

# An illustration of Bayesian approximate measurement invariance with longitudinal data and a small sample size

International Journal of  
Behavioral Development  
2020, Vol. 44(4) 371–382  
© The Author(s) 2019  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/0165025419880610  
journals.sagepub.com/home/ijbd



Sonja D. Winter<sup>1</sup> and Sarah Depaoli<sup>1</sup>

## Abstract

This article illustrates the Bayesian approximate measurement invariance (MI) approach in *Mplus* with longitudinal data and small sample size. Approximate MI incorporates zero-mean small variance prior distributions on the differences between parameter estimates over time. Contrary to traditional invariance testing methods, where exact invariance is tested, this method allows for some “wiggle room” in the parameter estimates over time. The procedure is illustrated using longitudinal data on college students’ academic stress as it changes in the period leading up to and right after an important midterm. Results show that traditional invariance testing methods come to a standstill due to the small sample size. Bayesian approximate MI testing was able to identify non-invariant parameters, after which a partially invariant model could be estimated.

## Keywords

Bayesian statistics, approximate measurement invariance, longitudinal data, small sample size

Research within the behavioral sciences commonly examines growth, or change, over time in the context of a variety of constructs. Many of these constructs are measured through scales that tap into underlying behavioral phenomena of interest. A common approach for assessing whether these behavioral constructs are stable over time is measurement invariance (MI) testing (e.g., see Edossa, Schroeders, Weinert, & Artelt, 2018). However, a new approach has been recently developed that is more flexible and can be applied to cases with smaller sample sizes than the traditional approach. This is called *approximate* MI (Muthén & Asparouhov, 2013), and it is conducted through Bayesian methods. The current article will introduce Bayesian MI testing and show an application related to changes in college student academic stress surrounding a midterm. Recommendations for researchers interested in implementing this approach will also be provided throughout.

## Longitudinal MI

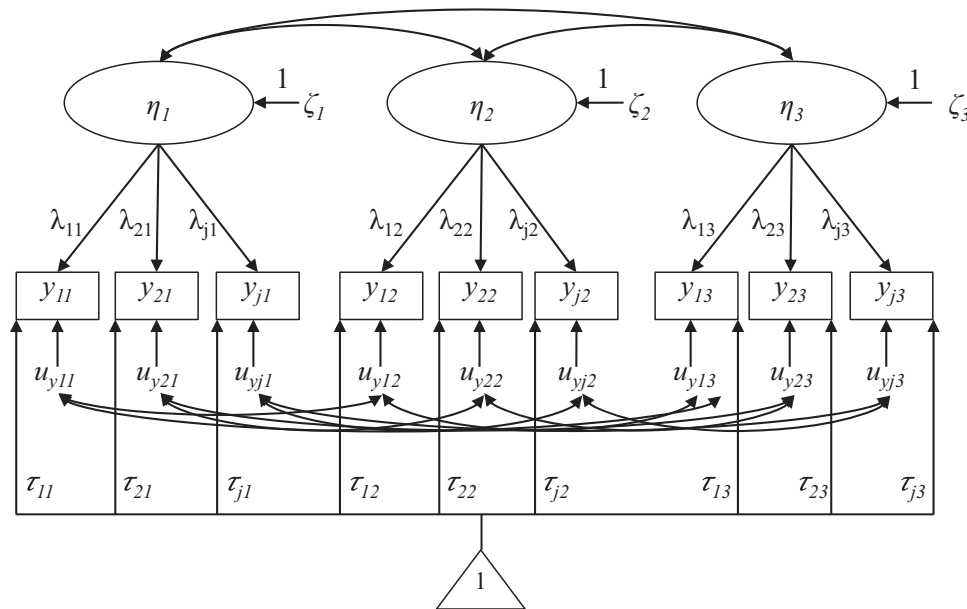
MI testing was originally developed to assess differences and similarities in how a construct is measured across groups through multiple-group confirmatory factor analysis (MGCFA; Jöreskog, 1971; Meredith, 1964a; Meredith, 1964b; Meredith, 1993). However, MI is also a critical assumption of any model assessing change over time. If a scale is invariant over time, it means that participants’ answers to the questions on the scale are related to the underlying construct in the same way at each observed time point (i.e., they do not *vary* across groups). If a scale is not invariant over time, using a composite score of that scale (i.e., the mean across all items) to assess change over time through a latent growth model (LGM) leads to significant bias in growth parameter estimates (Liu & West, 2018). Thus, the degree of non-invariance needs to be identified to ensure proper scale interpretation across time.

Assessment of MI occurs through a systematic series of steps (Vandenberg & Lance, 2000). Similar to MGCFA, longitudinal MI testing starts with a CFA for each time point included in the analysis (see Figure 1). A set of factor loadings ( $\lambda_{ji}$ ) and intercepts/thresholds ( $\tau_{ji}$ ) is estimated for each time point separately. This specification is called configural invariance as long as the same items are used at each time point. Each subsequent step in MI assessment involves constraining an additional set of parameters to be equal across time. Three levels of invariance are of interest: metric (factor loadings), scalar (intercepts/thresholds), and strict (residual variances) invariance. At each step, model fit can be compared to a less restrictive model to assess whether MI at that step holds or not. As these models are nested, a likelihood-ratio  $\chi^2$  difference test is often used to statistically test the difference in model fit. However, this test is sensitive to sample size (Bentler & Bonett, 1980) and has the potential of an inflated Type I error rate (Yuan & Chan, 2016). Change in the comparative fit index (CFI; Bentler, 1990) has been proposed as an alternative method for model comparison in invariance testing (Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008; Meade, Lautenschlager, & Hecht, 2005; Rutkowski & Svetina, 2014; Svetina, Rutkowski, & Rutkowski, 2019). Its advantages over the likelihood-ratio  $\chi^2$  difference test is that it is independent of model complexity and sample size (Cheung & Rensvold, 2002). Several cutoffs have been proposed for defining “significant” change in model fit, ranging from  $-.01$  (Cheung & Rensvold, 2002) to

<sup>1</sup> University of California, Merced, USA

## Corresponding author:

Sonja D. Winter, Psychological Sciences, University of California, Merced, 5200 Lake Road, Merced, CA 95343, USA.  
Email: swinter@ucmerced.edu



**Figure 1.** Example of a Longitudinal CFA Model With Three Measurement Occasions.

Note. CFA = confirmatory factor analysis.

–.0017 (Meade et al., 2005). A true consensus on which cutoff to use has not been reached, leaving the decision up to the researcher.

**A note on testing for partial invariance.** At each step in the MI testing procedure, it is possible to examine partial invariance, where some item parameters are constrained equal across time and others are freely estimated. Modification indices are used to examine which specific item parameters are variant over time. Starting with the item parameter that would improve model fit the most (i.e., the largest decrease in the  $\chi^2$ ), parameters are relaxed across time points in an iterative manner. Testing for partial invariance is a data-driven process that could be difficult to replicate (Maccallum, Edwards, & Cai, 2012). Researchers are urged to only release equivalence constraints for certain parameters if a preexisting theoretical foundation is present (Byrne, Shavelson, & Muthén, 1989; Vandenberg & Lance, 2000). Additionally, any MI testing results that involve some level of partial invariance should be replicated to support the validity of the measurement model. In addition, this approach is only feasible if at least a few items are invariant over time (Byrne et al., 1989).

**Challenges of traditional MI.** For items measured on a continuous scale, full scalar invariance is the best-case outcome of invariance testing for further model development; all observed items are considered measurement invariant at each time point. However, full invariance assumes all parameters are *exactly* the same at each time point, which is highly improbable (Vandenberg & Lance, 2000). Ignoring major departures from full invariance can lead to severely biased (i.e., inaccurate) results (Liu & West, 2018). Conversely, small deviations from MI over time often do not have harmful effects on substantive conclusions (Meuleman, 2012). Indices used for model comparison in invariance testing are not always sensitive enough to distinguish between minor and major departures from invariance, especially if sample sizes are small. At the other extreme, the likelihood-ratio  $\chi^2$  difference test becomes overly

sensitive to minor departures from invariance when the potential locations on non-invariance are increased, such as through inclusion of more items or more time points (Bentler & Bonett, 1980; Svetina et al., 2019; Yuan & Chan, 2016).

Switching to a partially invariant model might offer a solution, but as stated above, this solution might be too dependent on the specific sample observed. In addition, known limitations of this approach are that it is computationally more difficult because more parameters are being estimated. Both full and partial exact invariance have their disadvantages and can force a researcher to conclude that they cannot meaningfully compare covariances, regression estimates, and latent means over time. In such a situation, an approach that incorporates the small differences over time into the measurement model, while still estimating unbiased and meaningful growth parameters, would provide a solution. One such approach is Bayesian approximate MI.

### Bayesian Approximate Invariance

**A primer on Bayesian statistics.** Bayes theorem was developed over 250 years ago (Bayes, 1764; Price, 1765) and has been discussed in more detail elsewhere (e.g., Gelman et al., 2013). In short, Bayesian inference is made up of three parts: the prior distribution, the data, and the likelihood. Together, they create the posterior distribution. Prior information reflects the researcher's belief about a parameter. When a researcher has strong prior beliefs, this can be translated into a narrow, *informative* prior. If a researcher has uncertain prior beliefs, a wide, *diffuse* prior can be used. The posterior distribution finds a compromise between the prior distribution and the data likelihood. The more informative the prior, the closer the posterior will resemble its distribution. When models cannot be analytically derived and directly estimated, as is the case for structural equation modeling (SEM), Bayesian inference is implemented through a Markov chain Monte Carlo (MCMC; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) sampling

algorithm, using the Gibbs sampler (or akin; Gelfand & Smith, 1990). These samplers iteratively sample from multivariate posterior distributions of all parameters, sequentially updating the chain for each parameter. The posterior distribution does not rely on asymptotic arguments such as normality. To assess the posterior distribution, its mean or median can be used as a point estimate, while its standard deviation reflects precision. Bayesian methods allow for a more flexible approach to MI, which can greatly benefit the assessment of behavioral constructs over time. For a thorough introduction to the technical implementation of Bayesian methods, as well as a detailed explanation of the MCMC process, we refer to Depaoli and van de Schoot (2017), Gelman and colleagues (2013), Kaplan (2014), and Robert and Casella (2011).

There are many different reasons why a researcher may prefer to use Bayesian methods to traditional, frequentist (e.g., maximum likelihood (ML)) estimation (see also van de Schoot & Depaoli, 2014; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). The main reasons for using Bayesian methods include: (1) the models are too “complex” for traditional methods to handle (i.e., models are intractable under frequentist settings), (2) only relatively small sample sizes are available and prior information must be used to compensate, (3) the researcher *wants* to include background information into the estimation process, and (4) there is preference for the types of results that Bayesian methods produce.

What we mean by this last point is that Bayesian methods provide a more complete picture of population parameters, and researchers are able to narrate full distributions rather than a simple point estimate. As a result, Bayesian methods can be a rich source of information, regardless of how priors are used or what sort of model is being implemented. A nice example of this is provided in Kruschke (2013), which illustrates how Bayesian statistics can make a model as simple as a *t*-test more informative to applied researchers. Specifically, Bayesian methods allow researchers to interpret the *entire* posterior distribution created by the MCMC process, instead of just a point estimate (akin to the frequentist perspective). While the MCMC process is unlikely to sample from the complete posterior distribution, it does provide a representative approximation of the posterior distribution (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012). This feature means that the researcher will have a better understanding of believable values surrounding the parameter, conditional on the observed data, rather than relying on a single value that was converged upon akin to frequentist methods.

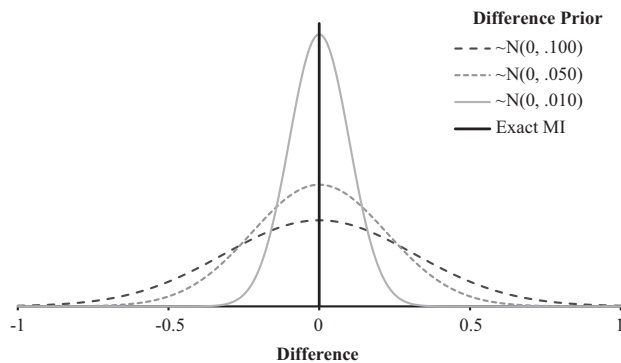
**Bayesian model comparison and fit.** There are several indices that can be used to compare models estimated using Bayesian methods. We will first discuss the two indices that are available in *Mplus*, after which we will also introduce two additional indices. In *Mplus*, Bayesian model assessment and comparison can be determined using the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). This index is computed for each competing model, and the model with the smallest DIC value is preferred.

Once a model is selected, model fit can be assessed with posterior predictive checks, which look at the predictive quality of the posterior estimates (Lynch & Western, 2004; Scheines, Hoijtink, & Boomsma, 1999). One such check is called the posterior predictive *p* value (PPP; Gelman, Meng, & Stern, 1996; Meng, 1994). At each MCMC iteration, a data set is generated based on the updated parameter estimates at that iteration. Both the observed

and the generated data set are compared to the model implied covariance matrix, resulting in two discrepancy statistics. *Mplus* uses the  $\chi^2$  goodness-of-fit statistic as their discrepancy statistic. The PPP value reflects the proportion of  $\chi^2$  values obtained from the generated data that exceed the  $\chi^2$  values obtained from the original data. A consistently lower  $\chi^2$  value for the generated data than for the original data (reflected in a PPP value close to 0) implies that the model does not fit the data well. PPP values that approach .5 are indicative of good model fit.

The use of the DIC and the PPP value for model selection has been criticized (e.g., Hoijtink & van de Schoot, 2017). In addition, both the DIC and PPP value are not completely Bayesian, in that their computation still relies on point estimates, instead of using the entire posterior distribution. The widely applicable information criterion (WAIC; Watanabe, 2010) and leave-one-out cross-validation (LOO; Geisser & Eddy, 1979; Gelfand & Dey, 1994) have been recommended for model comparison and fit assessment (Gelman, Hwang, & Vehtari, 2014; Vehtari, Gelman, & Gabry, 2017). In short, LOO estimates the out-of-sample predictive fit of a model by repeatedly partitioning the full data in a training partition of  $n - 1$  and a single holdout observation. The posterior is estimated for the training partition and used to predict the holdout observation. The WAIC can be interpreted in a similar manner to the DIC, however, it is more fully Bayesian as it uses the complete posterior distribution. In addition, the WAIC is asymptotically equivalent to LOO, but computationally more efficient. Furthermore, several approximate fit indices were recently proposed for Bayesian SEM (Garnier-Villareal & Jorgensen, 2019; Hoofs, van de Schoot, Jansen, & Kant, 2018). These fit indices can be computed for Bayesian analyses run through R packages such as *blavaan* (Merkle & Rosseel, 2018) and *Rstan* (Stan Development Team, 2018). Even though these packages are capable of implementing approximate MI, their implementation differs from *Mplus* and the methods described here. For example, in *blavaan*, the posterior estimate of a parameter in a reference group is used as the mean of the prior for that same parameter in the additional groups. Thus, approximate invariance is implemented through priors placed directly on the parameters, as opposed to priors placed on the differences between parameters (see *Approximate invariance with small variance priors* below). While this approach has been applied in a multiple-group context (Garnier-Villareal, 2019), it appears that it does not translate easily to a longitudinal context (i.e., the model does not converge). For this reason, and because of its widespread use and relatively user-friendly implementation, we conducted the current illustration in *Mplus*.

**Approximate invariance with small variance priors.** A promising new application of Bayesian methods is approximate MI, which works through applying very narrow prior distributions to difference parameters comparing the intercepts and loadings of like items over time (Muthén & Asparouhov, 2013). From the traditional exact MI standpoint, we would expect all of these difference parameters to be exactly zero (which is a *very* strict assumption). By applying a narrow prior distribution centered around zero, approximate MI allows for some “wobble room” around a difference of 0 (i.e., creating an approximate zero, rather than forcing an exact zero; see Figure 2), allowing researchers to draw conclusions about development over time even when full MI does not hold. In *Mplus*,



**Figure 2.** Illustration of Three Specifications of the Small Variance Difference Prior in Approximate MI as Compared to Full MI (Solid Line). Note. MI = measurement invariance.

this method can be implemented through placing priors on the differences between parameters. For example:

```
F1 by y11-y13 (lam1_1-lam1_3);
F2 by y21-y23 (lam2_1-lam2_3);
F3 by y31-y33 (lam3_1-lam3_3);
Model Priors:
DO(1,3) DIFF(lam1_#-lam3_#) ~ N(0, 0.01);
```

The above syntax places a normal prior distribution with a mean of zero and a variance of .01 on the difference between factor loadings of the first (y11, y21, and y31), second (y12, y22, and y32), and third item (y13, y23, and y33), respectively. By giving each factor loading a unique label (e.g., lam3\_1 for the first item measured at the third time point), they are allowed, within the “wiggle room” of the difference prior, to vary across time points.

Approximate MI (1) leads to more accurate parameter estimates, (2) detects parameters in specific groups or at specific times that cause the non-invariance without the use of modification indices, (3) is easily generalized to many more groups or time points, (4) allows for small cross-loadings in models with multiple factors, and (5) is not susceptible to typical MI estimation issues (e.g., negative variances; Muthén & Asparouhov, 2013; Pokropek, Davidov, & Schmidt, 2019). In addition, approximate MI can serve as a better performing alternative to partial MI when differences between parameter estimates are relatively small (van de Schoot et al., 2013).

Muthén and Asparouhov (2013) suggest implementing approximate MI by first estimating a model that assumes approximate MI for all parameters of interest (i.e., factor loadings and intercepts). Results from this initial analysis can inform which parameters are invariant and which show non-invariance that cannot be accounted for through a small variance prior.

The performance of approximate MI hinges on one important assumption: Deviations from invariance are small and non-systematic within and across groups or time points (Muthén & Asparouhov, 2013). This assumption is violated when, for example, an item intercept is invariant across all time points except for one, for which this intercept is substantially lower. The small variance difference prior will subsequently overestimate the non-invariant intercept and underestimate the remaining invariant intercepts. Estimation bias in the item intercepts subsequently biases estimation of the latent factor means and variances. This is known as the *alignment issue*. To resolve the impact of the alignment issue, one

can freely estimate parameters with a systematic deviation from invariance, resulting in a model that combines approximate and partial MI.

One issue that remains is how to specify the small variance difference prior so that it is small enough to accurately reflect small deviations from invariance and identify larger, more meaningful deviations from invariance. Asparouhov, Muthén, and Morin (2015) proposed a method that uses the DIC and PPP values to select a prior variance. They suggested running at least five models, each with a different prior variance specification, starting with a very small variance (e.g.,  $v = .001$ ). To decide which variance should be used, one needs to consider (a) speed of convergence, (b) the PPP value (or the 95% confidence interval for the difference between the observed and replicated  $\chi^2$  values), and (c) the DIC. The prior variance no longer needs to be increased further when model fit differences between models become small, or reverse direction.

It should be noted that some research has indicated that DIC and PPP are not able to compare different small variance specifications to each other in a consistent manner (Hojtink & van de Schoot, 2017). Specifically, as sample size increases, the fit indices, and also the posterior distributions, are no longer affected by the prior specification of these small variance priors, but solely reflect the model's fit to the data likelihood. In addition, the PPP lacks power in detecting model misfit of minor model parameters (i.e., differences, cross-loadings, residual covariances) when small variance priors are used (Jorgensen, Garnier-Villareal, Pomprasernan, & Lee, 2019). Thus, severe deviations from MI might be masked by small variance priors, resulting in a PPP that indicates acceptable model fit.

An alternative is the fit measure introduced by Hoijtink and van de Schoot (2017) called the prior-posterior predictive  $p$  value (PPPP). The PPPP value was recently implemented in *Mplus* version 8, as a test for minor model parameters (i.e., small variance priors). That is, if the test is not rejected (i.e., PPPP is  $>.05$ ), the minor parameters are assumed to come from a  $N(0, v)$  distribution (Asparouhov & Muthén, 2017). At this time, the PPPP value has only been implemented for priors that are directly applied to parameters and will not be computed for small variance priors placed on *differences* between factor loadings or intercepts. Thus, it cannot be used in the context of approximate MI.

**Use in applied literature.** Since the initial development of Bayesian approximate MI, it has been used several times in the applied literature. Through a (non-exhaustive) Google Scholar search of “Bayesian approximate invariance” on June 18, 2019, we identified 26 published papers and dissertations using the method (see Table 1). Almost all studies used approximate MI to test for invariance across groups (e.g., see Braeken & Blömeke, 2016; Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Kelcey, McGinn, & Hill, 2014). Four studies did include data measured at several time points, but only focused on MI across groups (Chan, Ivarsson, Stenling, Yang, & Chatzisarantis, 2015; Cieciuch, Davidov, Algesheimer, & Schmidt, 2018; Davidov et al., 2015; Zercher, Schmidt, Cieciuch, & Davidov, 2015). We found five notable exceptions that used the method in a longitudinal context (Falkenström, Hatcher, Skjulsvik, Larsson, & Holmqvist, 2015; Hawes et al., 2018; Jordalen, Lemyre, Solstad, & Ivarsson, 2018; Seddig & Leitgöb, 2018; Williams, Chandola, & Pendleton, 2018).

There is some variability in how the technique is applied (Table 1). Not only do the studies vary in the small variance prior specification, but they also apply this prior to different item parameters. In addition, some researchers (e.g., Seddig & Leitgöb, 2018)

**Table 1.** Overview of Application Papers Using Bayesian Approximate MI ( $N = 27$ ).

Article	Type of invariance	Small variance prior specification
Braeken and Blömeke (2016)	Multigroup: 12 countries	Factor loadings, residual (co)variances: $\sim N(0, .05)$ ; intercepts: $\sim N(0, .10)$
Bujacz, Vittersø, Huta, and Kaczmarek (2014)	Multigroup: 2 countries	Intercepts, cross-loadings: $\sim N(0, .01)$
Chan, Ivarsson, Stenling, Yang, and Chatzisarantis (2015)	Multigroup and longitudinal: 2 groups and 2 time points <sup>a</sup>	Factor loadings, intercepts, cross-loadings, residual (co)variances: $\sim N(0, .01)$
Chiorri, Day, and Malmberg (2014)	Dyadic: romantic couple <sup>b</sup>	Factor loadings, intercepts: a range from $\sim N(0, .050)$ to $\sim N(0, .500)$
Cieciuch, Davidov, Algesheimer, and Schmidt (2018)	Multigroup and longitudinal: 15 countries and 6 time points <sup>a</sup>	Factor loadings, intercepts: $\sim N(0, .05)$
Cieciuch, Davidov, Schmidt, Algesheimer, and Schwartz (2014)	Multigroup: 8 countries	Factor loadings, intercepts: $\sim N(0, .01)$
Davidov et al. (2015)	Multigroup and longitudinal: 35 countries and 6 time points <sup>a</sup>	Factor loadings, intercepts: $\sim N(0, .05)$
Desa, van de Vijver, Carsten, and Schulz (2019)	Multigroup: 38 countries crossed with 2 genders	Factor loadings, intercepts: $\sim N(0, .02)$
Elsworth, Beauchamp, and Osborne (2016)	Multigroup: 8 community agencies	Cross-loadings, residual (co)variances: $\sim N(0, .02)$
Falkenström, Hatcher, Skjulsvik, Larsson, and Holmqvist (2015)	Longitudinal: 10 therapy sessions	Factor loadings, intercepts: $\sim N(0, .05)$
Fong and Ho (2014)	Multigroup: 2 genders	Cross-loadings, residual (co)variances: $\sim N(0, .01)$
Gucciardi, Zhang, Ponnusamy, Si, and Stenling (2016)	Multigroup: 3 countries	Factor loadings, intercepts: $\sim N(0, .05)$
Hawes et al. (2018)	Multigroup and longitudinal: 2 ethnicities and 4–8 time points	Residual (co)variances: $\sim N(0, .006)$
He and Kubacka (2015)	Multigroup: 38 countries	Factor loadings, intercepts: $\sim N(0, .01)$
Higdon (2015)	Multigroup: 7 countries	Factor loadings, intercepts: $\sim N(0, .10)$ depending on the scale
Jordalen, Lemyre, Solstad, and Ivarsson (2018)	Longitudinal: 3 time points	Factor loadings, intercepts: $\sim N(0, .01)$
Kelcey, McGinn, and Hill (2014)	Multigroup: 39 raters and 2 units	Discrimination and threshold: $\sim IG(.005, .005)$ <sup>c</sup>
Lavenia (2016)	Multigroup: 2 treatment conditions	Not reported
Rammstedt, Danner, Soto, and John (2018)	Multigroup: 2 languages	Factor loadings, intercepts: $\sim N(0, .01)$
Rummel, Steindorf, Marevic, and Danner (2017)	Multigroup: 2 languages	Not reported
Seddig and Leitgöb (2018)	Longitudinal: 4 time points	Factor loadings, intercepts, thresholds: $\sim N(0, .01)$
Sokolov (2018)	Multigroup: 10 cultural zones	Factor loadings, intercepts: $\sim N(0, .01)$
Solstad (2017)	Multigroup: 2 countries	Factor loadings, intercepts: $\sim N(0, .05)$
Von Suchodoletz, Fäsche, and Skuballa (2017)	Multigroup: 2 genders and 3 age groups	Not reported
Williams, Chandola, and Pendleton (2018)	Longitudinal: 5 time points	Intercepts: $\sim N(0, .01)$
Zercher, Schmidt, Cieciuch, and Davidov (2015)	Multigroup and longitudinal: 15 countries and 6 time points <sup>a</sup>	Factor loadings, intercepts: $\sim N(0, .05)$

Note. MI = measurement invariance; CFA = confirmatory factor analysis.

<sup>a</sup>Only multigroup MI is assessed.

<sup>b</sup>Set up as a two-factor model, not a multigroup model.

<sup>c</sup>This article used a graded response model instead of a CFA.

used frequentist modification indices to locate the non-invariant parameters, while others followed the method proposed by Muthén and Asparouhov (2013)—the approach used in the current article. To select the specific small variance prior specification, we will follow the iterative procedure outlined by Asparouhov et al. (2015; Online Supplemental Appendix A). As our example involves a smaller data set, the PPP value and DIC should still reflect changes in the prior specification (Hojtink & van de Schoot, 2017).

## Methods

### Procedure

Data for this example were taken from a larger data set focused on psychosocial factors affecting academic performance of

undergraduate students at a designated Hispanic Serving institution. Specifically, students were asked to complete a self-report survey during the week leading up to a midterm, right after the midterm (before grades were posted) and a week after the midterm (when grades were posted).

### Sample

The total sample consisted of 144 undergraduate students (46 male, 92 female, 1 gender fluid, 5 unknown) enrolled in an Introduction to Psychology course. There was some missing data at each of the measurement occasions: completion rates for the three measurement occasions were 140 (97.2%), 102 (70.8%), and 115 participants (79.9%), respectively. Overall, 127 students participated in at least two measurement occasions (88.2%).

**Table 2.** Descriptive Statistics and Correlation Matrix of Observed Items ( $N = 144$ ).

	y11	y12	y13	y14	y15	y21	y22	y23	y24	y25	y31	y32	y33	y34	y35
y11	1														
y12	0.382	1													
y13	0.355	0.396	1												
y14	0.482	0.411	0.436	1											
y15	0.385	0.433	0.32	0.455	1										
y21	0.454	0.192	0.334	0.439	0.244	1									
y22	0.339	0.753	0.367	0.374	0.305	0.335	1								
y23	0.357	0.289	0.742	0.325	0.352	0.426	0.423	1							
y24	0.488	0.323	0.415	0.635	0.406	0.554	0.417	0.427	1						
y25	0.309	0.306	0.166	0.338	0.665	0.417	0.308	0.342	0.467	1					
y31	0.393	0.193	0.188	0.314	0.247	0.472	0.278	0.304	0.279	0.230	1				
y32	0.206	0.552	0.359	0.331	0.293	0.363	0.622	0.394	0.325	0.312	0.536	1			
y33	0.241	0.335	0.718	0.391	0.425	0.376	0.372	0.780	0.37	0.356	0.389	0.511	1		
y34	0.424	0.364	0.421	0.603	0.487	0.387	0.294	0.351	0.579	0.341	0.371	0.452	0.415	1	
y35	0.368	0.377	0.294	0.499	0.686	0.385	0.344	0.334	0.458	0.745	0.334	0.382	0.419	0.497	1
Mean	1.42	2.28	2.18	2.62	2.02	1.42	2.28	1.84	2.32	1.83	1.41	2.02	1.9	2.49	1.68
Var.	0.69	1.64	1.96	1.99	1.72	0.56	1.36	1.15	1.61	1.15	0.64	1.57	1.42	1.64	0.97

Note. The 15 y-variables in this table reflect the measurement of the 5 items across three time points, such that the first number equals the time point and the second number equals the item number. Items were measured on a 5-point Likert-type response options ranging from *none of the time* (1) to *all of the time* (5). Higher values indicate the presence of more stress-related responses.

## Measure

To illustrate the use of approximate MI in a longitudinal context, we highlight a measure used to assess college student academic stress. Academic stress was hypothesized to vary in the time surrounding an important midterm, but the question remained whether such variation over time was due to actual changes in academic stress, or whether it was due to changes in how students interpreted and answered questions related to academic stress at different stages of their semester.

Specifically, participants in this study completed the Lakaev Academic Stress Response Scale (Lakaev, 2009), a 21-item scale with 5-point Likert-type response options ranging from *none of the time* to *all of the time*. In this study, we focused on the Physiological Stress subscale, which consists of 5 items: (1) *I couldn't breathe*, (2) *I had headaches*, (3) *my hands were sweaty*, (4) *I have had a lot of trouble sleeping*, and (5) *I had difficulty eating*.

One of the current limitations of *Mplus*' implementation of Bayesian approximate MI is that it can only handle continuous or dichotomous items. Based on results from earlier simulation work indicating that categorical variables with five or more answer categories can be analyzed as if they were continuous through ML estimation (Rhemtulla, Brosseau-Liard, & Savalei, 2012), we treated the items as continuous. Item means, variances, and correlations are reported in Table 2.

## Analytical Procedure

For all estimated models, we used a longitudinal CFA (as in Figure 1). For model identification, we estimated all factor loadings but fixed the latent factor means to zero and the latent factor variances to one. We allowed the factors to correlate freely. In addition, we allowed residual covariances for each item with itself across measurement occasions (Liu & West, 2018).

We first examined the data using the classic approach to invariance testing with (full information) ML estimation to illustrate what conclusions could be drawn from such an analysis. We tested

for configural, metric, and scalar invariance, and decided to explore partially invariant models only as needed.

Next, the model was reestimated using a Bayesian approximate MI approach for all factor loadings and intercepts. After a prior variance hyperparameter was selected, we identified non-invariant parameters. Finally, we reestimated the model, with invariant parameters fixed over time and variant parameters estimated freely. Results were compared across different approaches. An example of an *Mplus* input file is included in the Online Supplemental Material.

For all analyses, we used *Mplus* version 8.2 (Muthén & Muthén, 2017). For the ML estimated models, we use the MLR (robust) estimator. For the Bayesian analyses, we used the Gibbs sampler with two MCMC chains. Initially, a minimum of 20,000 iterations was requested, of which half were discarded as burn-in. Convergence was assessed through the Brooks and Gelman convergence diagnostic (1998), potential scale reduction factor (PSRF), which is close to 1 when convergence is reached. *Mplus* ends model estimation when the PSRF drops below a value of  $1 + \epsilon$ , with a default  $\epsilon$  of .05 (multiplied by a constant that takes into account the number of parameters in the model). We used a more stringent convergence criterion of .01 for  $\epsilon$  (i.e., the PSRF was set at 1.01, where values exceeding this were classified as non-converged). We then ran the model with twice the iterations to ensure convergence remained. In addition, parameter trace plots and density plots were inspected to visually confirm convergence. Model fit and parameter estimates reported are based on the results from the longer chain. Other than the small variance difference priors, default prior specifications in *Mplus* were used for all parameters in the model (see Asparouhov & Muthén, 2010, p. 34).

## Analysis of MI

### Classic Stepwise Approach

Results of the classic stepwise approach are reported in Table 3 (the "Partial" row is discussed below). It appears that, through this process, the metric level of invariance was reached. However, once

**Table 3.** Maximum Likelihood (MLR) Longitudinal CFA Model Fit ( $N = 144$ ).

Model	$\chi^2$ (df)	$\Delta\chi^2$ (df) <sup>a</sup>	CFI	$\Delta$ CFI
Configural	99.93 (72)		.960	
Metric	103.14 (80)	3.91 (8) ns	.967	.007
Scalar	122.56 (88)	20.38 (8)*	.950	-.017
Partial <sup>b</sup>	108.84 (84)	5.74 (4) ns	.964	-.003

Note. MI = measurement invariance; CFA = confirmatory factor analysis.

<sup>a</sup>Using the correction formula for MLR estimation reported on the *Mplus* website.

<sup>b</sup>Based on results from Bayesian approximate MI, the intercepts of items y13, y32, y15, y25, and y35 are estimated freely.

\* $p < .01$ .

**Table 4.** Approximate MI Model Comparison ( $N = 144$ ).

Model	Prior	DIC	PPP	95% CI
Approximate	N(0, .001)	4,719.54	.046	-5.73; 86.26
Approximate	N(0, .005)	4,712.91	.092	-16.74; 77.15
Approximate	N(0, .010)	4,712.01	.117	-20.65; 73.86
Approximate	N(0, .050)	4,715.69	.146	-23.22; 71.49
Approximate	N(0, .100)	4,718.03	.143	-23.15; 72.05
Approximate	N(0, .500)	4,720.70	.135	-18.14; 70.57
Configural		4,722.10	.138	-23.00; 73.52
Metric		4,713.73	.118	-17.10; 73.98
Scalar		4,721.32	.042	-5.74; 86.90
Partial <sup>a</sup>		4,705.49	.131	-20.77; 70.74
Metric + approximate <sup>b</sup>	N(0, .050)	4,712.68	.114	-16.95; 73.37

Note. DIC = deviance information criterion; PPP = posterior predictive  $p$  value; 95% CI = 95% credible interval for the difference of observed and replicated  $\chi^2$  values; MI = measurement invariance.

<sup>a</sup>Partial specification based on results from Bayesian approximate MI, the intercepts of items y13, y32, y15, y25, and y35 are estimated freely.

<sup>b</sup>Metric + approximate specification included small variance priors on all intercept differences.

intercepts were constrained to be equal over time, model fit worsened significantly. This implied our measure was only partially invariant over time. Likely due to the relatively small sample size, *Mplus* did not return any suggestions for parameter modifications. This means the location of non-invariance cannot be assessed through existing frequentist methods. Next, we will demonstrate that Bayesian methods can be used to find the location of non-invariance.

### Bayesian Approximate Invariance Testing

First, the width of the difference prior was decided. The results of the iterative procedure (Asparouhov, Muthén, & Morin, 2015) are presented in the top half of Table 4. After balancing the information provided by the DIC and PPP, a prior variance of .05 was selected for future models.

The difference output provided by *Mplus* is reported in Table 5. For each item, this output reports the average posterior estimate (and its standard deviation) of the particular parameter (factor loading or intercept) across all included time points. In addition, it reports the difference of the posterior estimates for that parameter at each time point from the overall average. A \* next to a value indicates that this parameter's posterior estimate falls outside of the 95% credible interval of the average posterior estimate across all time points. A value outside of the 95% credible interval can be

**Table 5.** *Mplus* Difference Output ( $\nu = .05$ ;  $N = 144$ ).

	Average	Std. dev.	Deviations from the Mean		
1	0.540	.056	LAM1_1 .019	LAM2_1 .018	LAM3_1 -.937
2	0.813	.090	LAM1_2 -.006	LAM2_2 -.047	LAM3_2 .053
3	0.756	.089	LAM1_3 .068	LAM2_3 -.060	LAM3_3 -.009
4	0.991	.092	LAM1_4 .028	LAM2_4 .015	LAM3_4 -.044
5	0.691	.088	LAM1_5 .073	LAM2_5 -.028	LAM3_5 -.046
6	1.431	.059	TAU1_1 -.013	TAU2_1 .018	TAU3_1 -.005
7	2.194	.099	TAU1_2 .057	TAU2_2 .070	TAU3_2 -.127 <sup>a</sup>
8	1.996	.099	TAU1_3 .147 <sup>a</sup>	TAU2_3 -.066	TAU3_3 -.080
9	2.496	.101	TAU1_4 .086	TAU2_4 .109	TAU3_4 .022
10	1.845	.090	TAU1_5 .134 <sup>a</sup>	TAU2_5 -.018	TAU3_6 -.115 <sup>a</sup>

Note. LAM1\_1 = factor loading of Item 1 at Time 1; TAU1\_1 = intercept of Item 1 at Time 1.

<sup>a</sup>The parameter's posterior estimate falls outside of the 95% credible interval of the average posterior estimate across all time points.

interpreted as unlikely to occur based on the prior distributions placed on the differences between the parameter estimates across time (i.e., based on how much wiggle room was allowed through the prior distributions).

For this example, the results show that none of the factor loadings (i.e., "LAM") fell outside the 95% credible interval. However, for 3 items, the intercept (i.e., "TAU") fell outside the 95% credible interval on at least one measurement occasion. These intercept estimates are unlikely given our prior expectation of acceptable differences and imply that they deviate meaningfully from parameter estimates at other time points (i.e., measurement non-invariance). Specifically, Item 2 had a lower intercept at the third occasion, Item 3 had a higher intercept at the first occasion, and Item 5 had a higher intercept at the first occasion and a lower intercept at the third occasion—the intercept of Item 5 was non-invariant across all time points.

Now that the location of non-invariance has been identified, it is possible to estimate a partially invariant model using both MLR and Bayesian methods. As can be seen in the last row of Table 3, a model estimated with MLR that freely estimates the parameters identified as non-invariant fit the data well and did not result in a significant decrease in model fit as compared to the metric invariance model, both in terms of the  $\chi^2$  difference test and the change in CFI. In addition, the bottom half of Table 4 indicates that the same model estimated through the Bayesian approach had the lowest DIC when compared to configural, metric, scalar, and a combination of metric and approximate invariance model specifications. Thus, it appears the approximate MI procedure was successful in identifying and modeling the parameters that are non-invariant over time.

### Next Steps

After MI testing is completed, it is common to run additional models that test structural hypotheses of interest. One model which can



**Table 6.** Model Parameter Estimates (SE or SD) of Second-Order LGM Based on Various Measurement Models ( $N = 144$ ).

	MLR		Bayes		Metric + approximate
	Scalar	Partial	Scalar	Partial	
Means					
Intercept	—	—	—	—	—
Slope	−0.22 (0.07)	−0.08 (0.10)	−0.34 (0.11)	−0.09 (0.13)	−0.35 (0.21)
Variances					
Intercept	3.05 (1.50)	4.77 (2.13)	7.12 (2.47)	6.91 (2.47)	6.90 (2.45)
Slope	0.10 (0.12)	—	0.18 (0.22)	0.20 (0.21)	0.21 (0.22)
Intercept with slope	−0.11 (0.23)	—	−0.58 (0.46)	−0.54 (0.46)	−0.55 (0.45)
Model fit					
$\chi^2$ (df)	141.35 (91)	109.83 (89)			
CFI	.927	.970			
DIC			4,704.92	4,699.18	4,701.19
PPP			.117	.171	.190
95% CI			−18.03; 71.17	−23.68; 64.92	−26.62; 64.49

Note. CFI = comparative fit index; DIC = deviance information criterion; PPP = posterior predictive  $p$  value; 95% CI = 95% credible interval for the difference of observed and replicated  $\chi^2$  values; LGM = latent growth model.

be estimated with longitudinal data is a second-order LGM (Ferrer, Balluerka, & Widaman, 2008; Geiser, Keller, & Lockhart, 2013; Grimm & Ram, 2009). This model combines the longitudinal CFA with an LGM to incorporate the measurement model into the estimation of development over time. To illustrate how a different choice of measurement model can affect second-order LGM parameter estimates, a second-order LGM was estimated for five measurement models tested in this article: MLR estimation of scalar or partial invariance, and the implementation of Bayesian methods to examine for scalar, partial, or a combination of metric and approximate invariance. The parameter estimates and model fit results are reported in Table 6.

From Table 6, we can see that the partial invariance model fit the data best across both MLR and Bayesian methods. When MLR estimation is combined with a partial invariance model specification, the slope variance (and thus also the intercept-slope covariance) cannot be estimated due to a non-positive definite covariance matrix. Non-positive definite covariance matrices do not occur when models are estimated using Bayesian methods. This is because Bayesian inference is estimated through an iterative MCMC procedure, in combination with a set of prior distributions that prevent the variance estimates from becoming negative (i.e., through use of the inverse Gamma and inverse Wishart distributions, which result in a zero probability on any values  $< 0$ ). In addition, the partial invariance second-order LGM estimated through Bayesian methods provides more information about the distribution of the model parameters of interest.

## Discussion

The aim of the current article was to illustrate Bayesian approximate MI testing for longitudinal data with a small sample size. We assessed longitudinal MI of physiological academic stress of undergraduate students at three measurement occasions surrounding an important midterm. The results indicate that using the classic approach to invariance testing is not always feasible with small sample sizes, and Bayesian approximate MI testing offers a flexible alternative that is feasible.

## Overview of MI Techniques

In order to help place some of the concepts from this illustration into further context, and to aid in solidifying the different approaches, we are including an overview of MI techniques. This overview highlights the main MI techniques discussed, what situation each technique works well with, as well as drawbacks for each. This section should be viewed as a simple guide to these techniques, and when each is (or is not) appropriate to use.

Classic exact approach to MI, including testing for full and partial MI:

- Works well for: Limited number of groups or measurement occasions and when the sample size per group or measurement occasion is large.
- Drawbacks: Expecting exact invariance across groups or time is not always realistic.

Bayesian approximate MI:

- Works well for: Large number of groups or measurement occasions and small sample sizes. Allows the researcher to control an acceptable amount of non-invariance through difference priors.
- Drawbacks: Leads to bias in latent factor means and variances when non-invariance is systematic (e.g., factor loadings for items systematically decrease as time passes) and substantive in nature. In addition, it is unclear what the best method is for specifying the difference prior.

Alignment method:

- Works well for: Large number of groups, and in situations where most items are approximately invariant, with only some items showing large deviations from invariance. It only requires a configural model and uses a component loss function to minimize the amount of non-invariance across groups in the factor loadings and intercepts.
- Drawbacks: Highly data-dependent method, not implemented for longitudinal invariance in any software package.



## Limitations of the Current Illustration

One of the main limitations of the current illustration is that we opted to narrow the discussion down to a single software program, namely, *Mplus*. We used this program for two main reasons. First, the *Mplus* software program is relatively user-friendly, and it has already been heavily adopted by empirical researchers working in the frequentist setting. Second, a recent systematic review of the empirical Bayesian literature published within the Psychological Sciences indicated that *Mplus* is the most often implemented software program for applied Bayesian work since 2013 (van de Schoot et al., 2017). These reasons led us to using the *Mplus* program here.

One limitation to this software selection is that *Mplus* does not have certain features that other R-based programs can offer. In particular, *Mplus* can only handle continuous or binary items when approximate MI is assessed, and other item types cannot yet be implemented in the program. In addition, there are several superior indices that are not included in the software program, including the PPPP value, WAIC, and LOO, which have been recommended over the DIC and PPP value for model comparison and fit assessment. Overall, although *Mplus* is a valuable program for implementing Bayesian approximate MI, the model comparison and fit measures offered in the program are not without their limitations.

## Future Areas of Research

Bayesian approximate MI testing is just one new tool in the general MI testing toolbox. For MI testing across groups, the alignment method can be used in situations where approximate MI specification would lead to biased latent variable means (Asparouhov & Muthén, 2014; Flake & McCoach, 2017). This method is especially effective when measurement non-invariance is systematic and substantive. The alignment method has not yet been implemented for use with longitudinal data and could prove to be a fruitful area for future research. Likewise, Bayesian practical invariance has recently been introduced as an alternative to frequentist and Bayesian approximate approaches to invariance testing (Shi et al., 2018). In this approach, researchers compare obtained uncertainty intervals (i.e., 95% highest posterior density) of the size of the invariance to a previously defined region of practical equivalence.

In addition, research on the performance of Bayesian approximate MI with multiple groups implies that the method is especially suitable when there are many groups and thus many small deviations from invariance (e.g., Kim, Cao, Wang, & Nguyen, 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013). It seems reasonable to generalize this conclusion to longitudinal data and assume it is especially suited for data with many time points. However, the existing simulation literature on the performance of Bayesian approximate MI with longitudinal data is still limited. Moreover, studies that have focused on approximate MI over time have at most included five measurement occasions (Liang, Yang, & Huang, 2018; Muthén & Asparouhov, 2013) and 10 items. Findings imply that Bayesian approximate MI can be advantageous even with a limited number of time points, so long as the overall pattern of non-invariance across items and time points is non-systematic. In addition, results of our own simulation study on this topic (Winter & Depaoli, in press) indicate that this approach is especially suitable for studies with a smaller sample size (e.g.,  $n = 150$ ). However, this is an area that is in need of further exploration in order to fully understand the benefits of the technique.


Finally, with the emergence of ecological momentary assessment (EMA), future research might focus on the particular utility of Bayesian approximate MI for data with many time points. As of now, there is limited understanding with respect to how the approximate MI approach may benefit EMA data. Expanding the topics covered in this illustration, and exploring the impact with many more time points, would be a beneficial next step toward the potential application of Bayesian approximate MI to EMAs.

In conclusion, this article illustrated the use and advantages of Bayesian approximate MI testing with longitudinal data and a small sample size. While this method does not solve issues of bad measurement, it does provide researchers with more flexibility to explore and assess MI over time, where traditional methods fail. Finally, given the potential benefits of the approach, Bayesian approximate MI lends itself nicely to a rich future of methodological research.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Sonja D. Winter  <https://orcid.org/0000-0002-2203-002X>

## Supplemental material

Supplemental material for this article is available online.

## References

- Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation*. Retrieved from <https://www.statmodel.com/download/Bayes3.pdf>
- Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 495–508. doi:10.1080/10705511.2014.919210
- Asparouhov, T., & Muthén, B. (2017). Prior-posterior predictive p-values. *Mplus Web Notes: No. 22*. Retrieved from <https://www.statmodel.com/download/PPPP.pdf>
- Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeier et al. *Journal of Management*, 41, 1561–1577. doi:10.1177/0149206315591075
- Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.
- Bentler, P. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bentler, P., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in analysis of covariance structures. *Psychological Bulletin*, 88, 558–606. doi:10.1037/0033-2909.88.3.588
- Braeken, J., & Blömeke, S. (2016). Comparing future teachers' beliefs across countries: Approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning. *Assessment & Evaluation in Higher Education*, 41, 733–749. doi:10.1080/02602938.2016.1161005
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.
- Bujacz, A., Vittersø, J., Huta, V., & Kaczmarek, L. D. (2014). Measuring hedonia and eudaimonia as motives for activities: Cross-national investigation through traditional and Bayesian

- structural equation modeling. *Frontiers in Psychology*, 5, 1–10. doi:10.3389/fpsyg.2014.00984
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456–466. doi:10.1037/0033-2909.105.3.456
- Chan, D. K. C., Ivarsson, A., Stenling, A., Yang, X. S., & Chatzisarantis, N. L. D. (2015). Response-order effects in survey methods: A randomized controlled crossover study in the context of sport injury prevention. *Journal of Sport and Exercise Psychology*, 37, 666–673. doi:10.1123/jsep.2015-0045
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 464–504. doi:10.1080/10705510701301834
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9, 233–255. doi:10.1207/S15328007SEM0902\_5
- Chiorri, C., Day, T., & Malmberg, L. E. (2014). An approximate measurement invariance approach to within-couple relationship quality. *Frontiers in Psychology*, 5. doi:10.3389/fpsyg.2014.00983
- Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2018). Testing for approximate measurement invariance of human values in the European Social Survey. *Sociological Methods & Research*, 47, 665–686. doi:10.1177/0049124117701478
- 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, 982. doi:10.3389/fpsyg.2014.00982
- Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M. (2015). The comparability of measurements of attitudes toward immigration in the European Social Survey: Exact versus approximate measurement equivalence. *Public Opinion Quarterly*, 79, 244–266. doi:10.1093/poq/nfv008
- Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, 22, 240–261. doi:10.1037/met0000065
- Desa, D., van de Vijver, F., Carsten, R., & Schulz, W. (2019). Measurement invariance in international large-scale assessments: Integrating theory and method. In T. P. Johnson, B. E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology* (pp. 881–910). New York, NY: Wiley.
- Edossa, A. K., Schroeders, U., Weinert, S., & Artelt, C. (2018). The development of emotional and behavioral self-regulation and their effects on academic achievement in childhood. *International Journal of Behavioral Development*, 42, 192–202. doi:10.1177/0165025416687412
- Elsworth, G. R., Beauchamp, A., & Osborne, R. H. (2016). Measuring health literacy in community agencies: A Bayesian study of the factor structure and measurement invariance of the health literacy questionnaire (HLQ). *BMC Health Services Research*, 16, 508. doi:10.1186/s12913-016-1754-2
- Falkenström, F., Hatcher, R. L., Skjulsvik, T., Larsson, M. H., & Holmqvist, R. (2015). Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy. *Psychological Assessment*, 27, 169–183. doi:10.1037/pas0000038
- Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, 4, 22–36. doi:10.1027/1614-2241.4.1.22
- Flake, J. K., & McCoach, D. B. (2017). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 25, 56–70. doi:10.1080/10705511.2017.1374187
- Fong, T. C. T., & Ho, R. T. H. (2014). Testing gender invariance of the Hospital Anxiety and Depression Scale using the classical approach and Bayesian approach. *Quality of Life Research*, 23, 1421–1426. doi:10.1007/s11136-013-0594-3
- Garnier-Villareal, M. (2019). Specification of priors in blavaan. [Online forum comment]. Retrieved May 14, 2019, from <https://groups.google.com/forum/#!msg/blavaan/clwG87spt2Q/Izwo84MBAwAJ>
- Garnier-Villareal, M., & Jorgensen, T. D. (2019). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*. Epub ahead of print 10 June 2019. doi:10.1037/met0000224
- Geiser, C., Keller, B., & Lockhart, G. (2013). First versus second order latent growth curve models: Some insights from latent state-trait theory. *Structural Equation Modeling: A Multidisciplinary Journal*, 20, 479–503. doi:10.1080/10705511.2013.797832
- Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74, 153–160. doi:10.1080/01621459.1979.10481632
- Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 501–514.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York, NY: Chapman & Hall/CRC.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016. doi:10.1007/s11222-013-9416-2
- Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–760.
- Grimm, K. J., & Ram, N. (2009). A second-order growth mixture model for developmental research. *Research in Human Development*, 6, 121–143. doi:10.1080/15427600902911221
- Gucciardi, D. F., Zhang, C. Q., Ponnusamy, V., Si, G., & Stenling, A. (2016). Cross-cultural invariance of the mental toughness inventory among Australian, Chinese, and Malaysian athletes: A Bayesian estimation approach. *Journal of Sport & Exercise Psychology*, 38, 187–202. doi:10.1123/jsep.2015-0320
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97.
- Hawes, S. W., Byrd, A. L., Kelley, S. E., Gonzalez, R., Edens, J. F., & Pardini, D. A. (2018). Psychopathic features across development: Assessing longitudinal invariance among Caucasian and African American youths. *Journal of Research in Personality*, 73, 180–188. doi:10.1016/j.jrp.2018.02.003
- He, J., & Kubacka, K. (2015). Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013. OECD Education Working Papers, No. 124, OECD Publishing, Paris, <https://doi.org/10.1787/5jrp6fwtmh2-en>.

- Higdon, J. D. (2015). *Measuring and modeling intercultural attitudes among adolescents across Europe: A multi-level, multiple-group analysis examining student attitudes, intergroup contact, and school climate*. Doctoral dissertation, Harvard Graduate School of Education.
- Hojtink, H., & van de Schoot, R. (2017). Testing small variance priors using prior-posterior predictive  $p$  values. *Psychological Methods*, 23, 561–569. doi:10.1037/met0000131
- Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement*, 78, 537–568. doi:10.1177/0013164417709314
- Jordalen, G., Lemyre, P., Solstad, B. E., & Ivarsson, A. (2018). The role of self-control and motivation on exhaustion in youth athletes: A longitudinal perspective. *Frontiers in Psychology*, 9, 1–13. doi:10.3389/fpsyg.2018.02449
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426. doi:10.1007/BF02291366
- Jorgensen, T. D., Garnier-Villareal, M., Pomprasernan, S., & Lee, J. (2019). Small-variance priors can prevent detecting important misspecifications in Bayesian confirmatory factor analysis. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology: IMPS 2017. Springer Proceedings in Mathematics & Statistics* (pp. 255–263). Cham, Switzerland: Springer. doi:10.1007/978-3-030-01310-3\_23
- Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York, NY: Guilford Press.
- Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology*, 5, 1–13. doi:10.3389/fpsyg.2014.01469
- Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 24, 524–544. doi:10.1080/10705511.2017.1304822
- Kruschke, J. K. (2013). Bayesian estimation supersedes the  $t$  test. *Journal of Experimental Psychology: General*, 142, 573–603. doi:10.1037/a0029146
- Lakaev, N. (2009). Validation of an Australian academic stress questionnaire. *Australian Journal of Guidance and Counselling*, 19, 56–70. doi:10.1375/ajgc.19.1.56
- Lavenia, M. (2016). *Mathematics formative assessment system: Testing the theory of action based on the results of a randomized field trial*. Tallahassee, FL: Florida State University.
- Liang, X., Yang, Y., & Huang, J. (2018). Evaluation of structural relationships in autoregressive cross-lagged models under longitudinal approximate invariance: A Bayesian analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 25, 558–572. doi:10.1080/10705511.2017.1410706
- Liu, Y., & West, S. G. (2018). Longitudinal measurement non-invariance with ordered-categorical indicators: How are the parameters in second-order latent linear growth models affected? *Structural Equation Modeling: A Multidisciplinary Journal*, 25, 762–777. doi:10.1080/10705511.2017.1419353
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. J. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. New York, NY: Chapman and Hall/CRC.
- Lynch, S. M., & Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods & Research*, 32, 301–335. doi:10.1177/0049124103257303
- Maccallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*, 17, 340–345. doi:10.1037/a0027131
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, 93, 568–592. doi:10.5465/AMBPP.2006.27182124
- Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, 5, 279–300. doi:10.1207/s15327574ijt0503
- Meng, X. L. (1994). Posterior predictive  $p$  values. *The Annals of Statistics*, 22, 1142–1160.
- Meredith, W. (1964a). Notes on factorial invariance. *Psychometrika*, 29, 177–185.
- Meredith, W. (1964b). Rotation to achieve factorial invariance. *Psychometrika*, 29, 187–206.
- Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*, 58, 525–543. doi:10.1007/Bf02294825
- Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85, 1–30. doi:10.18637/jss.v085.i04
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fact computing machines. *Journal of Chemical Physics*, 21, 1087–1092.
- Meuleman, B. (2012). When are item intercept differences substantively relevant in measurement invariance testing? In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences: Festschrift for Peter Schmidt* (pp. 97–104). Heidelberg: Springer VS. doi:10.1007/978-3-531-18898-0
- Muthén, B. O., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus Web Notes: No. 17*. Retrieved from <http://www.statmodel.com/examples/webnotes/webnote17.pdf>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 26, 724–744. doi:10.1080/10705511.2018.1561293
- Price, R. (1765). A demonstration of the second rule in the essay toward the solution of a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 54, 296–325.
- Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2018). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*. Epub ahead of print 3 August 2018. doi:10.1027/1015-5759/a000481
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354–373. doi:10.1037/a0029315.supp
- Robert, C., & Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26, 102–115. doi:10.1214/10-STS351
- Rummel, J., Steindorf, L., Marevic, I., & Danner, D. (2017). A validation study of the German complex-span tasks and some general considerations on task translation procedures in cognitive psychology.

- European Journal of Psychological Assessment*, 1–12. Epub ahead of print 15 December 2017. doi:10.1027/1015-5759/a000444
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74, 31–57. doi:10.1177/0013164413498257
- Scheines, R., Hoijsink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 64, 37–52.
- Seddig, D., & Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: Concept and application with panel data. *Survey Research Methods*, 12, 29–41. doi:10.18148/srm/2018.v12i1.7210
- Shi, D., Song, H., Distefano, C., Maydeu-Olivares, A., McDaniel, H. L., & Jiang, Z. (2018). Evaluating factorial invariance: An interval estimation approach using Bayesian structural equation modeling. *Multivariate Behavioral Research*, 54, 224–245. doi:10.1080/00273171.2018.1514484
- Sokolov, B. (2018). The index of emancipative values: Measurement model misspecifications. *American Political Science Review*, 112, 395–408. doi:10.1017/S0003055417000624
- Solstad, B. E. (2017). *Towards a better understanding of the dynamics of sports coaching at the youth level: The coach's perspective*. Oslo: Norwegian School of Sport Sciences.
- 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, 583–639.
- Stan Development Team (2018). RStan: The R interface to Stan. R package version 2.18.2. Retrieved from <http://mc-stan.org/>
- Svetina, D., Rutkowski, L., & Rutkowski, D. (2019). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal*. Epub ahead of print 29 April 2019. doi:10.1080/10705511.2019.1602776
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70. doi:10.1177/109442810031002
- van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist*, 16, 75–84.
- van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4, 1–15. doi:10.3389/fpsyg.2013.00770
- van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22, 217–239.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413–1432. doi:10.1007/s11222-016-9696-4
- Von Suchodoletz, A., Fäsche, A., & Skuballa, I. T. (2017). The role of attention shifting in orthographic competencies: Cross-sectional findings from 1st, 3rd, and 8th grade students. *Frontiers in Psychology*, 8, 1665. doi:10.3389/fpsyg.2017.01665
- 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.
- Williams, B. D., Chandola, T., & Pendleton, N. (2018). An application of Bayesian measurement invariance to modelling cognition over time in the English Longitudinal Study of Ageing. *International Journal of Methods in Psychiatric Research*, 27, e1749. doi:10.1002/mpr.1749
- Winter, S. D., & Depaoli, S. (in press). *Measurement Invariance of Continuous and Dichotomous Indicators in Second-Order Latent Growth Models: Assessing Frequentist and Bayesian Approaches*.
- Yuan, K. H., & Chan, W. (2016). Measurement invariance via multi-group SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, 21, 405. doi:10.1037/met0000080
- Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact vs. approximate measurement invariance. *Frontiers in Psychology*, 6, 1–11. doi:10.3389/fpsyg.2015.00733