N > 300$, particularly using prior $\sigma = 0.1$. The dotted lines show that in the absence of DIF, Type I error rates are close to zero, and negligible DIF ($\Delta\lambda_4 = 0.1$) is also typically detected in less than 5% of samples. The bottom panels show that other factor loadings without DIF also have very low Type I error rates (0% in over half of the conditions).

The top panels of Figure 28 show very high power (100% for $\Delta\tau > 0.2$ when $N > 200$), even for negligible DIF when (a) $N > 400$ and prior $\sigma = 0.05$ or (b) $N = 800$ and prior $\sigma = 0.1$. Again, Type I error rates are close to 0% in the absence of DIF. However, the bottom panels reveal that when there is substantial DIF ($\Delta\tau > 0.2$), Type I errors are highly inflated for non-DIF items ($\Delta\tau_2$ and $\Delta\tau_4$ show similar results to $\Delta\tau_1$). The reason for this can be seen in Figure 23. Because τ_3 is actually lower in Group 2 (or Occasion 2), the constraint imposed on τ_3 by the informative priors causes positive biases in Group 2's τ_3. As DIF increases, the latent mean in the second group (or occasion) becomes more negatively biased to compensate for the invalid near-equality constraint, so that the observed item means can be more accurately reproduced ($\bar{x}_3 = \tau_3 + \lambda_3 \times \overline{factor}$). The negatively biased Group-2 factor mean in turn causes the other item intercepts to become positively biased, so that their means can be more accurately reproduced. The effect of the invalid constraint on the DIF estimate is distributed among multiple non-DIF items, so Type I errors can be controlled by only releasing the constraint for the largest DIF estimate. Figure 29 shows Type I errors near 0%, with no loss of power.

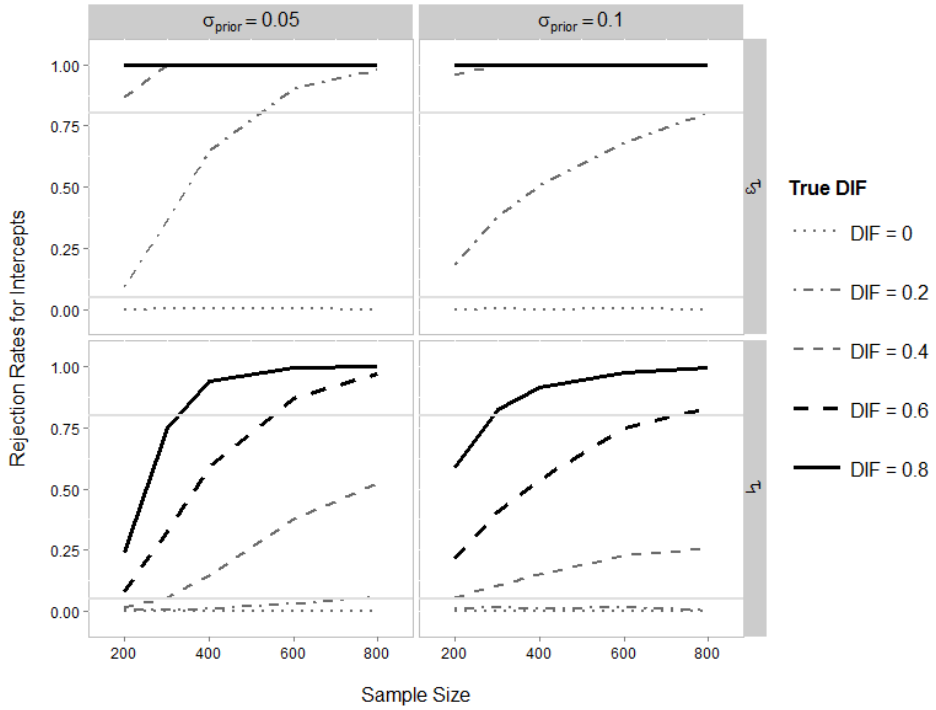

Figure 28. Rejection rates for Δτs by DIF, prior σ, and *N*. The bottom panels depict only one parameter (τ_1) for which DIF = 0, but the same pattern was observed for τ_2 and τ_4.

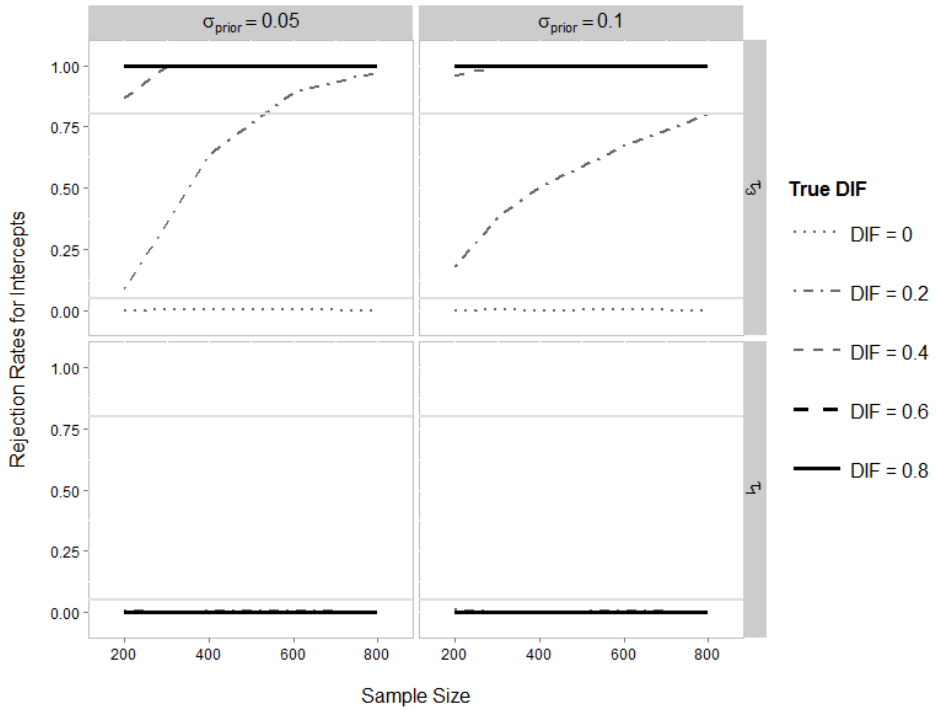

Figure 29. Maximum-DIF rejection rates for Δτs by DIF, prior σ, and *N*. The bottom panels depict only τ_1, but the same pattern was observed for τ_2 and τ_4.

The FDRs in Figure 30 reinforce the conclusion that constraints on intercepts must be released sequentially rather than all at once. As DIF and *N* increase, mean FDR for intercepts approaches 75% because DIF would be detected in all four intercepts, yet only one of those parameters actually has DIF (i.e., three out of four discoveries are false). FDR may be more acceptable using a sequential method, but fitting a sequence of models (as is required when using MIs in MLE) was not part of the current investigation. Figure 30 also shows that FDRs for factor loadings are much more acceptable (typically below or close to 5%).

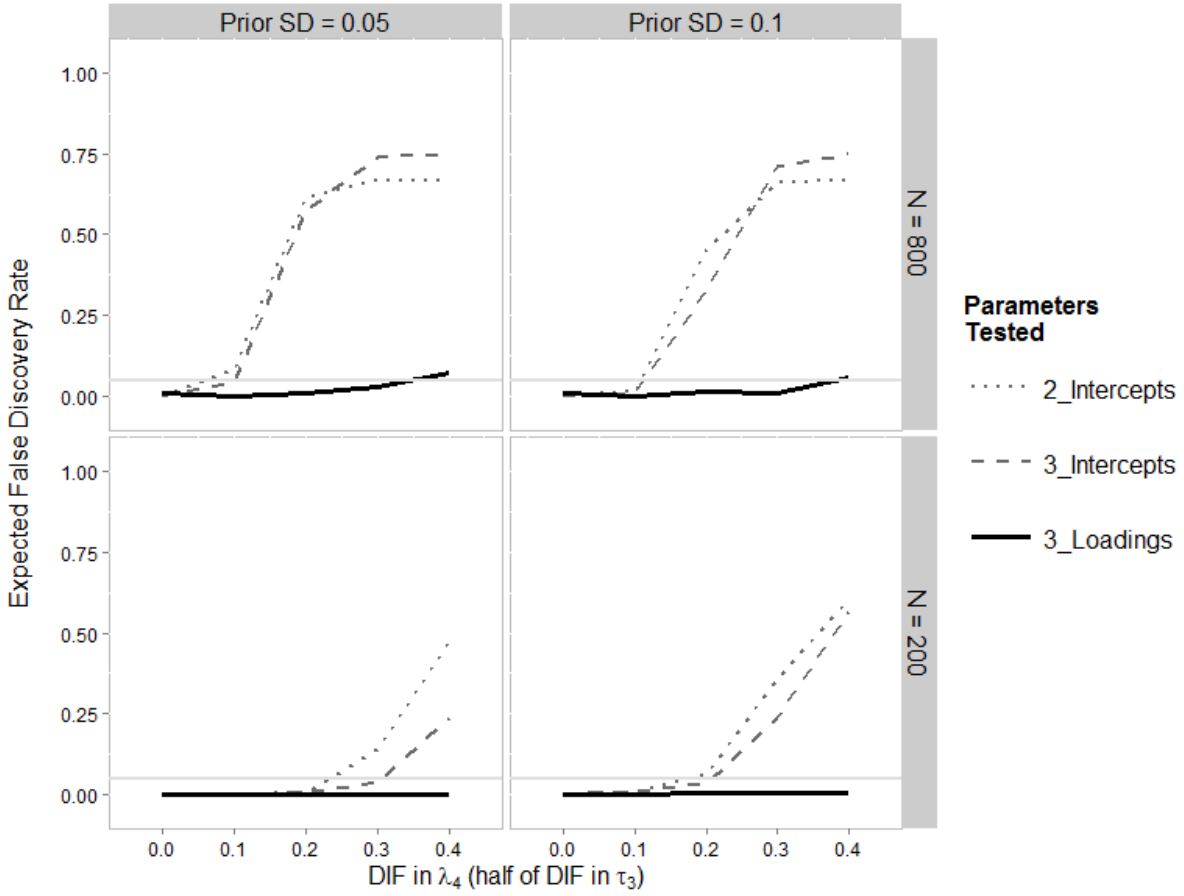

Figure 30: False discovery rates (FDR) for DIF in intercepts and factor loadings by population DIF, prior σ, and *N*.

The current best practice for detecting DIF in CFA is to use MIs to search for equality constraints that should be released, with a Bonferroni correction for the number of constraints being tested. For Models 1 and 2b (testing four loadings and all four intercepts, respectively), the corrected $\alpha = .05 / 4 = .0125$, which is associated with a critical $\chi^2(1)$ value of 6.24, assuming the research focuses only on MIs for measurement parameters constrained to equality across groups or occasions. For Model 2f (testing three intercepts, assuming nonuniform DIF in Item 4 was detected in Model 1), the corrected $\alpha = .05 / 3 = .0167$, which is associated with a critical $\chi^2(1)$ value of 5.73.

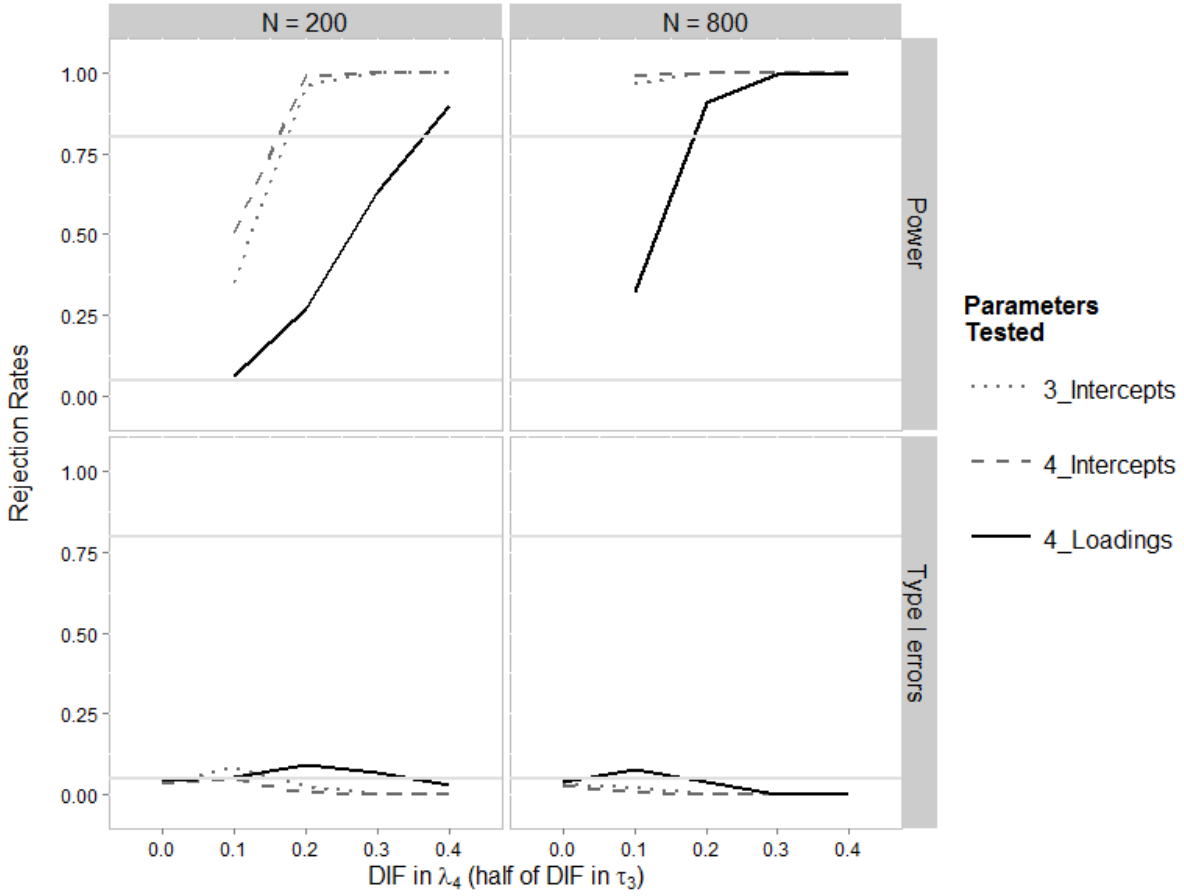

Figure 31: Power and Type I error rates for detecting DIF in intercepts and factor loadings using modification indices in MLE.

To contrast the Bayesian method under investigate to current best practices in CFA, the top row of Figure 31 presents the power to detect true DIF in each model. The bottom row of Figure 31 shows Type I error rates, which indicate the proportion of replications in which the largest significant MI among validly constrained measurement parameters showed evidence of DIF. In the presence of large DIF, MIs had near 100% power to detect it, in which case the Type I error rates were near zero. But when DIF was small or absent, Type I error rates were closer to nominal levels (i.e., around 5%), which are higher than the error rates using the Bayesian small-variance priors (see Figure 32). Just as model type had negligible effect on DIF estimates and rejection rates using Bayes, MIs were similar for longitudinal and multiple-group models, so rejection rates in Figure 31 are collapsed across those conditions.

Figure 32 directly compares the Type I error rates using MLE and Bayesian estimation. Because MLE does not incorporate priors, the grey lines are the same in the left and right panels. The grey lines (indicating Type I error rates for MIs) are only close to zero when DIF and *N* are large, in which case the largest MI typically corresponds to the parameter with DIF. Black lines (indicating Type I error rates using BCIs) remain much closer to zero, so fewer Type I error rates will be made using small-variance priors with Bayesian estimation than using MIs with MLE.

Figure 33 directly compares power using MLE and Bayesian estimation. Again, grey lines are the same in the left and right panels. For factor loadings, MIs and BCIs have similar power, except that MIs have much better power to detect large DIF with smaller *N* when priors are more informative (prior $\sigma = 0.05$). For intercepts, MIs typically have greater power to detect smaller DIF when *N* is small, but this discrepancy is negligible when *N* is large and priors are more informative.

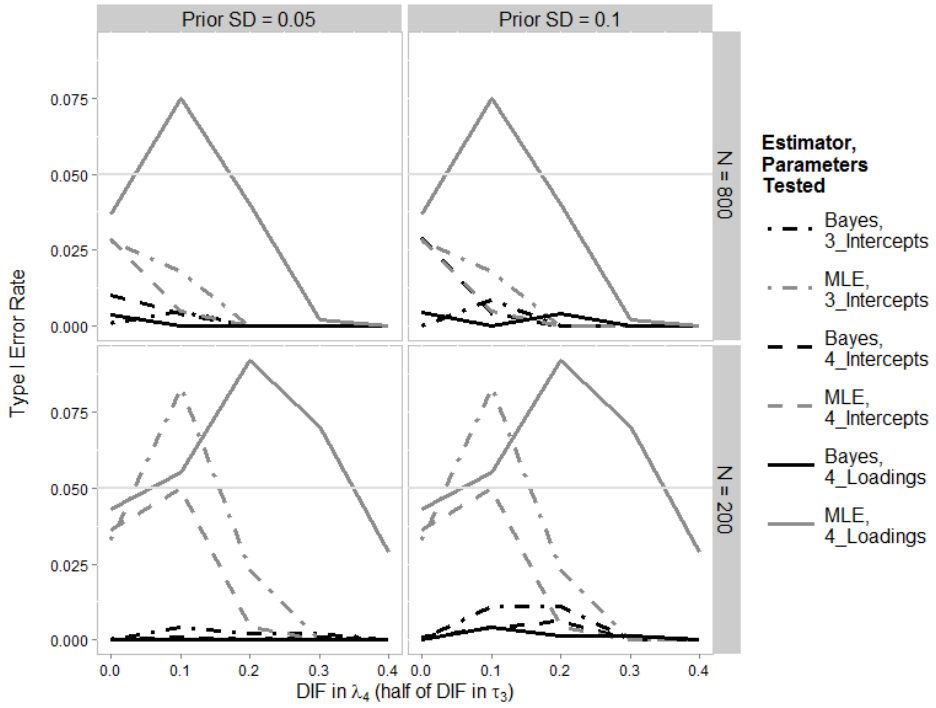

Figure 32. Type I error rates by DIF, prior σ, and *N*.

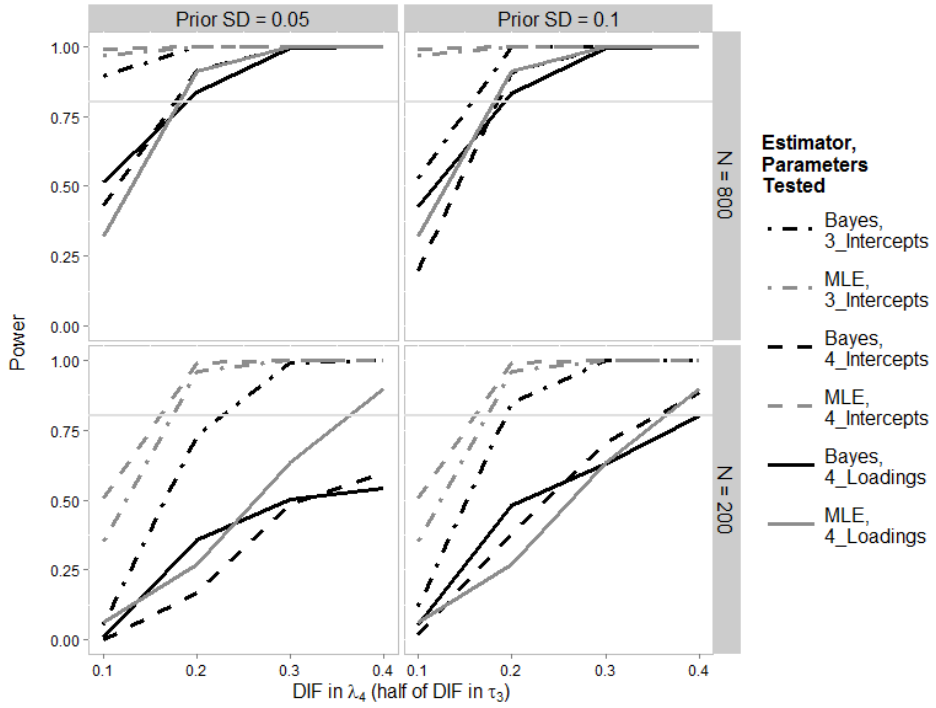

Figure 33. Power by DIF, prior σ, and *N*.

As expected, power to detect DIF—using MIs or BCIs—increases with sample size (*N*) and effect size (DIF). In the case of BCIs, prior variance had the expected effect on nonuniform DIF detection: less informative priors yielded greater power. This was generally the case with intercepts as well, except that for small *N* and small DIF, more informative priors yielded slight greater power to detect uniform DIF. No noticeable differences were found between multiple-group and longitudinal models.

PART IV: General Discussion

Measurement equivalence is a necessary assumption if latent parameters are to be compared across different contexts (e.g., subpopulations or occasions of measurement), so testing this assumption is the focus of much research. Much practical advice can be found for testing degrees of measurement equivalence using CFA, particularly when using MLE for continuous indicators. For example, configural invariance can be tested by inspecting whether the model fits well, using the χ^2 test statistic or alternative fit indices (AFIs) as criteria. Metric and scalar invariance can be tested by calculating $\Delta\chi^2$ statistics or ΔAFIs (e.g., ΔCFI < .01; Cheung & Rensvold, 2002) as criteria for judging whether equality constraints are tenable. Advice is also available for other special cases, such as when using WLS to estimate CFA with binary and ordinal indicators (e.g., Kim & Yoon, 2011), and although many questions remain unanswered, this is a very active area of research.

Bayesian estimation methods (e.g., Gibbs sampling) are becoming more popular due to their availability and ease of use in population SEM software packages such as M*plus*, but in contrast to the quantity of advice available using MLE or WLS, very little advice has been offered for testing measurement equivalence when using Bayesian methods. As of the time of this writing, I found only one methodological article about Bayesian estimation for testing

measurement equivalence in CFA (Asparouhov & Muthén, 2014), and it involves using a new technique called *alignment*, which is similar to rotation methods in EFA. The alignment method appears to perform well in initial simulation studies, but for a real scale administered in eight countries, Cieciuch, Davidov, Schmidt, Algesheimer, and Schwartz (2014) found strikingly different results compared to traditional methods (i.e., full scalar invariance for all 19 items using Bayesian alignment vs. partial invariance of only 12 items found in previous studies). More simulation studies are needed to discover under what conditions the alignment method might produce invalid results. Additional Bayesian methods have been developed in an IRT framework (e.g., multilevel parameterization; Verhagen & Fox, 2013), which could also be adapted for CFA.

The question of how best to test measurement equivalence in Bayesian CFA is therefore an open one. Although an advanced user of general Bayesian modeling software could manage to program posterior distributions of differences in model fit between nested models, most practicing researchers would only use statistical tools that are easily calculated or readily available. The only such tool for model comparison in Bayesian SEM is the DIC provided in standard output of M*plus* and Amos. The motivation for Study 1 is to assess how often researchers using DIC would correctly prefer the most parsimonious scalar invariance model when equality constraints for measurement parameters are valid, as well as how often the least parsimonious configural model would be preferred when those equality constraints are invalid. Because DIC is a generalization of AIC, DIC's performance was compared with MLE results, as well as with the newly proposed WAIC, which is greater generalization of AIC than DIC because it utilizes the full posterior distribution rather than a central-tendency point estimate.

Study 1 showed that AIC has lower Type I error rates than DIC or WAIC, in the sense

that in the absence of DIF, AIC is more likely to prefer the most parsimonious (scalar invariance) model. Nonetheless, Type I errors using DIC and WAIC are rare, so they would all be good tools to selecting the appropriate measurement model when measurement equivalence is a valid assumption. However, AIC makes more Type II errors than WAIC and DIC, particularly at smaller *N* and small-to-moderate DIF. That is, in the presence of DIF, DIC and WAIC more consistently prefer the configural model than AIC does, leading to the appropriate conclusion that some measurement parameters differ across contexts. These results were similar regardless of model type (longitudinal and multiple-group) or presence of parsimony error (a correlated residual in the population that was constrained to zero in the model). Although WAIC has less sampling variability than DIC, model choices would be nearly equivalent in practice, so researchers can confidently use the readily available DIC to choose an appropriate measurement model when fitting a CFA with a Bayesian estimation method.

After concluding that DIF exists, a researcher's next goal would be to identify which equality constraints are invalid, to establish partial measurement invariance so latent parameters could still be compared across contexts. In MLE, only MIs are available to identify items with DIF, but past research has noted several limitations of MIs, notably that specification searches often lead to an incorrect final model (MacCallum et al., 1992). Muthén and Asparouhov (2012, 2013) suggested that using Bayesian estimation methods, parameters could be approximately (rather than precisely) constrained with small-variance priors. Study 2 was an investigation of how well this method would work for identifying whether indicators had uniform or nonuniform DIF.

Bayesian DIF parameters are all estimated simultaneously with other model parameters, so rather than searching for items to free one at a time, Muthén and Asparouhov (2012)

insinuated that all invalid constraints could be identified in one step. But because invalid constraints are compensated for by multiple other parameters[8], Bayesian estimates of truly nonzero DIF parameters are not independent of DIF estimates that are truly zero. Because the effects of unmodeled DIF are distributed across other model parameters, the best practice is to only release the constraint for the largest DIF estimate, and then fit the model again. Thus, one predicted advantage over MIs is lost: the ability to identify all DIF parameters in a single step, rather than one at a time in a specification search (Muthén & Asparouhov, 2012).

The only other bases for discriminating between Bayesian posterior DIF estimates and ML MIs are to compare (a) their power to detect an invalid constraint, (b) their Type I error rates in the presence of an invalid constraint, and (c) their Type I error rates in the absence of invalid constraints. In each case, using BCIs would result in less frequent Type I errors than using MIs. Bayesian and ML methods both have high power when *N* and DIF are at least moderately large, but when *N* or DIF are small, MIs have noticeably greater power to detect uniform DIF than BCIs do; however, power to detect nonuniform DIF is typically similar for BCIs and MIs.

Limitations and Future Directions

Although the current investigations included several conditions that would be commonly encountered in practice, results presented here may not generalize to other conditions. Limitations of the current investigation are discussed below.

Measurement equivalence was tested for multiple-group and longitudinal models, and

[8] A parameter could be constrained to equal a constant (typically zero or one), or two or more free parameters could be constrained to equality. The degree to which these constraints are invalid can bias other parameter estimates. For example, an omitted cross-loading can exaggerate the correlation between the two factors on which the item truly loads. In the context of testing measurement equivalence, invalid equality constraints across groups (or occasions) introduce bias not only in the measurement parameter but also in the associated latent parameters. For example, if a factor loading truly is truly lower in Group 2 than in Group 1, constraining the groups' loadings to equality will result in an estimate that is a compromise between the two true values. To minimize the effect of that invalid constraint on model fit, Group 2's factor variance will be underestimated, which in turn causes all other loadings for that factor to be overestimated. The fewer items there are to share that balance, the worse the bias.

results differed only negligibly. However, only two-group or two-occasion situations were investigated, and the two groups had equal sample sizes. This mimics common situations such as pretest–posttest designs or studies comparing men to women or treatment to control groups. But situations with unbalanced groups are also common (e.g., comparing clinical to general populations). Short (2014) found that fit indices in MLE (including AIC) tend to have lower power to detect nonequivalence of intercepts as the discrepancy between group sizes increases, although the power to detect nonequivalence of factor loadings was generally unaffected by sample-size ratio, especially at larger *N*. The degree to which sample-size ratio affects DIC, WAIC, or estimates of DIF parameters is a topic that warrants further investigation.

Studies comparing more than two groups (e.g., several races or countries to whom a scale was administered) or more than two occasions are also common. In these cases, the method for detecting DIF that was the focus of Study 2 may be too unwieldy in practice. A single reference group or occasion would be chosen (e.g., the group on which the scale was normed, or the first occasion of measurement), and DIF parameters would be specified to characterize how measurement parameters in each other group (or occasion) differ from those of the reference group. If, for example, two groups both have significantly lower factor loadings than the reference group, then another model may need to be specified to test whether those two groups differ from each other. Because small-variance priors identify DIF well only when the largest is released, multiple comparisons for each parameter make the sequential specification search more complex, just as it does for using MIs. The Bayesian method results in far fewer Type I errors than using MIs, but future research is required to establish whether that result generalizes to several groups or occasions.

Only four indicators per construct were used, which is near the minimum necessary (i.e.,

three) for just-identification of a construct, but scales with many more items are common, especially during scale development or assessment, when items with DIF can be identified for removal. The number of indicators was shown to influence how much impact DIF has on other model parameters. It is possible that with enough indicators per construct, the impact of DIF will be distributed across so many other loadings or intercepts that a sequential search would be unnecessary to control Type I errors. Because more steps in a specification search lead to more potential errors (MacCallum et al., 1992), discovering the number of items necessary to simplify the process of DIF detection is an important avenue for further research.

Model 1 in Study 2 was under-identified enough to cause as low as 21% convergence in conditions with large *N*. If a practicing researcher is confronted with evidence of underidentification, then the model must include weakly informative priors for the Group-1 factor loadings and Group-2 latent *SD*. The most problematic condition ($N = 800$, $\sigma_{prior} = 0.1$, DIF = 0) had 21% convergence for the multiple-group model and 22% convergence for the longitudinal model. I specified weakly informative priors for factor loadings ~ lognormal($\mu = -0.2$, $\sigma = 0.3$), with the bulk of its density roughly between $\lambda = 0.5$ and 1.2, and for the factor *SD* ~ lognormal($\mu = 0$, $\sigma = 0.25$), with the bulk of its density roughly between 0.6 and 1.6. These weakly informative priors led to 100% convergence for the multiple-group and longitudinal models. Parameter estimates and rejection rates were similar the results found in Part III, with bias ranging between −0.005 and 0.007. Future research would be helpful to establish guidelines for choosing appropriately weak priors in cases when an applied researcher has little or no information about differences in factor variances across groups or occasions.

Choosing priors for the DIF parameters themselves would also be helpful, so I offer some practical advice here. When no clear substantive or theoretical choice is apparent, priors for DIF

parameters may need to be chosen based on characteristics of the data. I used priors that seemed appropriate for the scale of the observed indicators and latent constructs (both were standard normal). Because the total indicator variances were close to one (exactly one in the population), the error variance estimates could not exceed one, nor could the factor loadings. Thus, factor loadings and error variances were originally specified with a uniform prior between zero and one. Because the observed variances between groups or occasions did not differ greatly, it would be reasonable to assume that the factor variances differ minimally, so I originally specified a uniform prior between zero and two for the second factor variance. As the follow-up simulation showed, convergence problems caused by the "hard" boundaries in a uniform prior can be solved with priors whose high-density regions correspond roughly to the same limits but are unbounded. Priors for DIF parameters were specified as $N(\mu = 0, \sigma = 0.05 \text{ or } 0.10)$. The values for σ seemed appropriate because 95% of values would be within ± 0.1 or ± 0.2, respectively. Allowing that much DIF to occur corresponds to allowing small amounts of DIF to be considered approximately equal, which would have negligible effect on latent parameter estimates (see Figures 12 and 13).

Assuming measurement equivalence holds, "impact" has been a term used to indicate true differences in latent parameters (Stark et al., 2006). No true differences in latent parameters existed in the population, yet impact can commonly be expected in practice and is certainly an important research question. Impact could complicate the effect of DIF on latent parameters. For example, if Group 2 has a higher factor loading than Group 1, but Group 2's latent *SD* is also higher, then the factor loadings may appear equivalent if both latent variances are both fixed to one. When using constrained Bayesian estimates of DIF parameters to identify DIF, latent parameters are allowed to differ, just as they are allowed to differ in constrained models when

using MIs to identify DIF with MLE; therefore, the magnitude and direction of impact relative to DIF should not prevent adequate DIF detection with good control of Type I errors. However, this is an open question left for future research.

Lastly, although little practical difference was found between the model choices using $WAIC_1$, $WAIC_2$, and DIC_1, it is unclear whether these results generalize to other situations in which competing models are compared in terms of fit and parsimony. Configural, metric, and scalar invariance models share an identical functional form; they differ only in terms of equality constraints. Thus their claims about how data are generated from a population process are nearly identical. It is entirely possible that WAIC and DIC model choices would differ more substantially when the models being compared are not so similar (e.g., common-factor vs. simplex models to describe a large set of similar items responded to in a sequence). Because the theoretical support for preferring WAIC to DIC is so strong (Gelman et al., 2013; Vehtari & Gelman, 2014; Vehtari & Ojanen, 2012), it is surprising to find so little practical difference in their applied behavior, especially given the evidence in Study 1 confirming the predicted smaller sampling variability of WAIC. Further research is warranted to distinguish between the relative practical values of DIC and WAIC.

Conclusions

In conclusion, practicing researchers interested in using Bayesian CFA to investigate measurement equivalence can be confident that WAIC and DIC are useful tools for deciding whether a search for DIF is warranted. If the scalar invariance model is not the optimal model, then small-variance priors for DIF parameters can be added to the metric and scalar models in order to identify nonuniform and uniform DIF, respectively. Contrary to Muthén and Asparouhov's (2012) suggestion, if any DIF parameter's BCI excludes zero, it is best to release

the prior constraint only for the largest DIF estimate and fit the model again to search for any additional DIF parameters. This appears especially important when testing for uniform DIF, as Type I errors were very frequent when testing all intercepts simultaneously rather than testing only the largest $\Delta\tau$; however, this may not generalize to conditions other than those investigated in Study 2. As long as only the parameter with the largest DIF estimate is freed, then Type I errors are unaffected by whether the forward or backward approach is used; however, the forward approach (i.e., first identifying DIF in factor loadings, then searching for uniform DIF only among items with equal loadings) yields greater power than the backward approach. Care should be taken to assign appropriate priors to DIF parameters, by taking into account the variability of the data and any information available in past research. A sensitivity analysis can reveal whether results are influenced by the location of the prior distribution.

When both measurement and latent parameters are allowed to differ (to some degree) across contexts, lack of identification may manifest in multiple chains stabilizing on different posterior distributions. Convergence to a target posterior distribution can be helped by specifying reasonable, weakly informative priors with "soft" boundaries; specifying a fixed boundary, even if is near a logical limit for the parameter, can cause convergence problems. In Study 2, such a constraint was apparently necessary for Model 1 (testing factor loadings for DIF), particularly with larger *N*. If more informative priors are required to identify the model, a sensitivity analysis may also be appropriate to ascertain the degree to which results are influenced by the choice of prior distribution(s).

Bayesian methods investigated in Study 2 are more complicated to implement than using MIs in MLE, but the advantage is that the incorporation of small-variance priors for DIF parameters result in fewer Type I errors. Thus, specification searches using BCIs would result in

fewer deviations from the correct model than specification searchers using MIs. However, both methods are tedious, necessitating several models to be fit in sequence. This disadvantage of both MIs and BCIs makes the recently proposed alignment method (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2013) highly desirable because it simplifies the process to fitting only one or two models. Future research must, however, indicate the degree to which the alignment method provides valid results under a wide variety of conditions.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology, 66*, 1–7. doi:10.1111/bmsp.12004

Arbuckle, J. L. (2012). *IBM SPSS Amos 21 user's guide*. Chicago, IL: IBM.

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21*(4), 495–508. doi:10.1080/10705511.2014.919210

Bayarri, M. J., & Berger, J. O. (2000). *P* values for composite null models. *Journal of the American Statistical Association, 95*, 1127–1142. doi:10.1080/01621459.2000.10474309

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606. doi:10.1037/0033-2909.88.3.588

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods, 15*(2), 111–123. doi:10.1037/a0019625

Bollen, K. A. (1989). *Structural equations with latent variables*. Hoboken, NJ: Wiley.

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation*

Modeling, 21(1), 1–19, doi:10.1080/10705511.2014.856691

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83. doi:10.1111/j.2044-8317.1984.tb00789.x

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*, 230–258. doi:10.1177/0049124192021002005

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*(article 982), 1–10. doi:10.3389/fpsyg.2014.00982

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences, 33*, 137–193. doi:10.1017/S0140525X09991567

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research, 32*(2), 208–252. doi:10.1177/0049124103256130

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12*(3), 343–367. doi:10.1207/s15328007sem1203_1

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529. doi:10.1080/00273170701382864

Fan, X., & Sivo, S. A. (2009). Using Δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling, 16,* 54–69. doi:10.1080/10705510802561311

Fox, J.-P., & Glas, C. A. W. (2005). Bayesian modification indices for IRT models. *Statistica Neerlandica, 59*(1), 95–106. doi:10.1111/j.1467-9574.2005.00282.x

Gelman, A. (2013, December 29). RE: a Waic example! [Online forum comment]. Retrieved from https://groups.google.com/d/msg/stan-users/FJUDq2ALtEg/w4w-MGjSno8J

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*. doi:10.1007/s11222-013-9416-2

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–807. doi:10.1.1.142.9951

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. doi:10.1214/ss/1177011136

Gelman, A., & Shalizi, C. (2013a). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66,* 8–38. doi:10.1111/j.2044-8317.2011.02037.x

Gelman, A., & Shalizi, C. (2013b). Rejoinder to discussion of "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology, 66,* 76–80. doi:10.1111/j.2044-8317.2012.02066.x

Gelman, A., & Stern, H. S. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328–331. doi:10.1198/000313006X152649

Hu, L. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453. doi:10.1037/1082-989X.3.4.424

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi:10.1080/10705519909540118

Iversen, G. R. (1984). *Bayesian statistical inference*. Newbury Park, CA: Sage.

Johnson, P. E. (2015). rockchalk: Regression estimation and presentation (version 1.8.91) [R package]. Available from the Comprehensive R Archive Network: http://cran.r-project.org/

Johnson, V. E. (2004). A Bayesian χ^2 test for goodness-of-fit. *The Annals of Statistics, 32*(6), 2361–2384. doi:10.1214/009053604000000616

Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Jorgensen, T. D., Garnier-Villarreal, M., Pornprasertmanit, S., & Lee, J. (2014). *Detecting misspecification in Bayesian confirmatory factor analysis: A Monte Carlo simulation study*. Manuscript submitted for publication.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*(2), 137–155. doi:10.1207/s15327906mbr2502_1

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. doi:10.1080/01621459.1995.10476572

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212–228. doi:10.1080/10705511.2011.557337

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*(4), 457–474. doi:10.1007/BF02296338

Kruschke, J. K. (2013). Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology, 66*, 45–56. doi:10.1111/j.2044-8317.2012.02063.x

Levy, R. (2011). Bayesian data–model fit assessment for structural equation modeling. *Structural Equation Modeling, 18*(4), 663–685. doi:10.1080/10705511.2011.607723

Levy, R., & Hancock, G. R. (2007). A framework of statistical tests for comparing mean and covariance structure models. *Multivariate Behavioral Research, 42*(1), 33–66. doi:10.1080/00273170701329112

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine, 28*(25), 3049–3067. doi:10.1002/sim.3680

MacCallum, R. C. (1986). Specification searches in covariance structural modeling.

Psychological Bulletin, 100, 107–120. doi:10.1037/0033-2909.100.1.107

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113–139. doi:10.1207/S15327906MBR3801_5

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods, 11*(1). 19–35. doi:10.1037/1082-989X.11.1.19

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490–504. doi:10.1037/0033-2909.111.3.490

Maiti, S. S., & Mukherjee, B. N. (1990). A note on distributional properties of the Jöreskog–Sörbom fit indices. *Psychometrika, 55*(4), 721–726. doi:10.1007/BF02294619

Marsh. H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341. doi:10.1207/s15328007sem1103_2

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling, 14*(4), 611–635. doi:10.1080/10705510701575461

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592. doi:10.1037/0021-9010.93.3.568

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: Model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology, 66,* 68–75. doi:10.1111/j.2044-8317.2012.02067.x

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419. doi:10.1037/a0024377

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling, 7*(1), 36–73. doi:10.1207/S15328007SEM0701_02

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. doi:10.1037/a0026802

Muthén, B. O., & Asparouhov, T. (2013). *BSEM measurement invariance analysis* (Web note 17). Retrieved May 14, 2013, from http://www.statmodel.com/examples/webnote.shtml

Muthén, L. K., & Muthén, B.O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Plummer, M. (2013). *JAGS version 3.4.0 user manual*. Retrieved from http://sourceforge.net/projects/mcmc-jags/files/Manuals/

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41*(3), 227–259. doi:10.1207/s15327906mbr4103_1

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17*(1), 1–14. doi:10.1037/a0026804

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification

errors in structural equation models. *Sociological Methodology, 17*, 105–129. doi:10.2307/271030

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561–582. doi:10.1080/10705510903203433

Short, S. D. (2014). *Power of alternative fit indices for multiple-group longitudinal tests of measurement invariance* (Unpublished doctoral dissertation). University of Kansas, Lawrence, KS.

Song, X.-Y., & Lee, S.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. Hoboken, NJ: Wiley.

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371–384. doi:10.1007/BF02294623

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639. doi:10.1111/1467-9868.00353

Stan Development Team. (2014). *Stan modeling language: User's guide and reference manual* (version 2.4.0). Retrieved from http://mc-stan.org/

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. doi:10.1037/0021-9010.91.6.1292

van de Schoot, R., Hoijtink, H., Hallquist, M. N., & Boelen, P. A. (2012). Bayesian evaluation of inequality-constrained hypotheses in SEM models using M*plus*. *Structural Equation Modeling, 19*(4), 593–609. doi:10.1080/10705511.2012.713267

Vehtari, A., & Gelman, A. (2014). *WAIC and cross-validation in Stan*. Unpublished manuscript. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection, and comparison. *Statistics Surveys, 6*, 142–228. doi:10.1214/12-SS102

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology, 66*, 383–401. doi:10.1111/j.2044-8317.2012.02059.x

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243. doi:10.1037/a0027127

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*, 3571–3594.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle, *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford.

Whitaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education, 80*(1), 26–44. doi:10.1080/00220973.2010.531299

Widamin, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*(1), 16–37. doi:10.1037/1082-989X.8.1.16

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*(1), 42–57. doi:10.1177/0146621607314044

Appendix

Prior Distributions for Model Parameters

Parameter	Prior Distribution
Factor loadings (λ)	$U(0, 1)$
Nonuniform DIF($\Delta\lambda$)	$N(0, \sigma_{prior})$
Intercepts (τ)	$N(0, 5)$
Uniform DIF($\Delta\tau$)	$N(0, \sigma_{prior})$
Residual variances (θ)	$U(0, 1)$
Group-2 Factor Mean	$N(0, 5)$
Group-2 Factor *SD*	$U(0, 2)$
Group-2 Factor Correlation	$U(-1, 1)$