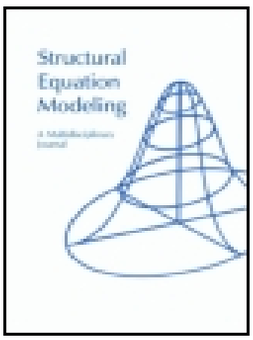

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

On Using Bayesian Methods to Address Small Sample Problems

Daniel McNeish

To cite this article: Daniel McNeish (2016): On Using Bayesian Methods to Address Small Sample Problems, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2016.1186549

To link to this article: http://dx.doi.org/10.1080/10705511.2016.1186549

Published online: 14 Jun 2016.

Structural Equation Modeling: A Multidisciplinary Journal, 00: 1–24, 2016

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2016.1186549

Routledge
Taylor & Francis Group

TEACHER'S CORNER

On Using Bayesian Methods to Address Small Sample Problems

Daniel McNeish
Utrecht University

As Bayesian methods continue to grow in accessibility and popularity, more empirical studies are turning to Bayesian methods to model small sample data. Bayesian methods do not rely on asympotics, a property that can be a hindrance when employing frequentist methods in small sample contexts. Although Bayesian methods are better equipped to model data with small sample sizes, estimates are highly sensitive to the specification of the prior distribution. If this aspect is not heeded, Bayesian estimates can actually be worse than frequentist methods, especially if frequentist small sample corrections are utilized. We show with illustrative simulations and applied examples that relying on software defaults or diffuse priors with small samples can yield more biased estimates than frequentist methods. We discuss conditions that need to be met if researchers want to responsibly harness the advantages that Bayesian methods offer for small sample problems as well as leading small sample frequentist methods.

Keywords: Bayes, prior distribution, small sample

In the last decade or so, Bayesian methods have become vastly more popular in nearly all scientific fields (van de Schoot, 2016). In fact, based on a comprehensive review of Bayesian studies over the last 15 years, van de Schoot (2016) noted that the number of empirical papers in psychology using Bayesian methods increased nearly fivefold between 2010 and 2015. This rapid increase is partially attributable to the decrease in the cost of computational resources needed to estimate the increasingly complex models applied researchers are fitting to their data as well as increased accessibility to Bayesian software (Dunson, 2001), M*plus* in particular, whose Bayesian module was available beginning in Version 6, which, not coincidently, was introduced in 2010.

In addition to greater ease of implementation, several recent methodological papers and Monte Carlo simulation studies have noted the potential advantages of Bayesian methods over frequentist maximum likelihood (ML) methods with small samples (e.g., Baldwin & Fellingham, 2013; Depaoli, 2013; Depaoli & van de Schoot, 2015; Dunson, 2000; Gelman, 2006; Hox, van de Schoot, & Matthijsse, 2012; Kadane, 2015; Lambert, Sutton, Burton, Abrams, & Jones, 2005; Lee & Song, 2004; McNeish, 2016; McNeish & Stapleton, 2016; Muthén & Asparouhov, 2012; Price, 2012; Scheines, Hoijtink, & Boomsma, 1999; Soares, Gonçalves, & Gamerman, 2009; Stegmueller, 2013; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015). As researchers are certainly aware, sample sizes in a vast array of behavioral science fields are frequently quite small. For instance, two meta-analyses by Roberts and colleagues found that about 33% of growth models investigating personality traits over time had fewer than 100 people, a review by Russell (2002) found that about 40% of exploratory factor analysis studies in psychology had samples less than 100, 18% of structural equation modeling (SEM) applications in psychology had samples less than 100 in a review by MacCallum and Austin (2000), 50% of meta-analyses in education had fewer than 40 studies (Ahn, Ames, & Myers, 2012), and the average number of clusters in registered primary care randomized trials was 29 (Eldridge, Ashby, Feder, Rudnicka, & Ukoumunne, 2004). The issue of small sample size is highly

Correspondence should be addressed to Daniel McNeish, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands.
E-mail: d.m.n.mcneish@uu.nl

relevant and the promise of Bayesian methods to handle these situations in a more-or-less straightforward manner without small sample corrections or adjustments is undoubtedly alluring. Using this justification, an increasing number of empirical studies have noted that Bayesian methods are explicitly selected over frequentist methods to better accommodate reduced sample sizes. Recent occurrences of this rationale from a wide variety of fields can be found quite readily. Here are five such examples from diverse fields such as abnormal psychology, exercise science, cross-cultural studies, and prevention science.

- "A Bayesian estimator was used to account for the small sample size … providing a more trustworthy result than a traditional maximum likelihood estimator" (Wanless, Rimm-Kaufman, Abry, Larsen, & Patton, 2015, p. 1111).
- "Several reasons favored the utilization of the Bayesian estimator rather than the more traditional maximum likelihood estimator. First, accurate Bayesian estimation can be obtained with very small samples" (Doron & Gaudreau, 2014, p. 7).
- "This procedure relies on Bayesian estimation of the overall effect size, which has been shown to be more appropriate in meta-analyses with a small number of studies" (Kliem, Kröger, & Kosfelder, 2010, p. 938).
- "When comparing ML-based SEM with BSEM, some advantages of the BSEM approach have been suggested … better small-sample performance can be obtained, as the Gibbs sampler has been found reliable for all *N*s" (Stenling, Ivarsson, Johnson, & Lindwall, 2015, p. 412).
- "Compared to traditional CFA models, the sample size requirement for a Bayesian CFA is generally more relaxed" (McNab et al., 2010, p. 1274).

However, as we discuss in this article, although Bayesian methods are better equipped to handle small sample situations, simply switching to a Bayesian framework alone does not necessarily mitigate the small sample problem in and of itself. That is, just like in the frequentist framework, when modeling small sample data, special considerations must be taken to ensure that the resulting inferences are as trustworthy as possible. In this article, we discuss circumstances in which Bayesian analyses can indeed outperform frequentist analyses as well as situations in which switching to Bayesian methods without due diligence can make inferences from small samples worse than those obtained with ML. We provide demonstrative simulations and applied examples of a growth model, multilevel model, and meta-analysis to demonstrate how simply changing TYPE = ML to TYPE = BAYES in M*plus*, for instance, is not sufficient to capitalize on these advantages and that Bayesian methods are not an instant cure for small sample sizes.

BAYESIAN METHODS WITH SMALL SAMPLES

As an overview, there are many philosophical differences between how data and parameters are considered between the Bayesian and frequentist frameworks, namely that Bayesian methods consider that all parameters are random (and therefore have distributions) and that the data are fixed. Conversely, frequentists consider all parameters to be fixed and that the data are random. As a result, each framework approaches statistical inferences from very different perspectives. Frequentists aim to answer this question: Given a set of parameter values, how likely is it to obtain the values seen in the data? On the other hand, Bayesians approach statistical modeling from the viewpoint that, given that the data are collected and the experiment will not be repeated, what are the values for the parameters? As advantages of the Bayesian approach, researchers can include prior information into the model through prior distributions (discussed in the next section), which can help the accuracy of predictions; interval estimates have a more intuitive interpretation because the data are viewed as fixed rather than random (Jackman, 2009); hierarchical models can be easily extended to data with multilevel levels, missing data, or for multiple comparisons (Gelman, Hill & Yajima, 2012); and models that are difficult or inestimable with frequentist methods can be fit straightforwardly with Bayesian methods (Levy, 2009; Muthén & Asparouhov, 2012). As a disadvantage, all parameters are random, meaning that their interpretation is probabilistic. This can cause difficulties with model comparison, model selection, and hypothesis testing that has come to be the norm in statistical analysis due to the relative ease with which these aspects can be carried out in a frequentist framework (Jordan, 2011; Levy & Choi, 2013).

Although some statisticians staunchly pledge allegiance to one faction or the other based on the reasons just outlined, many applied researchers are open to either philosophical orientation, provided that it is well-suited for their data or analytical needs and that the resulting conclusions are trustworthy for the purpose of testing their theory. With small sample sizes, the goal of obtaining trustworthy estimates might come into question with ML estimation (the primary frequentist estimation method). ML is known to have desirable properties such as consistency (the parameter estimates are unbiased on expectation as sample size approaches infinity) and asymptotic normality (which is the foundation for frequentist confidence interval calculations). However, notice that both of these properties require large samples to take effect. As a result, ML estimates in smaller samples can be quite poor, which has been noted on several occasions in the literature (e.g., Browne & Draper, 2006; Lee & Song, 2004; McNeish & Stapleton, 2014).

Nonetheless, this problem is well-known in the statistical literature and several remedies have been proposed. For

instance, methods like restricted maximum likelihood (REML) estimation (Harville, 1977), the Kenward–Roger standard error and degree of freedom correction (Kenward & Roger, 1997, 2009), or Skene–Kenward corrections (Skene & Kenward, 2010a, 2010b) that have been found to perform well in the case of small sample sizes with models for clustered data such as growth models or multilevel models (continuous outcomes are required for most of these methods and remedies for discrete outcomes are less apparent). Methods for obtaining better small sample data–model fit indexes for structural equation models are also abundant including algebraic corrections by Bartlett (1950), Swain (1975), and Yuan (2005), or more statistical rigorous corrections such as those proposed in a series of papers by Yuan and Bentler (Bentler & Yuan, 1999; Yuan & Bentler, 1997, 1999), among others.

Due to their differential theoretical underpinnings, Bayesian methods do not rely on large samples. Recall that Bayesian methods consider the data to be fixed while the parameters are random. With sampling-based Bayesian methods such as Markov chain Monte Carlo (MCMC), which is becoming increasingly synonymous with Bayesian estimation in empirical studies, this means that the quality of inference is controlled not by sample size approaching infinity, but rather by the number of samples taken approaching infinity (Kruschke, 2010; Lee & Song, 2004). This has led to recommendations that MCMC methods are better equipped to model data with small samples because MCMC does not rely on asymptotics in the same way that frequentist methods do (Depaoli & van de Schoot, 2015; Dunson, 2001; Kadane, 2015; Lee & Song, 2004; Muthén & Asparouhov, 2012; van de Schoot et al., 2014).

For instance, whereas frequentist confidence intervals rely on asymptotic normality, MCMC intervals are taken from percentiles of the posterior distribution for each parameter (sometimes referred to as *exact inference* in the statistical literature). This obviates the need for degree of freedom adjustments (e.g., Kenward–Roger) or finite sample corrections (e.g., REML) that are necessary in the frequentist framework: Assumptions about the sampling distribution of parameters need not be made because the MCMC framework provides a distribution of each parameter via resampling algorithms. The ability of such a distribution also eliminates the need for complex mathematical operations (e.g., computation of first and second derivatives of the likelihood) that are required to compute standard errors with ML, which is a prime culprit of convergence issues with smaller samples (Levy & Choi, 2013).

In addition, previous knowledge about the phenomenon under investigation can be incorporated in the analysis via the prior distributions with so-called informative or subjective priors. This allows an MCMC analysis to base results on more information than what is strictly provided in the data, which is especially helpful at smaller sample sizes where each additional piece of information has a more pronounced impact. More generally, MCMC can also prevent inadmissible estimates (e.g., negative variances) through selection of a prior distribution that is restricted to have nonnegative support. With smaller samples, out of bounds solutions are a frequent cause for nonconvergence with ML (Depaoli & Clifton, 2015; Schoeneberger, 2015).

Although it is certainly true that MCMC is better equipped theoretically to handle small sample problems, being theoretically better equipped to handle these situations is not the same as being uniformly more appropriate. To be more specific, Muthén and Asparouhov (2012) stated that "Point 2 [MCMC is advantageous with small sample sizes] is illustrated by better Bayesian small-sample performance for factor analyses prone to Heywood cases and better performance when a small number of clusters are analyzed in multilevel models. This, however, requires a judicious choice of prior" (p. 314). Kadane (2015) noted that "a Bayesian analysis with a proper, subjective prior can be done comfortably at any sample size" (p. 1018) and van de Schoot et al. (2014) echoed this sentiment: "It should be noted that the smaller the sample size, the bigger the influence of the prior specification and the more can be gained from specifying subjective priors" (p. 856). The crucial (and often overlooked) portion of these statements (and related statements found in several other papers) concerns the choice of the prior distribution: MCMC methods have no issue with small samples, provided that strong priors are provided. Herein lies the potential danger of blindly pursuing MCMC with small samples.

PRIOR DISTRIBUTIONS WITH SMALL SAMPLES

As a quick review, in Bayesian statistics, the posterior distribution (which is summarized to yield the Bayesian equivalents of frequentist point estimates and standard errors) is a combination of the prior distribution (determined by the researcher) and the likelihood (determined by the data). The contribution of these two quantities to the posterior distribution is not equal. When the data contain many observations, the likelihood is given much more relative weight in calculating the posterior distribution than when the data have fewer observations. By similar logic, with small sample data, the prior distribution is given more weight in calculating the posterior distribution, relative to whether the data came from a larger sample (van de Schoot et al., 2014).

Although this seems trivial, it can have enormous implications with small samples, as the posterior distribution is highly reliant on how the prior was specified. As Gelman (2006) noted and as echoed in simulation work by McNeish (2016), with small samples, the idea of noninformative priors is more myth than reality: when the information contained in the likelihood is relatively small due to a limited sample size, the prior will necessarily play a key role in the posterior distribution. Moreover, the diffuse prior distributions that are used when researchers want to "let their data do the talking" can be highly problematic with small samples (Gelman, 2006). To explain, with small

samples, the prior is given more relative weight. Diffuse priors have very wide support, meaning that the algorithm can sample an extreme value. Unlike with large sample sizes, the likelihood is not able to negate the effect of sampling an extreme value and an extreme value is given more relative weight in the calculation of the posterior distribution with small samples than with larger samples. In practice, variance parameters and posterior standard deviations can be inflated, a consequence that can be dire in small samples because power will be limited from the onset and inefficient variance estimates will exacerbate the problem by making estimates appear less precise than they might be in actuality.

This is not to say that MCMC methods consistently perform poorly with small samples. Several methodological studies have shown that MCMC methods can indeed outperform frequentist methods in small sample contexts as theory would prescribe (McNeish, 2016; Moore, Reise, Depaoli, & Haviland, 2015; van de Schoot et al., 2015; Zhang, Hamagami, Wang, Nesselroade, & Grimm, 2007). However, these studies obtained improved performance uniformly by specifying informative priors based on expert opinions or previous studies. The problem lies in that applied researchers typically do not consider their prior distributions as carefully as these methodological studies recommend, if at all. For instance, in a review of 99 empirical psychology studies using Bayesian methods by van de Schoot, Ryan, Winter, Zondervan-Zwijenburg, and Depaoli (under review), over half of the studies (56%) failed to report information about the prior distributions and presumably used software defaults. M*plus* documentation also does not alert applied users to this potential pitfall and also seems to imply that the default priors are suitable in a variety of situations by stating, "in Mplus, simple analysis specifications with convenient defaults allow easy access to a rich set of analysis possibilities" (Muthén, 2010, pp. 3–4). If researchers are using the increasingly popular BAYES module in M*plus*, the default prior distributions are intended to apply to a wide variety of modeling contexts but are quite poor choices with small samples (McNeish, 2016).

The broader implication is that a disconnect exists between what researchers think MCMC is accomplishing compared to what is actually being done. There is a growing illusion that MCMC is impervious to small samples and that simply changing the estimation method mitigates the small sample problem straightaway. If wide or diffuse priors that are the default in software programs like M*plus* are employed, then not only are the advantages of MCMC methods not being tapped to address the small sample problem, but MCMC can actually exhibit worse properties than frequentist methods, especially when frequentist small sample corrections are considered. That is, the commonly held adage that using diffuse priors cannot hurt and that results from diffuse priors will at worst be equal to ML is unfounded when sample sizes are small; the idea of a noninformative prior becomes a myth when sample size is small (Gelman, 2006).

For the remainder of this article, we demonstrate that switching to an MCMC framework based on potential advantages with sample sizes without careful consideration of the prior distribution will not address the small sample issue and that ML with small sample alterations typically produce estimates with quality that can equal or often surpass MCMC methods that do not carefully consider the prior distributions. Examples are provided for three common modeling situations where researchers routinely encounter small sample sizes: growth modeling, multilevel modeling, and meta-analysis. Software code for running these models in M*plus* (or SAS, where appropriate) is provided in the Appendix.

THE CONTEXT OF THIS STUDY

Extant studies in the methodological literature have addressed issues associated with undisciplined selection of prior distributions including small sample contexts specifically. Despite the existence of these studies, the current literature on small sample Bayesian methods is rather diffuse.

Many important findings on small sample Bayesian issues stem from secondary questions embedded within studies with broader aims. As recent examples, Depaoli and Clifton (2015) aptly demonstrated that diffuse priors show quite poor performance with smaller samples for multilevel structural equation models, Depaoli (2013) showed that frequentist methods outperform MCMC with diffuse priors in growth mixture models with smaller samples, and McNeish and Stapleton (2016) compared several different methods for modeling small sample clustered data and reported that some priors were ineffective. These studies include highly relevant findings on small sample Bayesian issues but each has an entirely different primary goal: how to implement Bayesian methods for multilevel structural equation models, latent class recovery in growth mixture models, or comparing the performance of several different methods for clustered data. Moreover, substantive researchers searching the literature for guidance on small sample issues might also not even come across the information contained in these studies if they are working with a model that is not the main focus of these studies. For instance, a researcher with small sample growth data might not locate Depaoli and Clifton (2015) or, if they do locate this study, they might judge it to be unrelated because the title ("A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes") does not readily reveal that it contains relevant information to small sample problems generally. Furthermore, these studies are intended to appeal to methodological researchers and delve into many specifics about

the particular model under investigation and also contain simulation studies with several manipulated conditions. The broad take-home message regarding issues with either small samples or MCMC can easily be lost in the shuffle because these conditions are often moderators among many other conditions and dedicated conclusions pertaining to small samples and MCMC are either brief or implied.

Relatedly, Bayesian methods and small sample sizes are relevant in a wide variety of disciplines. As a result, the relevant methodological literature is fairly scattered and some articles might appear in journals outside of the field within which substantive researchers work. Although numerous previous studies appear in broad methodological journals (i.e., *Structural Equation Modeling, Psychological Methods, Multivariate Behavioral Research*), there are several that appear in discipline-specific journals or specialist methodological journals that might be unknown to substantive researchers working in different areas. For instance, highly relevant and informative small sample Bayesian studies can be found in journals such as the *European Journal of Psychotraumatology, Survey Research Methods, Statistics in Medicine, The American Journal of Political Science, Prevention Science*, and *Bayesian Analysis*. That is not to say that these journals are unimportant; in fact, the opposite is true, as many of these journals have enviable impact factors. Nonetheless, major databases might not include papers from these more targeted journals: PsycInfo, for instance, only indexes two of the seven journals mentioned previously in this paragraph (the *European Journal of Psychotraumatology* and *Prevention Science*). This could result in substantive researchers being unable to locate these studies even if they are expressly looking for them.

This study aims to take a step back and approach the issue of small samples with MCMC from a broad perspective to make awareness of this issue more widespread by focusing on it exclusively. We show examples from several different types of models that are applicable to different areas of research, both within and across disciplines, rather than narrowing in on specific details and considerations for one particular type of model (for which previous studies can be far more informative). Although we use simulation, our simulations have very few manipulated conditions and serve more as a demonstration of the general issue rather than digging into nuances and recommendations for specific models.

GROWTH MODEL

Data that capture growth over time are often very rich, but as a consequence, can be rather expensive to collect and therefore frequently result in small sample sizes (as noted in the introduction). As is widely known, with small samples, the variance components and fixed effect standard error estimates are highly downwardly biased with ML (Browne & Draper, 2006; McNeish & Stapleton, 2014). However, switching to an MCMC framework and using highly diffuse prior distributions with small samples can lead to highly upwardly biased estimates of these quantities. In fact, the bias of such prior distributions can actually exceed the bias incurred with ML (McNeish, 2016). This is demonstrated in the following subsections.

Illustrative Simulation

Consider 1,000 data sets generated for sample sizes of 20, 30, and 50 from the following model,

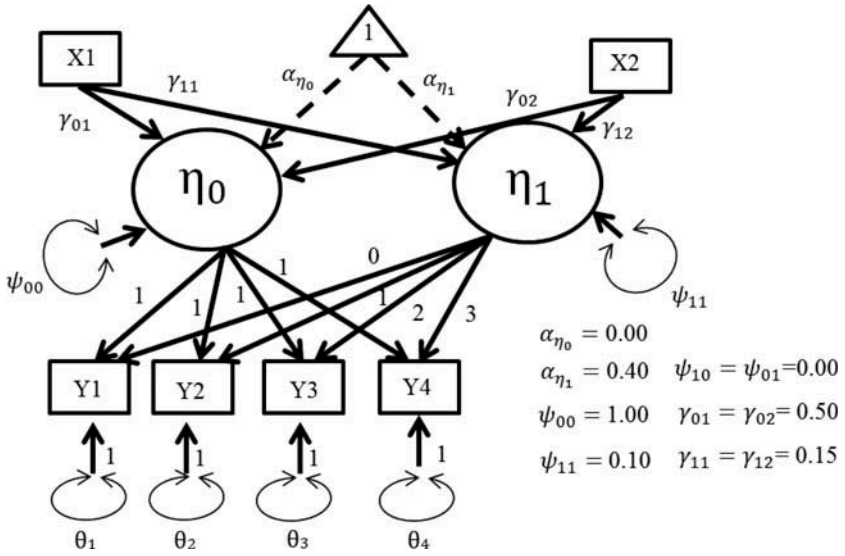

FIGURE 1 Path diagram for growth model simulation with population values.

$$\begin{aligned} Y_{it} &= \eta_{0i} + \eta_{1i} t_t + \varepsilon_{it} \\ \eta_{0i} &= \alpha_{\eta_0} + \gamma_{01} X_{1i} + \gamma_{02} X_{2i} + \zeta_{0i} \\ \eta_{1i} &= \alpha_{\eta_1} + \gamma_{11} X_{1i} + \gamma_{22} X_{2i} + \zeta_{1i} \end{aligned} \quad (1)$$

where Y_{ij} is a continuous outcome variable for the ith person at the tth time, $\eta_{0i}$is the latent intercept for the ith person, $\eta_{1i}$is a latent slope for the ith person, X are time-invariant exogenous predictor variables that predict the values of the latent growth factors, γ are coefficients from the exogenous predictors to the latent growth factors, ζ are random effects such that $\zeta \sim MVN\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\psi_{00} & \psi_{01}\\ \psi_{10} & \psi_{11}\end{bmatrix}\right)$, α are latent growth factor means, and $\varepsilon_{it}$is the error term for the ith person at the tth time where $\varepsilon \sim MVN\left(\mathbf{0}, \begin{bmatrix}\theta_1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \theta_4\end{bmatrix}\right)$. The generated data will feature four time points such that $t = 0, 1, 2, 3$. In the generation model, both X variables will be generated from a standard normal distribution and then dichotomized so that they are binary predictors. X_{1i}will have a 50–50 split reminiscent of biological sex and X_{2i}will have a 60–40 split reminiscent of an ethnic minority indicator. A path diagram and population values are shown in Figure 1. The residual variances are not listed in Figure 1 but were set such that 50% of the variance was explained at each time point as in Bauer and Curran (2003), which would be equal to [1.00, 1.25, 1.75, 2.25] based on the other population values.

Then, each generated data set is modeled under six separate modeling conditions: two frequentist methods, two MCMC conditions with diffuse priors for the variance components, and two MCMC with informative priors for the variance components. All conditions except Condition 2 were conducted in M*plus*: Condition 2 was conducted in SAS PROC MIXED because M*plus* does not offer the necessary estimation method as an option. The six estimation methods are as follows:

1. Full maximum likelihood.
2. Restricted maximum likelihood with a Kenward–Roger correction.
3. MCMC with the M*plus* default improper inverse Wishart prior, $W^{-1}(\mathbf{0}, -p-1)$, for the growth parameter covariance matrix (where p is the dimension of the covariance matrix). This is a widely diffuse prior and allows the variance parameters to be any nonnegative value from 0 to ∞ and the covariance parameters to be any value from $-\infty$ to ∞.
4. MCMC with $\Gamma^{-1}(0.01, 0.01)$ marginal prior[1] for the growth parameters variances and the residual variances and a diffuse normal prior for growth parameter covariance. The $\Gamma^{-1}(0.01, 0.01)$distribution can technically take any value from 0 to ∞ although the distribution is concentrated near 0 and the probability of drawing a value greater than 10 is small but nonzero. This prior is included due to its desirable properties in a recent simulation study by Liu, Zhang, and Grimm (2016) when the variance of the growth parameters is small, the correlation between intercept and slope variances is close to 0, or the growth trajectory is nonlinear.
5. MCMC with a $W^{-1}\left(\begin{matrix}3 & 0\\ 0 & .3\end{matrix}, 6\right)$. prior for the growth parameter covariance matrix. Based on the hyperparameters selected, the prior distribution will have a mean at the population values (1.00 for the intercept, 0.10 for the slope) but with a reasonably large standard deviation (1.41 for the intercept, 0.14 for the slope).[2] This corresponds to a weakly informative prior because the values are centered around the population value but with fairly large uncertainty.
6. MCMC with $W^{-1}\left(\begin{matrix}22 & 0\\ 0 & 2.2\end{matrix}, 25\right)$prior for the growth parameter covariance matrix. Similar to the weakly informative prior, based on the hyperparameters selected, the prior distribution will have a mean at the population values (1.00 for the intercept, 0.10 for the slope). However, the standard deviations are much smaller (0.32 for the intercept, 0.03 for the slope), making this prior distribution a strong prior.

The number of conditions for this simulation (and the simulation in the next section) is purposefully limited because these simulations are intended to serve a didactic purpose. The studies referenced in the introductory sections are more rigorous and demonstrate the concepts we are targeting more comprehensively. Imposing fewer conditions allows us to summarize the main findings of these studies in a manner that can be expressed simply and succinctly.

[1] Marginal prior refers to a prior distribution on an individual element of a matrix. This is opposed to multivariate priors (e.g., the inverse Wishart prior) that place one prior on the entire matrix. In general, the marginal approach can be problematic because a matrix formed by draws from the multiple marginal distributions are not guaranteed to be nonpositive definite whereas this property is guaranteed by drawing from an inverse Wishart distribution. In the context of growth models where the random effect covariance matrix is typically of very low dimension, this issue is less likely to be a concern (Liu et al., 2016).

[2] The inverse Wishart distribution is rather complex so we do not delve into full details given the intended focus of this article (see Muthén & Asparouhov, 2012, for a detailed explanation). As a quick introduction, the first argument is a scale matrix ($\mathbf{\Psi}$) and the second number is the degrees of freedom (v). The larger the degrees of freedom, the more informative the prior will be. The mean of the inverse Wishart distribution is $\frac{\psi}{v-p-1}$ for $v > p+1$ and the variance is $\frac{2\psi^2}{(v-p-1)^2(v-p-3)}$ where p is the dimension of the scale matrix and ψ is an element of the scale matrix $\mathbf{\Psi}$.

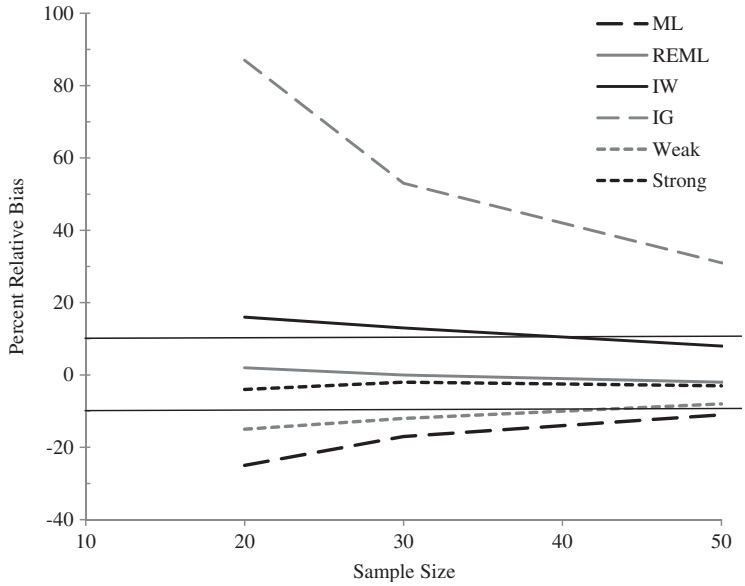

FIGURE 2 Growth model intercept variance relative bias for six different methods across sample size conditions. Superimposed horizontal lines represent the ±10% cutoff recommended by Hoogland and Boomsma (1998). ML = maximum likelihood; REML = restricted maximum likelihood with Kenward–Roger correction; IW = Markov chain Montel Carlo (MCMC)with M*plus* default $W^{-1}(\mathbf{0}, -p-1)$ for the growth parameter covariance matrix; IG = MCMC with $\Gamma^{-1}(0.01, 0.01)$ priors for the diagonal elements of the growth parameter covariance matrix; Weak = $W^{-1}\left(\begin{pmatrix}3 & 0\\ 0 & .3\end{pmatrix}, 6\right)$ prior for the growth parameter covariance matrix; Strong = $W^{-1}\left(\begin{pmatrix}22 & 0\\ 0 & 2.2\end{pmatrix}, 25\right)$ prior for the growth parameter covariance matrix.

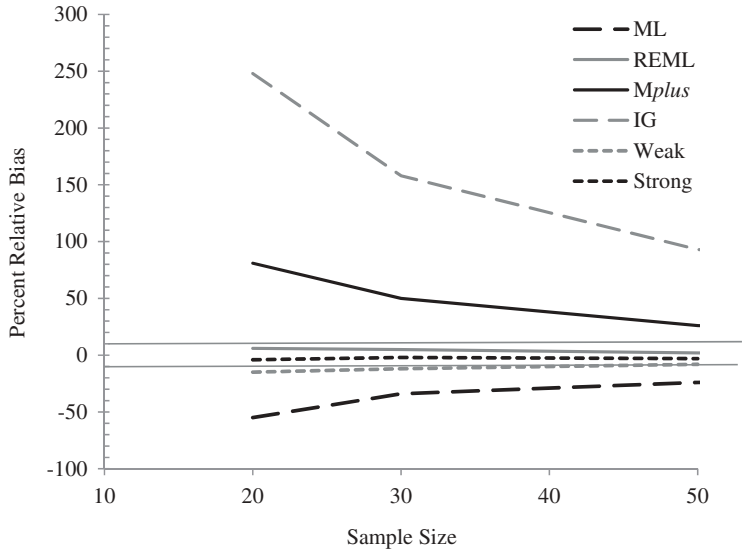

FIGURE 3 Growth model slope variance relative bias for six different methods across sample size conditions. Superimposed horizontal lines represent the ±10% cutoff recommended by Hoogland and Boomsma (1998). ML = maximum likelihood; REML = restricted maximum likelihood with Kenward–Roger correction; M*plus* = Markov chain Monte Carlo (MCMC) with M*plus* default $W^{-1}(\mathbf{0}, -p-1)$ prior for the growth parameter covariance matrix; IG = MCMC with $\Gamma^{-1}(0.01, 0.01)$ priors for the diagonal elements of the growth parameter covariance matrix; Weak = $W^{-1}\left(\begin{pmatrix}3 & 0\\ 0 & .3\end{pmatrix}, 6\right)$ prior for the growth parameter covariance matrix; Strong = $W^{-1}\left(\begin{pmatrix}22 & 0\\ 0 & 2.2\end{pmatrix}, 25\right)$ prior for the growth parameter covariance matrix.

TABLE 1
95% Confidence/Credible Interval Coverage and Power for γ_{01} Parameter by Method

95% Coverage Interval						
Sample Size	ML	REML	IG	M*plus*	Weak	Strong
20	**82**	95	94	**98**	95	96
30	**86**	95	95	97	94	95
50	**89**	95	95	96	94	94
Power						
Sample Size	ML	REML	IG	M*plus*	Weak	Strong
20	*20*	11	10	6	12	13
30	*23*	17	14	10	17	19
50	*29*	26	23	20	25	26
SD						
Sample Size	ML	REML	IG	M*plus*	Weak	Strong
20	0.644	0.644	0.72	0.714	0.624	0.618
30	0.521	0.521	0.559	0.555	0.519	0.508
50	0.390	0.390	0.426	0.425	0.391	0.392

Note. Based on recommendations in Bradley (1978), coverage rates outside [92.5, 97.5] suggest biased standard error estimates. Power for ML is in italics because the Type I error rate was not well controlled. The REML condition fixed variance components estimate to 0 to achieve convergence in 25%, 21%, and 11% of replications for the 20-, 30-, and 50-person conditions, respectively. These replications were removed from the results reported. ML = maximum likelihood; REML = restricted maximum likelihood; M*plus* = Markov chain Monte Carlo (MCMC) with M*plus* default $W^{-1}(\mathbf{0}, -p-1)$prior for the growth parameter covariance matrix; IG = MCMC with $\Gamma^{-1}(0.01, 0.01)$ priors for the diagonal elements of the growth parameter covariance matrix; Weak = $W^{-1}\left(\begin{smallmatrix}3 & 0\\ 0 & .3\end{smallmatrix}, 6\right)$ prior for the growth parameter covariance matrix; Strong = $W^{-1}\left(\begin{smallmatrix}22 & 0\\ 0 & 2.2\end{smallmatrix}, 25\right)$ prior for the growth parameter covariance matrix.

Figure 2 shows the relative bias for the intercept variance and Figure 3 shows the relative bias for the slope variance. Horizontal lines are placed at ±10% based on recommendations in Hoogland and Boomsma (1998) for severely biased estimates. Despite the recurrent assertion that MCMC is more appropriate with small samples, the bias observed in estimates of the intercept variance from the M*plus* default prior is essentially identical (except that the bias is upward rather than downward) to the ML bias. Furthermore, the marginal prior approach performed worse than other methods. REML with Kenward–Roger did not exhibit meaningful bias even with as few as 20 people (although the quality of REML estimates has been shown to progressively deteriorate if the model becomes more complex or sample size is even further reduced; Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; McNeish & Stapleton, 2016).[3] The weakly informative prior showed some downward bias at smaller sample sizes but entered into the reasonable range when sample size increased. The performance of the strong prior was similar to REML and showed very little bias at any sample size.

With small sample growth models, it is also important to consider the standard errors for exogenous predictors as well as power to detect true effects for the exogenous predictors. Table 1 shows 95% confidence or credible interval coverage rates for γ_{01}, the number of replications in which a nonnull effect was found for γ_{01} (with a standardized path coefficient for the population values that was 0.20), and the standard deviation of the coefficient estimates across replications. The last measure is related to the statistical efficiency of the estimates: when estimates are efficient, the sampling variability is reduced, which helps in detecting nonnull effects. This metric is assessed based on relative comparisons where lower values indicate that the method is more efficient.

As has been shown in previous studies, MCMC methods yield proper coverage rates even with small samples (although note that the M*plus* default condition had coverage intervals that were borderline too large), whereas ML estimates yielded coverage intervals that were far too short (the operating Type I error

[3] We want to also note that REML is not immune to convergence issues with smaller sample sizes. SAS will fix variance components estimates to 0 if an estimate is negative or if there is a convergence issue. The percentages of replications in which this occurred for the 20-, 30-, and 50-cluster conditions were 25%, 21%, and 11%, respectively. A preponderance of the issues occurred for the slope variance with a population value (0.10) that was fairly close to 0. These replications are excluded from the results, which might result in the REML results appearing slightly better than they are in actuality because the replications that would conceivably perform the worst might be excluded. Also note that the Kenward–Roger correction cannot operate when the variance components are inadmissible or fixed to zero.

[4] Note that these methods are not available in M*plus* and require that the models be fit in SAS (used in this study) or Stata. This requires that the model be considered a linear mixed model rather than a latent growth model. Curran (2003) noted that these two models are interchangeable mathematically.

rates can be approximated by 1 – [coverage interval], indicating that ML's Type I error rate was nearly four times the nominal 0.05 rate). However, small-sample-specific methods like REML and Kenward–Roger[4] yielded coverage rates that showed no issues. Furthermore, the REML with the Kenward–Roger condition resulted in higher power for the exogenous predictor compared to the diffuse prior MCMC methods, which is likely attributable to the increased efficiency of the estimates. Although the differences in power are rather small, keep in mind that with small samples, power will be low from the onset of the study and it is tantamount to retain the highest amount of power possible. The informative prior MCMC conditions slightly surpass REML with a Kenward–Roger correction in terms of both power and efficiency, showing that MCMC methods can be advantageous for smaller samples. However, this gain in performance is not acquired automatically and prior information about the data needs to be incorporated for this gain to be realized.

Example Analysis

Consider the classical Potthoff and Roy (1964) data that are often used to demonstrate the basics of growth modeling and are used for example analyses in the SAS 9.3 User's Manual. These data contain four measurements of the distance between children's pituitary gland and pteryomaxillary fissure at age 8, 10, 12, and 14. Based on Example 58.8 in the SAS 9.3 User's Guide, Child 20 and Child 24 are outliers in the data and this example analysis will exclude them so that the sample size for the analysis is 25, a small sample by almost any definition. The model features linear growth, covarying random effects for the intercept and (linear) slope that are each predicted by sex, and the residual structure is a homogenous diagonal such that the error variance at each time point can be constrained to be examples. These data will be modeled with the same six estimation conditions used in the previous simulation. To align with the scale of the outcome variable in the Potthoff and Roy data, the weakly informative prior is $W^{-1}\left(\begin{pmatrix}14 & 0\\ 0 & .14\end{pmatrix}, 7\right)$ and the strongly informative prior is $W^{-1}\left(\begin{pmatrix}45 & 0\\ 0 & .45\end{pmatrix}, 15\right)$. The weakly informative prior corresponds to a mean intercept variance of 3.5 with a standard deviation of 3.5 and a mean slope variance mean of 0.035 with a standard deviation of 0.035. The strongly informative prior corresponds to a mean intercept variance of 3.75 with a standard deviation of 1.7 and a mean slope variance mean of 0.0375 with a standard deviation of 0.017. For the MCMC analyses, we require M*plus* to iterate for at least 50,000 iterations with a maximum of 100,000 iterations. Convergence will be determined by using the M*plus* potential scale reduction (PSR; Gelman & Rubin, 1992) and we made the convergence criteria stricter at 0.01 rather than the default 0.05 value. Convergence was also verified by inspecting trace plots of parameters. The posterior is summarized by the median.

Table 2 shows the model estimates for the six estimation methods and the standard errors and posterior standard deviations for the fixed effects. Similar to the simulation study, using the diffuse MCMC priors for the growth factor covariance matrix led to much higher variance estimates and higher posterior standard deviations compared to the frequentist methods or the informative MCMC methods. The diffuse prior MCMC conditions also were unable to detect that either sex or the sex × age interaction had nonnull

TABLE 2
Model Estimates for Six Different Methods From Potthoff and Roy Data With Outliers Removed

	Point Estimate						SE/Posterior SD					
Predictor	ML	REML	IG	M*plus*	Weak	Strong	ML	REML	IG	M*plus*	Weak	Strong
Mean structure parameters												
Intercept	22.95	22.95	22.97	22.94	22.93	22.93	0.511	0.533	0.586	0.626	0.531	0.517
Age	0.69*	0.69*	0.69*	0.69*	0.69*	0.69*	0.064	0.067	0.072	0.080	0.066	0.065
Sex	−1.74*	−1.74*	−1.74	−1.73	−1.84*	−1.83*	0.770	0.803	0.883	0.942	0.831	0.829
Sex × Age	−0.21*	−0.21*	−0.21	−0.21	−0.22*	−0.22*	0.096	0.100	0.108	0.121	0.104	0.104
Covariance structure parameters												
Var (Int.)	3.05	3.36	3.82	4.39	3.44	3.41						
Cov (Int., Slope)	0.06	0.06	0.05	0.03	0.03	0.01						
Var (Slope)	0.01	0.02	0.02	0.04	0.03	0.03						
Residual variance	0.86	0.86	0.93	0.89	0.89	0.86						

Note. ML = maximum likelihood; REML = restricted maximum likelihood; M*plus* = Markov chain Monte Carlo (MCMC) with M*plus* default $W^{-1}(\mathbf{0}, -p-1)$prior for the growth parameter covariance matrix; IG = MCMC with $\Gamma^{-1}(0.01, 0.01)$ priors for the diagonal elements of the growth parameter covariance matrix; Weak = $W^{-1}\left(\begin{pmatrix}14 & 0\\ 0 & .14\end{pmatrix}, 7\right)$; Strong =$W^{-1}\left(\begin{pmatrix}45 & 0\\ 0 & .45\end{pmatrix}, 15\right)$

*Indicates that 0 is not included in the 95% confidence/credible interval.

effects. This is likely attributable to the decreased efficiency of the diffuse priors that can be directly seen by the magnitude of the posterior standard deviations relative to the other methods. The frequentist and informative prior MCMC conditions were each able to detect that these effects were nonnull.

MULTILEVEL MODELING

Cross-sectionally clustered multilevel data (e.g., students clustered within schools, cluster randomized trials) are also notorious for having small sample sizes at the cluster level due to difficulties such as high financial costs associated with including each cluster, geographic sparsity of clusters such that it is logistically challenging to collect data, or sparsity of the population in general (e.g., schools that specialize in students with particular disabilities). In the case of cluster randomized trials where assignment of an intervention is assigned to entire higher level units (e.g., school, hospital), the number of clusters tends to be quite small, especially for studies that have limited funding and are trying to demonstrate proof of concept, as is common in educational sciences. As with the growth model example, we demonstrate that it is not sufficient to merely switch to an MCMC framework and rely on software defaults with small samples using an illustrative simulation and an example analysis.

Illustrative Simulation

Consider a standard cluster randomized trial where there is a treatment effect at Level 2 and the primary objective of the model is to determine whether the treatment effect is nonnull. We generate 1,000 data sets from the following model:

$$\begin{aligned} Y_{ij} &= \beta_{0\mathrm{j}} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \mathrm{r}_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}W_{1X} + \mathrm{u}_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}W_{1j} \\ \beta_{3j} &= \gamma_{30} + \gamma_{31}W_{1j} \end{aligned} \tag{2}$$

where Y_{ij} is a continuous outcome variable for the ith individual in the jth cluster, X_{kij} is the value of the kth Level 1 predictor for the ith person in the jth cluster, W_{1j} is a Level 2 predictor for the jth cluster, γ are fixed effect coefficients that are constant across all clusters, and u_{0j} is a cluster-specific random effect for the intercept (a.k.a. Level 2 residual) of the jth cluster where $u_{0j} \sim N(0, g_{00})$, and r_{ij} is the Level 1 residual for the ith person in the jth cluster where $r_{ij} \sim N(0, \sigma^2)$. The W_{1j} predictor represents the treatment effect (assigned at the cluster level in cluster randomized trials) and is generated as a binary variable with 50–50 prevalence (half of the generated clusters were assigned to the control group and half to the treatment group). X_{1ij} is generated from a standard normal distribution and is reminiscent of a pretest score predictor (where Y_{ij} would be the posttest score), $X_{2ij}$is generated as a binary variable with 50–50 prevalence similar to biological sex, and X_{3ij} is generated as a binary variable with 25–75 prevalence similar to English language learner status in the applied example that follows the simulation. Figure 4 shows a path diagram and population values for the data generation model using Curran and Bauer's (2007) notation. Population values for fixed effects were selected based on the applied example that is presented in the next section and variance component population values were selected so that the unconditional intraclass correlation would be approximately 0.20, which is a typical value for contexts where cluster randomized trials are carried out in behavioral science research (Hedges & Hedberg, 2007).

The same six estimation methods used in the growth model example are used for this illustrative example with a few minor changes. The generated cluster randomized trial data only has a single random effect for the intercept, meaning that the M*plus* default prior for the random effect is an improper inverse gamma prior, $\Gamma^{-1}(-1, 0)$, instead of an improper inverse Wishart. The default inverse gamma prior is the univariate version of the improper inverse Wishart, so it has similar properties albeit in the univariate case. Similarly, the second noninformative prior will again be $\Gamma^{-1}(0.01, 0.01)$for the random intercept variance. Because these data have only a single random effect, there is no worry about drawing covariance matrix elements from separate distributions as was present in the growth model section. The weakly informative prior is a $\Gamma^{-1}(3, 3.25)$ that has a mean of 1.625 and a standard deviation of 1.625.[5] The strongly informative prior is $\Gamma^{-1}(12, 18)$which has a mean of 1.64 and a standard deviation of 0.52. To more closely mimic properties of cluster randomized trials with smaller samples, the number of clusters is generated to be 8, 10, and 14 and the number of individuals within each cluster will be unbalanced between 7 and 14 per cluster (Eldridge et al., 2004, reported that only 60% of cluster randomized trials have an average cluster size greater than 20).

Figure 5 shows the relative bias of the intercept variance for each estimation method. As expected, even for a fairly straightforward model, the ML estimates were highly downwardly biased and the magnitude of the bias continues to exceed 25% with 14 clusters. However, using small sample methods in the frequentist framework, the bias was much reduced and acceptable estimates were obtained with as few as 10 clusters. The weakly informative prior performed almost identically to REML (the lines are difficult to distinguish in Figure 5 because they are nearly on top of one another) and the strongly

[5] The inverse gamma distribution has a mean equal to $\frac{\beta}{\alpha-1}$ for $\alpha > 1$ and a variance equal to $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$ for α the shape parameter (the first hyperparameter) and β the scale parameter (the second hyperparameter).

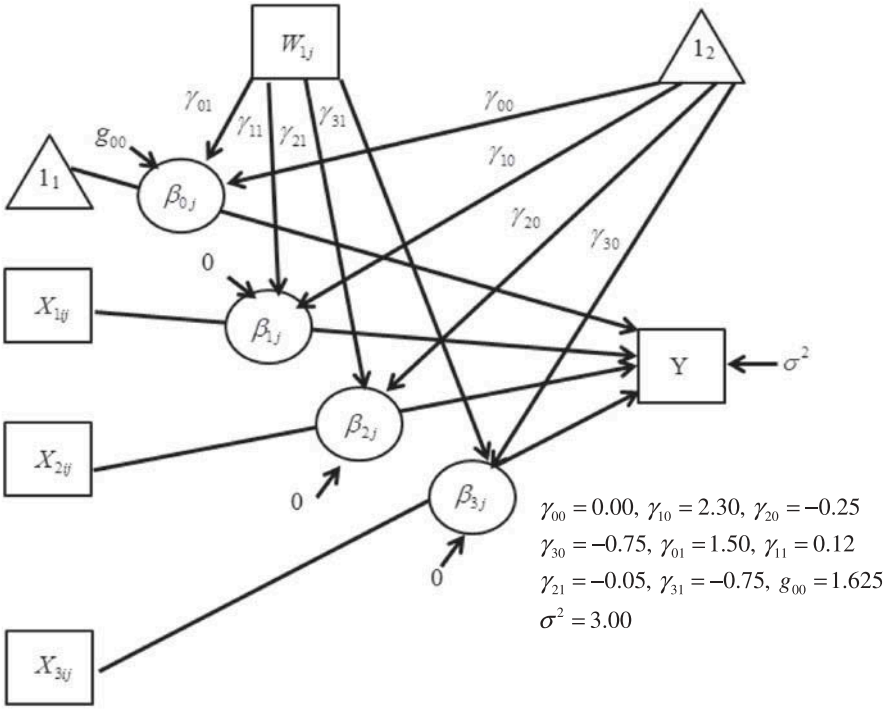

FIGURE 4 Path diagram for multilevel model data generation and population values. Population values were selected to roughly correspond to values present in the applied example that follows.

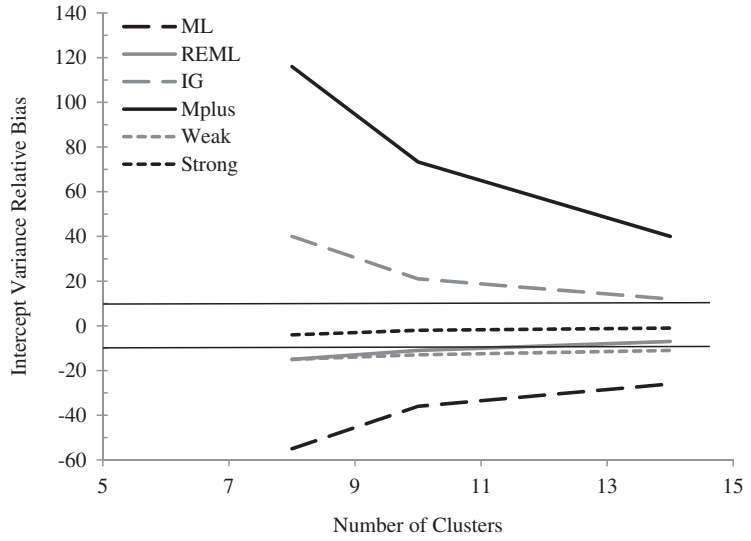

FIGURE 5 Cluster randomized trial intercept variance relative bias for four different methods across sample size conditions. Superimposed horizontal lines represent the ±10% cutoff recommended by Hoogland and Boomsma (1998). ML = maximum likelihood; REML = restricted maximum likelihood with Kenward–Roger correction; M*plus* = Markov chain Monte Carlo (MCMC) with M*plus* default $\Gamma^{-1}(-1, 0)$ prior for the intercept variance; IG = MCMC with $\Gamma^{-1}(0.01, 0.01)$prior for the intercept variance; Weak = MCMC with $\Gamma^{-1}(3, 3.25)$ prior for intercept variance; Strong = MCMC with $\Gamma^{-1}(12, 18)$ prior for intercept variance.

TABLE 3
95% Confidence/Credible Interval Coverage Rates by Predictor for 0.40 Effect Size Condition

	8 Clusters						10 Clusters						14 Clusters					
Predictor	ML	REML	IG	M*plus*	Weak	Strong	ML	REML	IG	M*plus*	Weak	Strong	ML	REML	IG	M*plus*	Weak	Strong
X_{1ij}	93	95	95	94	94	94	94	95	95	95	95	95	94	95	95	95	94	93
X_{2ij}	94	96	96	97	94	94	95	96	96	96	95	96	94	96	95	96	95	94
X_{3ij}	94	96	96	96	93	93	95	96	96	96	93	94	95	95	96	96	94	94
$X_{3ij} \times W_{1j}$	**92**	94	96	95	94	94	96	96	97	97	93	93	95	96	96	97	94	94
W_{1j}	**88**	93	95	**98**	93	94	**92**	96	96	97	94	95	**92**	95	95	95	93	94
$X_{1ij} \times W_{1j}$	**91**	93	95	95	95	95	94	95	95	95	94	94	94	96	95	95	94	95
$X_{2ij} \times W_{1j}$	94	95	96	96	95	95	94	95	95	95	95	95	94	95	95	95	94	95
Intercept	95	95	96	97	93	95	96	96	96	96	95	95	95	95	94	95	95	95

Note. Based on recommendations in Bradley (1978), coverage rates outside [92.5, 97.5] suggest biased standard error estimates. Bold entries correspond to values outside Bradley's range. ML = full maximum likelihood; REML = restricted maximum likelihood with Kenward–Roger correction; IG = Markov chain Monte Carlo (MCMC) with $\Gamma^{-1}(0.01, 0.01)$ prior for intercept variance; M*plus* = MCMC with $\Gamma^{-1}(-1, 0)$ prior for intercept variance; Weak = MCMC with $\Gamma^{-1}(3, 3.25)$ prior for intercept variance; Strong = MCMC with $\Gamma^{-1}(12, 18)$ prior for intercept variance.

informative prior outperformed REML. In contrast, opting for MCMC with diffuse priors using either the M*plus* default or $\Gamma^{-1}(0.01, 0.01)$is less effective with respect to obtaining unbiased variance estimates and more clusters are actually needed to reach unbiased estimates compared to REML with a Kenward–Roger correction.[6]

Table 3 shows the 95% confidence interval coverage of each parameter, for each estimation type across sample size conditions. As expected, the ML coverage intervals are too short, which indicates downwardly biased standard error estimates. The coverage intervals with the M*plus* default prior tended to be fairly wide with smaller samples (although, with a few exceptions, they were mostly within the Bradley [1978] bounds and were closer to nominal rate than the ML intervals). The coverage intervals for REML, the weakly informative prior, the strongly informative prior, and $\Gamma^{-1}(0.01, 0.01)$ were not problematic for any conditions and gave more desirable coverage intervals than the M*plus* default.

Because the primary interest in a cluster randomized trial is in testing the treatment effect, we manipulated the effect size of the treatment and calculated the empirical power for REML, the weakly informative prior, the strongly informative prior, and the $\Gamma^{-1}(0.01, 0.01)$ prior because these methods maintained consistently good coverage intervals across sample sizes. We included four Cohen's *d* effect sizes for the difference between the treatment and control groups: 0.10, 0.20, 0.30, and 0.40 (this value is used in Figure 4). Table 4 shows these results. From this illustrative simulation, the difference among all methods appears to be quite minimal (the strongly informative prior showed a very slight increase in power). Although not reported, the standard deviation of the treatment effect across replications also showed only very minor differences, indicating that the efficiency of the different methods is about equal.

TABLE 4
Empirical Power for Finding Nonnull Treatment Effect for Various Effect Sizes and Sample Sizes

Effect Size (d)	Number of Clusters	REML	IG	Weak	Strong
0.10	8	7	7	7	8
	10	8	8	8	9
	14	9	8	8	9
0.20	8	14	12	13	14
	10	14	14	14	16
	14	17	17	17	19
0.30	8	19	18	18	19
	10	26	25	26	28
	14	34	34	35	37
0.40	8	29	27	28	28
	10	40	41	41	42
	14	56	58	58	60

Note. REML = restricted maximum likelihood with Kenward–Roger correction; IG = Markov chain Monte Carlo (MCMC) with $\Gamma^{-1}(0.01, 0.01)$prior for intercept variance; Weak = MCMC with $\Gamma^{-1}(3, 3.25)$ prior for intercept variance; Strong = MCMC with $\Gamma^{-1}(12, 18)$ prior for intercept variance.

Example Analysis

We also demonstrate this point with a cluster randomized trial example from education. The data are from an Institute of Educational Sciences Development Grant that investigated the efficacy of a Reading Buddies intervention to assess whether a researcher-designed treatment applied at the classroom level affected students' reading vocabulary compared to

[6] The percentage of replications in which this occurred for the 8-, 10-, and 14-cluster conditions were 9%, 3%, and 1%, respectively. These replications are excluded from the results, which might result in the REML results appearing slightly better than they are in actuality because the replications that would conceivably perform the worst might be excluded.

students in a control group who did not receive the treatment. The example data that we expound on is used to address only one of several research questions posed within this grant and included 203 kindergarten students clustered within 12 classrooms in a semiurban, mid-Atlantic region school district. The outcome measure was students' posttest vocabulary scores (as measured by the Peabody Picture Vocabulary Test Growth Score Value [PPVT–GSV]) and is predicted by treatment group status, English language learner (ELL) status, PPVT–GSV pretest score, and interactions of the treatment effect with ELL and pretest scores. PPVT pretest score was grand-mean centered prior to being included in the model in accordance with recommendations from Enders and Tofighi (2007) because the primary interest was on the treatment effect (located at Level 2). In Raudenbush and Bryk (2002) notation, the model is written out as

$$\begin{aligned} PPVT\ Post - Test_{ij} &= \beta_{0j} + \beta_{1j}(\mathrm{ELL_{ij}}) + \beta_{2j} \\ (\mathrm{PPVT\ Pre - Test_{ij}} &- \overline{\mathrm{PPVT\ Pre - Test}}) + \mathrm{r_{ij}} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\mathrm{Treatment_j} + \mathrm{u_{0j}} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\mathrm{Treatment_j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}\mathrm{Treatment_j} \end{aligned} \tag{3}$$

All models but REML with a Kenward–Roger correction were estimated in M*plus*. For the MCMC models, the number of iterations was set to a minimum of 50,000, the convergence of the PSR criteria was set to 0.01, and the posterior was summarized with the median. Trace plots were also inspected to verify convergence.

Table 5 reports the estimates from the same six estimation methods included in the simulation in the previous section. The hyperparameters for the weakly and strongly informative priors differ between the example analysis and the simulation to account for differences in the scale of the outcome. The weakly informative prior for the intercept variance is $\Gamma^{-1}(5, 30)$, which results in a distribution with a mean of 7.5 and a standard deviation of 4.3. The strongly informative prior is $\Gamma^{-1}(15, 105)$, which results in a distribution with a mean of 7.5 and a standard deviation of 2.1. Similar patterns to the growth model example again emerge: The ML intercept variance is noticeably lower than the MCMC estimates or the REML estimate, which is consistent with the finding that ML underestimates this parameter. Similarly, both diffuse prior MCMC conditions resulted in estimates that were higher than the REML estimates. In particular, the M*plus* default prior estimated the intercept to be about 56% larger than the REML estimate, which had a noticeable effect of the posterior standard deviation that was about 17% higher than either REML or the $\Gamma^{-1}(0.01, 0.01)$ prior (suggestive of the decreased efficiency seen in the simulation results). This inflated posterior standard deviation is a cause for concern with small samples because the ability to detect nonnull effects will be limited from the onset due to the small sample, and reduced efficiency as reflected by an inflated posterior standard deviation will make such detection of a nonnull effect even more difficult (although in this particular case, all six methods were able to detect that the treatment effect was nonnull). The Cohen's *d* effect size for the treatment effect in the model is between the mid 0.30s and the low 0.40s, depending on the estimation method.

META-ANALYSIS

Meta-analyses are another commonly occurring data analytic situation in which sample sizes tend to be rather small

TABLE 5
Comparison of Estimates for Reading Buddies Data

	Point Estimates						SE/Posterior SD					
Predictor	ML	REML	IG	M*plus*	Weak	Strong	ML	REML	IG	M*plus*	Weak	Strong
Mean structure parameters												
Intercept	126.1	126.0	126.1	126.0	126.3	126.1						
ELL	3.14	3.26	3.10	3.36	3.21	3.17	2.33	2.43	2.38	2.44	2.39	2.37
Pretest	0.88*	0.88*	0.88*	0.88*	0.88*	0.88*	0.06	0.06	0.05	0.06	0.05	0.05
Treat	6.98*	6.96*	6.96*	6.95*	7.00*	6.91*	2.37	2.56	2.54	2.98	2.55	2.54
ELL × Treat	−6.68*	−6.70*	−6.57*	−6.70*	−6.71*	−6.61*	3.19	3.29	3.24	3.31	3.28	3.28
Pre × Treat	−0.19*	−0.19*	−0.19*	−0.19*	−0.19*	−0.19*	0.07	0.08	0.07	0.08	0.07	0.07
Covariance structure parameters												
Var (Int.)	5.00	7.17	8.30	11.19	6.44	6.70						
Residual var.	62.69	63.99	65.32	64.98	65.13	65.09						

Note. ML= full maximum likelihood; REML = restricted maximum likelihood with Kenward–Roger correction; IG = Markov chain Monte Carlo (MCMC) with $\Gamma^{-1}(0.01, 0.01)$prior for intercept variance; M*plus* = MCMC with $\Gamma^{-1}(-1, 0)$prior for intercept variance; Weak = MCMC with $\Gamma^{-1}(5, 30)$prior for intercept variance; Strong = MCMC with $\Gamma^{-1}(15, 105)$ prior for intercept variance; ELL = English language learner.

*Indicates that 0 is not included in the 95% confidence/credible interval.

because there are often not an overwhelming number of studies that are conducted on the same phenomenon (Cumming, 2013). Although there are multiple ways to analyze data from meta-analyses (e.g., fixed effect, random effects, robust regression), multilevel models remain a popular method due to the ability to partition the variance into between-study and within-study components (Goldstein, Yang, Omar, Turner, & Thompson, 2000; Hox, 2010; van den Noortgate & Onghena, 2003). This can help inform researchers whether there might be relevant covariates that explain why different studies produced effect sizes of different magnitudes.

In this section, in the interest of brevity, we do not conduct an illustrative simulation study because it would overlap considerably with the simulation in the previous section. Instead, this section only features an example analysis to again highlight that resorting to the MCMC methods without carefully considering the prior distributions does not solve the difficulties encountered with small samples.

Example Analysis

The data we use here are from Chapter 11 of Hox (2010), which is also used in Cheung (2008) and features 20 different studies that compare the Cohen's *d* effect size between experimental and control groups. The length of time in which the study was conducted (1–9 weeks) is also included as a possible covariate. First, we fit an unconditional model with Cohen's *d* as the outcome with the variance partitioned into a between-study and within-study components. This step of the modeling process helps determine whether the studies found approximately the same effect size or whether different aspects of each study might have led to different effect sizes. Then, we proceed by including the number of weeks the study was conducted as a covariate to see whether the covariate reduces the between-study variance. Each model is estimated with six different methods:

1. Full maximum likelihood.
2. Restricted maximum likelihood with a Kenward–Roger correction.
3. MCMC with an $\Gamma^{-1}(-1, 0)$ prior for the intercept random effect (the M*plus* default for models with a single random effect).
4. MCMC with an $\Gamma^{-1}(0.01, 0.01)$prior for the intercept random effect.
5. MCMC with an $\Gamma^{-1}(3, 0.07)$ prior for intercept variance for conditional model. This results in a prior distribution with a mean and standard deviation of 0.035. For the unconditional model, the prior is $\Gamma^{-1}(3, 0.30)$, which corresponds to a mean and standard deviation of 0.15. These priors represent weakly informative priors.
6. MCMC with an $\Gamma^{-1}(12, 0.39)$ prior for intercept variance for conditional model. This results in a prior distribution with a mean of 0.035 and a standard deviation of 0.01. For the unconditional model, the prior is $\Gamma^{-1}(12, 1.65)$, which corresponds to a distribution with a mean of 0.15 and a standard deviation of 0.04. These priors represent strongly informative priors.

The ML model is fit in M*plus* and the REML model (with a Kenward–Roger correction) is fit in SAS PROC MIXED because M*plus* does not offer these options. The MCMC models were fit in M*plus* with a minimum of 50,000 iterations and the posterior was summarized with the median. Convergence was assessed via a PSR value less than 0.01 and through visual inspection of trace plots for each parameter.

Because the sampling standard deviation for these studies was known, the model can be treated as a known-variance model. This can be done using one of two methods. First, the data can be weighted by the inverse of the sampling variance and the residual variance can be constrained to 1. Although very straightforward, this method is not viable for the interest at hand because M*plus* does not allow for multilevel models with weights and MCMC estimation. The second method (which we use) is outlined by Cheung (2008) and discussed in Hox (2010). In this method, the sampling standard deviation is added as a predictor of the residual error (but not the intercept random effects or as a predictor of the outcome variable). The variance of the sampling standard deviation predictor is then also constrained to 0. The model equation can be difficult to interpret, so Figure 6 provides the path diagram for the

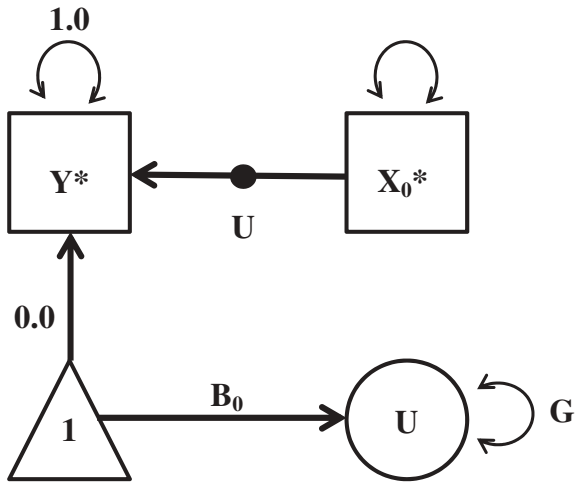

FIGURE 6 Path diagram for unconditional multilevel meta-analysis. Y^* are the transformed effect sizes such that $Y^* = Y/\sqrt{Var(Y)}$, $X_0^* = Var(Y)^{-1/2}$, B_0 is the overall average effect size, U is a random slope that is needed to take into account that each study has a different effect size variance, and G is the between-study intercept variance. Paths with numeric labels are constrained to that value.

TABLE 6
Comparison of Estimates for Meta-Analysis Data Across Estimation Methods

	Point Estimate						SE/Posterior SD					
Predictor	ML	REML	IG	M*plus*	Weak	Strong	ML	REML	IG	M*plus*	Weak	Strong
Unconditional model												
Intercept	0.579	0.58	0.579	0.583	0.579	0.581	0.105	0.108	0.111	0.123	0.106	0.109
Var (Int.)	0.132	0.145	0.136	0.185	0.134	0.147						
With weeks as a covariate												
Intercept	−0.214	−0.217	−0.214	−0.215	−0.211	−0.216	0.193	0.205	0.218	0.239	0.202	0.206
Weeks	0.139	0.140	0.139	0.140	0.138	0.140	0.032	0.033	0.036	0.039	0.033	0.034
Var (Int.)	0.023	0.036	0.038	0.064	0.031	0.035						

Note. ML= full maximum likelihood; REML = restricted maximum likelihood with Kenward–Roger correction; IG = Markov chain Montel Carlo (MCMC) with $\Gamma^{-1}(0.01, 0.01)$ prior for intercept variance; M*plus* = MCMC with $\Gamma^{-1}(-1, 0)$ prior for intercept variance; Weak = MCMC with $\Gamma^{-1}(3, 0.07)$ prior for intercept variance for conditional model and $\Gamma^{-1}(3, 0.30)$ prior for the intercept variance in the unconditional model; Strong = MCMC with $\Gamma^{-1}(12, 0.39)$ prior for intercept variance for conditional model and $\Gamma^{-1}(12, 1.65)$ prior for intercept variance for unconditional model. The 95% confidence/credibility interval for Weeks did not contain 0 for any estimation method.

unconditional model instead. For the REML model that will be fit in SAS PROC MIXED, the first method is used instead because it is far simpler to program and produces equivalent results.

Table 6 shows the results of these two models with the six estimation methods. To continue the common theme of this article, note the differences in the random intercept variance between the diffuse prior MCMC models and the other methods: With the M*plus* default prior, the intercept variance is about 20% larger than the other methods. When the number of weeks covariate is added to the model, the fixed effect estimates across all methods are nearly identical. However, again, the intercept variances are quite disparate between methods, with the M*plus* default prior being approximately 80% higher than the frequentist or informative Bayesian methods.

DISCUSSION

To reiterate, we are not saying that Bayesian methods are necessarily poor with small samples or that frequentist methods necessarily outperform Bayesian methods. In fact, as many methodological studies have shown, quite the opposite can be true. The points we are trying to emphasize are that (a) Bayesian methods do not effortlessly alleviate small sample problems, and (b) although Bayesian methods have properties that make them more conducive to modeling small sample data, prior distributions must be carefully considered to take full advantage of these properties. Furthermore, switching from a frequentist framework to a Bayesian framework without exercising due diligence in considering the prior distributions can exacerbate the small sample problem rather than ameliorate it.

When sample sizes are small, to obtain trustworthy estimates that maximize one's ability to find nonnull effects, some legwork must be carried out—merely switching modeling paradigms often does very little to address this situation. If the frequentist framework is initially desired for reasons of familiarity, convenience, philosophy, and so on, there are methods that can address small sample issues for many types of models. For most clustered data problems that can be addressed with linear models (i.e., multilevel models and growth models for normal outcomes), REML estimation and a Kenward–Roger correction can often yield unbiased estimates of variance components and regression coefficients (although, see McNeish, 2016, for types of growth models where REML is not available).

REML estimation and the Kenward–Roger correction have not been extended to the SEM framework broadly and currently only exist for models that can be translated to a regression framework. Therefore, these options are not always available. As a broader alternative to dealing with small samples, Staggs (2015) found that bootstrapped standard errors performed nearly as well as Kenward–Roger standard errors and are applicable to a much wider variety of models than the Kenward–Roger correction. To address variance component bias, Raudenbush and Bryk (2002) noted a heuristic approximation between ML and REML variance components such that REML variance estimates are approximately equal to $(J - K)/J$ where J is the number of clusters (or individuals in a longitudinal study) and K is the number of predictor variables in the model (including the intercept). In the SEM context, K would be equal to the number of paths from exogenous predictors to the latent variables plus the latent variable means.

By a similar token, Bayesian methods are capable of outperforming frequentist methods with small samples by providing less biased estimates, increased efficiency, and an increased ability to determine nonnull effects when they are present; however, a similarly nuanced analytic process must be followed. Rather than finite sample corrections to estimates, to augment performance with small samples

researchers need to include outside information in the form of informative priors. For instance, researchers might need to consult content experts, meta-analyses, or review studies in the area of interest to obtain informative, accurate priors that can meaningfully contribute to posterior distributions. It is important to ensure that the information on which informative priors are based is accurate with smaller samples, otherwise the resulting estimates and posterior standard deviations could be biased if misleading informative priors are utilized (e.g., Depaoli, 2014). Although this might seem like a somewhat pessimistic conclusion regarding the utility of MCMC with small samples, note that the simulation studies in the previous sections showed that the prior need not be very strong for performance to be desirable. In terms of bias, power, appropriateness of coverage intervals, and efficiency, the weakly informative prior performed nearly as well as REML with a Kenward–Roger correction or a strongly informative prior. This demonstrates that researchers need not have extensive prior information to obtain useful MCMC estimates; if the prior is set in the vague vicinity of the population value, even with a fairly large variance, then the advantages of MCMC with small samples can be realized. This finding is particularly helpful for models that cannot be treated with REML and the Kenward–Roger correction because weakly informative priors are widely applicable to various model types, are not overly difficult implement, and reduce the chance that the posterior will be misleading due to an overly precise prior distribution.

If one has very little prior information and wishes to pursue Bayesian methods or requires Bayesian estimation because, for instance, ML methods have not converged or produce inadmissible estimates (e.g., Heywood cases), other strategies can also be employed. Gelman (2006) recommended using half-Cauchy priors for variance components; both analytical work (Polson & Scott, 2012) and simulation work (McNeish & Stapleton, 2016) have shown that this type of prior is superior to other commonly recommended, noninformative priors with small samples. O'Malley and Zaslavsky (2005) also outlined a multivariate extension to the half-Cauchy prior. Unfortunately, this type of prior is not available in M*plus*.

The omission of prior distributions that are appropriate for small samples in M*plus* can be a little concerning, especially because much of the program's documentation on the BAYES module notes advantages with small sample sizes (e.g., Asparouhov & Muthén, 2010; Muthén, 2010; Muthén & Asparouhov, 2012). The half-Cauchy distribution (and other distributions) can be programmed manually (they are not available as preprogrammed options, however) in more general Bayesian software (e.g., see McNeish & Stapleton, 2016 for SAS code), although specification of more complex models can be rather difficult for nonexpert users in general Bayesian software. Although software users are ultimately responsible for their modeling choices, the program defaults and the limited number of prior distributions in M*plus* can facilitate uninformed modeling choices. Given the ease of Bayesian model specification in M*plus* and its popularity with applied researchers, prior distributions that perform well with small samples (e.g., half-Cauchy and its multivariate extensions) could be a valuable addition to future versions.

To conclude, we are not expressly advocating that researchers use either frequentist or Bayesian methods to combat small samples. Neither of these choices are necessarily "safer" for small sample problems than the other: Bayesian methods can go awry if an incorrect informative prior is chosen or if default priors are left intact, and no corrective procedure can prevent frequentist methods from eventually breaking down when the model becomes too complex for the data. Rather, we want researchers to be aware that small sample methods exist in either framework and that small samples can be addressed with either frequentist or Bayesian methods—the choice of modeling framework to employ is less important than the choices made within the selected framework—and that there are associated risks in either framework. The common adage that diffuse priors cannot do any harm within a Bayesian framework and at worst produce estimates on par with ML is decidedly misleading with smaller samples because the data do not contain enough information to override the chosen prior distribution in the posterior distribution.

REFERENCES

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, *82*, 436–476. doi:10.3102/0034654312458162

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using Mplus* (Tech. Rep.). Retrieved from http://www.statmodel2.com/download/BayesAdvantages18.pdf

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*, 151–164. doi:10.1037/a0030642

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3*, 77–85. doi:10.1111/bmsp.1950.3.issue-2

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*, 338–363. doi:10.1037/1082-989X.8.3.338

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*, 181–197. doi:10.1207/S15327906Mb340203

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152. doi:10.1111/bmsp.1978.31.issue-2

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473–514. doi:10.1214/06-BA117

Cheung, M. W. L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychological Methods*, *13*, 182–202. doi:10.1037/a0013163

Cumming, G. (2013). The new statistics why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*, 529–569. doi:10.1207/s15327906mbr3804_5

Curran, P. J., & Bauer, D. J. (2007). Building path diagrams for multilevel models. *Psychological Methods*, *12*, 283–297. doi:10.1037/1082-989X.12.3.283

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*, 186–219. doi:10.1037/a0031609

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling*, *21*, 239–252. doi:10.1080/10705511.2014.882686

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling*, *22*, 327–351. doi:10.1080/10705511.2014.937849

Depaoli, S., & van de Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*. Advance online publication. doi:10.1037/met0000065

Doron, J., & Gaudreau, P. (2014). A point-by-point analysis of performance in a fencing match: Psychological processes associated with winning and losing streaks. *Journal of Sport & Exercise Psychology*, *36*, 3–13. doi:10.1123/jsep.2013-0043

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B*, *62*, 355–366. doi:10.1111/1467-9868.00236

Dunson, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, *153*, 1222–1226. doi:10.1093/aje/153.12.1222

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials*, *1*, 80–90. doi:10.1191/1740774504cn006rr

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. doi:10.1037/1082-989X.12.2.121

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*, 372–384. doi:10.3758/BRM.41.2.372

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*, 515–534. doi:10.1214/06-BA117A

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211. doi:10.1080/19345747.2011.618213

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

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics*, *49*, 399–412.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–338. doi:10.1080/01621459.1977.10480998

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. doi:10.3102/0162373707299706

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367. doi:10.1177/0049124198026003003

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*, 87–93.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. New York, NY: Wiley.

Jordan, M. (2011). What are the open problems in Bayesian statistics. *The ISBA Bulletin*, *18*, 1–4.

Kadane, J. B. (2015). Bayesian methods for prevention research. *Prevention Science*, *16*, 1017–1025. doi:10.1007/s11121-014-0531-x

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. doi:10.2307/2533558

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*, 2583–2595. doi:10.1016/j.csda.2008.12.013

Kliem, S., Kröger, C., & Kosfelder, J. (2010). Dialectical behavior therapy for borderline personality disorder: A meta-analysis using mixed-effects modeling. *Journal of Consulting and Clinical Psychology*, *78*, 936–951. doi:10.1037/a0021015

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*, 2401–2428. doi:10.1002/(ISSN)1097-0258

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*, 653–686. doi:10.1207/s15327906mbr3904_4

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics*, *2009*, 1–18. doi:10.1155/2009/537139

Levy, R., & Choi, J. (2013). Bayesian structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 563–623). Charlotte, NC: Information Age.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling*, *23*, 354–367. doi:10.1080/10705511.2015.1057285

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226. doi:10.1146/annurev.psych.51.1.201

McNab, Y. C., Malloy, D. C., Hadjistavropoulos, T., Sevigny, P. R., McCarthy, E. F., Murakami, M., … Liu, P. L. (2010). Idealism and relativism across cultures: A cross-cultural examination of physicians' responses on the Ethics Position Questionnaire. *Journal of Cross-Cultural Psychology*, *42*, 1272–1278. doi:10.1177/0022022110383313

McNeish, D. M. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using M*plus*. *Journal of Educational and Behavioral Statistics*, *41*, 27–56. doi:10.3102/1076998615621299

McNeish, D., & Stapleton, L. M. (2014). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295–314. doi:10.1007/s10648-014-9287-x

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*. Advance online publication. doi:10.1080/00273171.2016.1167008

Moore, T. M., Reise, S. P., Depaoli, S., & Haviland, M. G. (2015). Iteration of partially specified target matrices: Applications in exploratory and

Bayesian confirmatory factor analysis. *Multivariate Behavioral Research*, *50*, 149–161. doi:10.1080/00273171.2014.973990

Muthén, B. (2010). Bayesian analysis in M*plus*: A brief introduction (Unpublished manuscript). Retrieved from https://www.statmodel.com/download/IntroBayesVersion%203.pdf

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

O'Malley, A. J., & Zaslavsky, A. M. (2005). Variance-covariance functions for domain means of ordinal survey items. *Survey Methodology*, *31*, 169–182.

Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*, 887–902. doi:10.1214/12-BA730

Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, *51*, 313–326. doi:10.1093/biomet/51.3-4.313

Price, L. R. (2012). Small sample properties of Bayesian multivariate autoregressive time series models. *Structural Equation Modeling*, *19*, 51–64. doi:10.1080/10705511.2012.634712

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in *Personality and Social Psychology Bulletin*. *Personality and Social Psychology Bulletin*, *28*, 1629–1646. doi:10.1177/014616702237645

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37–52. doi:10.1007/BF02294318

Schoeneberger, J. A. (2015). The impact of sample size and other factors when estimating multilevel logistic models. *The Journal of Experimental Education*, *84*, 3.

Skene, S. S., & Kenward, M. G. (2010a). The analysis of very small samples of repeated measurements I: An adjusted sandwich estimator. *Statistics in Medicine*, *29*, 2825–2837. doi:10.1002/sim.v29:27

Skene, S. S., & Kenward, M. G. (2010b). The analysis of very small samples of repeated measurements II: A modified Box correction. *Statistics in Medicine*, *29*, 2838–2856. doi:10.1002/sim.v29:27

Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*, 348–377. doi:10.3102/1076998609332752

Staggs, V. S. (2015). Comparison of naïve, Kenward–Roger, and parametric bootstrap interval approaches to small-sample inference in linear mixed models. *Communications in Statistics: Simulation and Computation*. Advance online publication. doi:10.1080/03610918.2015.1019002

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*, 748–761. doi:10.1111/ajps.2013.57.issue-3

Stenling, A., Ivarsson, A., Johnson, U., & Lindwall, M. (2015). Bayesian structural equation modeling in sport and exercise psychology. *Journal of Sport & Exercise Psychology*, *37*, 410–420. doi:10.1123/jsep.2014-0330

Swain, A. J. (1975). *Analysis of parametric structures for variance matrices* (Unpublished doctoral dissertation). Department of Statistics, University of Adelaide, Adelaide, Australia.

van de Schoot, R. (2016). *25 years of Bayes in psychology.* Paper presented at the 7th M*plus* Users' Meeting, Utrecht, The Netherlands. Retrieved from http://mplus.fss.uu.nl/wp-content/uploads/sites/24/2012/07/opening-review-short.pptx

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*. doi:10.3402/ejpt.v6.25216

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*, 842–860. doi:10.1111/cdev.12169

van de Schoot, R., Ryan, O., Winter, S., Zondervan-Zwijenburg, M. A. J., & Depaoli, S. (under review). *A systematic review of empirical Bayesian applications in psychology.*

van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*, 765–790. doi:10.1177/0013164403251027

Wanless, S. B., Rimm-Kaufman, S. E., Abry, T., Larsen, R. A., & Patton, C. L. (2015). Engagement in training as a mechanism to understanding fidelity of implementation of the responsive classroom approach. *Prevention Science*, *16*, 1107–1116. doi:10.1007/s11121-014-0519-6

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*, 115–148. doi:10.1207/s15327906mbr4001_5

Yuan, K. H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*, 767–774. doi:10.1080/01621459.1997.10474029

Yuan, K. H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*, 225–243. doi:10.2307/1165323

Zhang, Z., Hamagami, F., Wang, L. L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*, 374–383. doi:10.1177/0165025407077764

APPENDIX: SOFTWARE CODE FOR APPLIED EXAMPLE MODELS

Growth Model
Maximum Likelihood

```
VARIABLE:
NAMES = Y1 Y2 Y3 Y4 SEX;
USEVARIABLES Y1 Y2 Y3 Y4 SEX;

ANALYSIS:
ESTIMATOR=ML;

MODEL:
I S | Y1@0 Y2@2 Y3@4 Y4@6;
Y1-Y4;
[Y1-Y4@0];
[I S];
I;
S;
S WITH I;
I ON SEX;
S ON SEX;
```

Restricted Maximum Likelihood (SAS)

```
PROC MIXED DATA=PR;
MODEL Y=TIME|SEX/SOLUTION CL DDFM=KR ;
* DDFM= KR uses Kenward–Roger Correction;
RANDOM INT TIME/ SUB=PERSON TYPE=UN;
RUN;
```

Improper Inverse Wishart (M*plus* Default)

```
VARIABLE:
NAMES = Y1 Y2 Y3 Y4 SEX;
USEVARIABLES Y1 Y2 Y3 Y4 SEX;

ANALYSIS:
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
I S | Y1@0 Y2@2 Y3@4 Y4@6;
Y1-Y4;
[Y1-Y4@0];
[I S];
I;
S;
S WITH I;
I ON SEX;
S ON SEX;
```

Marginal Inverse Gammas

```
VARIABLE:
NAMES = Y1 Y2 Y3 Y4 SEX;
USEVARIABLES Y1 Y2 Y3 Y4 SEX;

ANALYSIS:
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
I S | Y1@0 Y2@2 Y3@4 Y4@6;
Y1-Y4;
[Y1-Y4@0];
```

```
[I S];
I (A);
S (B);
S WITH I (C);
I ON SEX;
S ON SEX;

MODEL PRIORS:
A~IG(0.01, 0.01); ! Set inverse gamma prior for intercept variance;
B~IG(0.01, 0.01); ! Set inverse gamma prior for slope variance;
C~N(0,10000); ! Set normal prior for intercept, slope covariance;
```

Weakly Informative Inverse Wishart

```
VARIABLE:
NAMES = Y1 Y2 Y3 Y4 SEX;
USEVARIABLES Y1 Y2 Y3 Y4 SEX;

ANALYSIS:
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
I S | Y1@0 Y2@2 Y3@4 Y4@6;
Y1-Y4;
[Y1-Y4@0];
[I S];
I (A);
S (B);
S WITH I (C);
I ON SEX;
S ON SEX;

MODEL PRIORS:
A~IW(14,7); ! Set inverse Wishart prior for intercept variance;
B~IW(.14,7); ! Set inverse Wishart prior for slope variance;
C~IW(0,7); ! Set normal prior for intercept, slope covariance;
```

Strongly Informative Inverse Wishart

```
VARIABLE:
NAMES = Y1 Y2 Y3 Y4 SEX;
USEVARIABLES Y1 Y2 Y3 Y4 SEX;

ANALYSIS:
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
I S | Y1@0 Y2@2 Y3@4 Y4@6;
Y1-Y4;
[Y1-Y4@0];
[I S];
I (A);
S (B);
S WITH I (C);
I ON SEX;
S ON SEX;

MODEL PRIORS:
A~IW(45,15); ! Set inverse Wishart prior for intercept variance;
B~IW(.45,15); ! Set inverse Wishart prior for slope variance;
C~IW(0,15); ! Set normal prior for intercept, slope covariance;
```

Multilevel Model
Maximum Likelihood

```
VARIABLE:
NAMES = PRE POST TRT ELL PREC TRTELL ID;
USEVARIABLES POST TRT ELL PREC TRTELL ID;
BETWEEN = TRT;
CLUSTER = ID;

ANALYSIS:
TYPE=TWOLEVEL RANDOM;
ESTIMATOR=ML;

MODEL:
%WITHIN%
POST;
%BETWEEN%
POST ON TRT ELL PREC TRTELL;
POST;
```

Restricted Maximum Likelihood (SAS)

```
PROC MIXED DATA=READ;
MODEL POST= ELL|TREAT PRE|TREAT /SOLUTION CL DDFM=KR;
* DDFM= KR uses Kenward–Roger correction;
RANDOM INT /SUB=SCHOOL;
RUN;
```

Improper Inverse Gamma (M*plus* Default)

```
VARIABLE:
NAMES = PRE POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
USEVARIABLES POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
BETWEEN = TRT;
CLUSTER = ID;

ANALYSIS:
TYPE=TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
POST ;
%BETWEEN%
POST ON TRT ELL PRE_C TRT_ELL PREC_TRT;
POST ;
```

Inverse Gamma (0.01, 0.01)

```
VARIABLE:
NAMES = PRE POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
USEVARIABLES POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
BETWEEN = TRT;
CLUSTER = ID;

ANALYSIS:
TYPE=TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
POST ;

%BETWEEN%
POST ON TRT ELL PRE_C TRT_ELL PREC_TRT;
POST (A);
```

```
MODEL PRIORS:
A~IG(0.01, 0.01); ! Set inverse gamma prior for between-cluster variance;
```

Weakly Informative Inverse Gamma

```
VARIABLE:
NAMES = PRE POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
USEVARIABLES POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
BETWEEN = TRT;
CLUSTER = ID;

ANALYSIS:
TYPE=TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
POST ;

%BETWEEN%
POST ON TRT ELL PRE_C TRT_ELL PREC_TRT;
POST (A);

MODEL PRIORS:
A~IG(5,30); ! Set inverse gamma prior for between-cluster variance;
```

Strongly Informative Inverse Gamma

```
VARIABLE:
NAMES = PRE POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
USEVARIABLES POST TRT ELL PRE_C TRT_ELL PREC_TRT ID;
BETWEEN = TRT;
CLUSTER = ID;

ANALYSIS:
TYPE=TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
POST ;

%BETWEEN%
POST ON TRT ELL PRE_C TRT_ELL PREC_TRT;
POST (A);

MODEL PRIORS:
A~IG(15,105); ! Set inverse gamma prior for between-cluster variance;
```

Meta-Analysis

Maximum Likelihood

```
VARIABLE:
NAMES = ID ES VAR WEEKS;
USEVARIABLES WEEKS Y X ;
WITHIN=Y X;
BETWEEN=WEEKS;
CLUSTER = ID;
DEFINE: Y= ES/SQRT(VAR); X=1/SQRT(VAR);
! Transform based on inverse of effect size sampling variance;

ANALYSIS: TYPE= TWOLEVEL RANDOM;
ESTIMATOR=ML;

MODEL:
%WITHIN%
```

```
[Y@0.0]; ! Mean at the within level constrained to 0;
Y@1; ! Residual variance constrained to 1;
THETA | Y ON X; !random slope;
%BETWEEN%
THETA on WEEKS;
[THETA]; ! average conditional effect size;
THETA ; ! between study variance;
```

Restricted Maximum Likelihood (SAS)

```
*** Create weight variable based on the inverse of the sampling variance*;
DATA HOX; SET HOX;
WV = 1/VARD;
RUN;

PROC MIXED DATA=HOX COVTEST;
MODEL ES = WEEKS /SOULTION CL DDFM=KR ;
* DDFM=KR uses Kenward–Roger correction;
RANDOM INT / SUB=ID ;
PARMS (.15) (1) / HOLD=2; *Constrains the residual variance to 1;
WEIGHT WV; *Weight calculated by inverse of sampling variance;
RUN;
```

Improper Inverse Gamma (M*plus* Default)

```
VARIABLE:
NAMES = ID ES VAR WEEKS;
USEVARIABLES WEEKS Y X ;
WITHIN=Y X;
BETWEEN=WEEKS;
CLUSTER = ID;
DEFINE: Y= ES/SQRT(VAR); X=1/SQRT(VAR);
! Transform based on inverse of effect size sampling variance;

ANALYSIS: TYPE= TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
[Y@0.0]; ! Mean at the within level constrained to 0;
Y@1; ! Residual variance constrained to 1;
THETA | Y ON X; !random slope;
%BETWEEN%
THETA on WEEKS;
[THETA]; ! average conditional effect size;
THETA ; ! between study variance;
```

Inverse Gamma (0.01, 0.01)

```
VARIABLE:
NAMES = ID ES VAR WEEKS;
USEVARIABLES WEEKS Y X ;
WITHIN=Y X;
BETWEEN=WEEKS;
CLUSTER = ID;
DEFINE: Y= ES/SQRT(VAR); X=1/SQRT(VAR);
! Transform based on inverse of effect size sampling variance;

ANALYSIS: TYPE= TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
[Y@0.0]; ! Mean at the within level constrained to 0;
```

```
Y@1; ! Residual variance constrained to 1;
THETA | Y ON X; !random slope;

%BETWEEN%
THETA ON WEEKS;
[THETA]; ! average conditional effect size;
THETA (A); ! between-study variance;
MODEL PRIORS:
A~ IG (0.01, 0.01); ! Set inverse gamma prior for between-study variance;
```

Weakly Informative Inverse Gamma

```
VARIABLE:
NAMES = ID ES VAR WEEKS;
USEVARIABLES WEEKS Y X ;
WITHIN=Y X;
BETWEEN=WEEKS;
CLUSTER = ID;
DEFINE: Y= ES/SQRT(VAR); X=1/SQRT(VAR);
! Transform based on inverse of effect size sampling variance;

ANALYSIS: TYPE= TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
[Y@0.0]; ! Mean at the within level constrained to 0;
Y@1; ! Residual variance constrained to 1;
THETA | Y ON X; !random slope;
%BETWEEN%

THETA ON WEEKS;
[THETA]; ! average conditional effect size;
THETA (A); ! between study variance;
MODEL PRIORS:
A~ IG (3,.07); ! Set inverse gamma prior for between-study variance;
```

Strongly Informative Inverse Gamma

```
VARIABLE:
NAMES = ID ES VAR WEEKS;
USEVARIABLES WEEKS Y X ;
WITHIN=Y X;
BETWEEN=WEEKS;
CLUSTER = ID;
DEFINE: Y= ES/SQRT(VAR); X=1/SQRT(VAR);
! Transform based on inverse of effect size sampling variance;

ANALYSIS: TYPE= TWOLEVEL RANDOM;
ESTIMATOR=BAYES;
BITERATIONS=100000 (50000);
! minimum of 50,000 iterations, maximum of 100,000;
BCONVERGENCE =.01; ! set more strict PSR convergence criteria of .01;

MODEL:
%WITHIN%
[Y@0.0]; ! Mean at the within level constrained to 0;
Y@1; ! Residual variance constrained to 1;
THETA | Y ON X; !random slope;

%BETWEEN%
THETA ON WEEKS;
[THETA]; ! average conditional effect size;
THETA (A); ! between study variance;
MODEL PRIORS:
A~ IG (12,.39); ! Set inverse gamma prior for between-study variance;
```