frontiers in Psychology

ORIGINAL RESEARCH
published: 25 January 2021
doi: 10.3389/fpsyg.2020.611267

Prior Specification for More Stable Bayesian Estimation of Multilevel Latent Variable Models in Small Samples: A Comparative Investigation of Two Different Approaches

Steffen Zitzmann[1]*, *Christoph Helm*[2] and *Martin Hecht*[3]

[1] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany, [2] Institute for the Management and Economics of Education, University of Teacher Education Zug, Zug, Switzerland, [3] Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

OPEN ACCESS

Edited by:
Christoph Koenig,
Goethe University Frankfurt, Germany

Reviewed by:
Esther Ulitzsch,
University of Kiel, Germany
Alexander Naumann,
Leibniz Institute for Research and Information in Education (DIPF), Germany

***Correspondence:**
Steffen Zitzmann
steffen.zitzmann@uni-tuebingen.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: 28 September 2020
Accepted: 22 October 2020
Published: 25 January 2021

Citation:
Zitzmann S, Helm C and Hecht M (2021) Prior Specification for More Stable Bayesian Estimation of Multilevel Latent Variable Models in Small Samples: A Comparative Investigation of Two Different Approaches. Front. Psychol. 11:611267. doi: 10.3389/fpsyg.2020.611267

Bayesian approaches for estimating multilevel latent variable models can be beneficial in small samples. Prior distributions can be used to overcome small sample problems, for example, when priors that increase the accuracy of estimation are chosen. This article discusses two different but not mutually exclusive approaches for specifying priors. Both approaches aim at stabilizing estimators in such a way that the Mean Squared Error (MSE) of the estimator of the between-group slope will be small. In the first approach, the MSE is decreased by specifying a slightly informative prior for the group-level variance of the predictor variable, whereas in the second approach, the decrease is achieved directly by using a slightly informative prior for the slope. Mathematical and graphical inspections suggest that both approaches can be effective for reducing the MSE in small samples, thus rendering them attractive in these situations. The article also discusses how these approaches can be implemented in M*plus*.

Keywords: Bayesian estimation, Markov chain Monte Carlo, multilevel modeling, structural equation modeling, small sample

As van de Schoot et al. (2017) pointed out, the number of applications of Bayesian approaches is growing quickly, mainly because software that is easy to use such as M*plus* (Muthén and Muthén, 2012) is providing Bayesian estimation as an option. Bayesian approaches can be beneficial in several respects, for example, by offering greater flexibility (e.g., Hamaker and Klugkist, 2011; Muthén and Asparouhov, 2012; Lüdtke et al., 2013) or fewer estimation problems (e.g., Hox et al., 2012; Depaoli and Clifton, 2015; Zitzmann et al., 2016), particularly when latent variable models are estimated. One major difference between Bayesian and traditional Maximum Likelihood (ML) estimation is that the former not only uses the information from the data at hand (i.e., the likelihood function) but combines it with additional information from what is called the prior distribution. Inferences are based on the result of this combination, that is, the posterior distribution. Scholars have advised researchers against the use of default priors in an automatic fashion and have encouraged them to specify priors on their own (e.g., McNeish, 2016; Smid et al., 2020). This may be an obstacle to some researchers. However, the prior can also be considered a feature of Bayesian estimation that can be used to improve estimation by choosing a favorable prior—a task that is particularly challenging but also particularly worth pursuing when the sample size is small.

The choice of prior has received a lot of attention in the methodological literature (e.g., Natarajan and Kass, 2000; Gelman, 2006; Chung et al., 2013), and scholars have made different suggestions about how priors can be specified in advantageous ways. Only recently, Smid et al. (2020) discussed how priors can be "thoughtfully" constructed on the basis of previous knowledge about the parameter of interest (e.g., on the basis of a previous study or a meta-analysis) in order to reduce small-sample bias. However, it has been argued that the variability of an estimator should not be ignored when evaluating the quality of a method (e.g., Greenland, 2000; Zitzmann et al., 2020), particularly when the sample size is small. Therefore, other suggestions for specifying the prior have been aimed at reducing the Mean Squared Error (MSE), which combines bias and variability: MSE $=$ bias^2 $+$ variability. One such approach was proposed by Zitzmann et al. (2015), who focused on the between-group slope in multilevel latent variable modeling. The authors suggested that researchers should suitably modify the estimator of the group-level variance of the predictor variable because this will result in a more stable (i.e., more accurate) estimator of the slope. To this end, a slightly informative prior is specified for the group-level variance of the predictor to pull the variance estimates away from zero (i.e., the indirect approach). By doing so, the estimates of the slope will not be too large, and the MSE of the estimator of the slope will be reduced. Notably, in contrast to Smid et al.'s (2020) suggestion, the prior does not need to match previous knowledge or the true value of the parameter in the population. Rather, an incorrect prior whose location deviates from the parameter in the population might reduce the MSE even more than a correct prior will. Zitzmann et al. (2015) found that for a standardized predictor (standardized at Level 1), a slightly informative inverse gamma prior for the group-level variance provided a somewhat biased but much more accurate (because it had a smaller MSE) estimator in small samples. Alternatively, to reduce the MSE of the estimator of the slope, one can specify a slightly informative prior directly for the slope in order to shrink the estimates and thereby ensure that they will not be too large (i.e., the direct approach).

In the present article, we mathematically work out the idea behind the direct approach for a simple multilevel latent variable model, and we contrast this approach with the indirect approach and with ML. Then, we graphically show the benefits that both approaches have over ML when the sample size is small. Finally, we discuss how these approaches can be implemented in M*plus*.

1. EXAMPLE MODEL

Before we go into detail, we present an example model that we will use later to illustrate the different strategies. The model was suggested by Lüdtke et al. (2008) as one way to yield (asymptotically) unbiased estimates of between-group slopes in contextual studies (see also Asparouhov and Muthén, 2019). To this end, on the group level in the model, the dependent variable Y is predicted by a latent variable (i.e., the latent group mean) instead of the unreliable manifest group mean of the predictor variable, which is why the model was named the *multilevel latent covariate model* (Lüdtke et al., 2008). Such latent group means have become part of many more complex multilevel structural equation models that are commonly applied in research practice (see Preacher et al., 2010, 2016, for overviews of such models).

More specifically, the individual-level predictor X splits into two uncorrelated and normally distributed parts: a between-group part X_b, which is the latent group mean, and a within-group part X_w, which is the individual deviation from X_b. For a person $i = 1, \ldots, n$ in group $j = 1, \ldots, J$, the decomposition thus reads:

$$X_{ij} = X_{b,j} + X_{w,ij} \tag{1}$$

$X_{b,j}$ is distributed around μ_X with variance τ_X^2, whereas the deviation $X_{w,ij}$ has variance σ_X^2. Hereafter, we will also call σ_X^2 and τ_X^2 the within-group and between-group variances of X, respectively.

Applying Raudenbush and Bryk's (2002) notation, the regression at the individual level reads:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_w X_{w,ij} + \varepsilon_{ij} \tag{2}$$

where β_w is the (fixed) within-group slope that describes the relationship between the predictor and the dependent variable at the individual level, and the ε_{ij} are normally distributed residuals. The residual variance is σ_Y^2. At the group level, the intercept β_{0j} is regressed on X_b:

$$\text{Level 2: } \beta_{0j} = \alpha + \beta_b X_{b,j} + \delta_j \tag{3}$$

where α is the overall intercept, and β_b is the between-group slope (i.e., the relationship between X and Y at the group level). The δ_j are normally distributed residuals with variance τ_Y^2.

Here, we focus on the between-group slope (β_b), which is of great interest in many applications of multilevel models (e.g., in the analysis of contextual effects). When the data are balanced (i.e., equal numbers of persons per group), the ML estimator of β_b is given by:

$$\hat{\beta}_b = \frac{\hat{\tau}_{YX}}{\hat{\tau}_X^2} \tag{4}$$

where $\hat{\tau}_X^2$ and $\hat{\tau}_{YX}$ are sample estimators of the group-level variance of X and the group-level covariance of X and Y, respectively.

Some statistical properties of the ML estimator in Equation 4 need to be discussed first to be able to compare this estimator with the Bayesian estimators later on. First, by using the first-order Taylor expansion (e.g., Casella and Berger, 2001; see also Grilli and Rampichini, 2011) and ignoring terms involving higher order factors such as $\frac{1}{n^2(n-1)}$ or $\frac{1}{n^2}$ for better readability, then the bias of $\hat{\beta}_b$ can roughly be approximated by:

$$\mathrm{E}\left(\hat{\beta}_b\right) - \beta_b \approx -\frac{2}{J-1}\left\{-\frac{2\,(1-\rho_X)}{n\rho_X} + \frac{1-\rho_X}{n\rho_X}\left(1+\frac{\beta_w}{\beta_b}\right)\right\}\beta_b \tag{5}$$

where $\rho_X = \frac{\tau_X^2}{\tau_X^2+\sigma_X^2}$ is the intraclass correlation (ICC) of X.[1] This equation could be simplified further, but we continue to use this expression here to emphasize formal similarities with the biases of the Bayesian estimators (see below). However, even in its current form, it is evident from Equation (5) that the bias critically depends on the sample size (because J occurs in the denominator) and that the bias is generally non-zero in small samples. However, if we let J become large, the bias diminishes because $\frac{1}{J-1}$ becomes very small—a property of the estimator that we refer to as "asymptotic unbiasedness." In a similar way, we can yield an approximation of the variability of $\tilde{\beta}_b$:

$$\mathrm{Var}\left(\hat{\beta}_b\right) \approx \frac{1}{J-1}\left\{\left[\frac{\rho_Y}{\rho_X} + \frac{1-\rho_X}{n\rho_X}\left(\frac{\rho_Y}{\rho_X} + \frac{1-\rho_Y}{1-\rho_X}\right)\right]\frac{\tau_Y^2+\sigma_Y^2}{\tau_X^2+\sigma_X^2} + \left[-1 - \frac{2\,(1-\rho_X)}{n\rho_X}\frac{\beta_w}{\beta_b}\right]\beta_b^2\right\} \tag{6}$$

where $\rho_Y = \frac{\tau_Y^2}{\tau_Y^2+\sigma_Y^2}$ is the ICC of Y. Similar to the bias, the variability depends on the sample size in such a way that the variability will be small when J is large. Because the MSE of $\hat{\beta}_b$ is the sum of the squared bias and the variability,

$$\mathrm{MSE}\left(\hat{\beta}_b\right) \approx \left[\mathrm{E}\left(\hat{\beta}_b\right) - \beta_b\right]^2 + \mathrm{Var}\left(\hat{\beta}_b\right) \tag{7}$$

this measure will be small as well. Taken together, the more information the data provide, the more the overall accuracy of the estimator improves.

Whereas the asymptotic properties are favorable, the ML estimator tends to be biased in small samples, and it has high variability and thus a large MSE in these situations (e.g., McNeish, 2017). This challenges the usefulness of the ML estimator when the sample size is small because the result from a single study might be highly inaccurate. Therefore, scholars have called for alternative estimators that are less variable and thus more accurate (i.e., they have a smaller MSE), although they might be more biased than ML. In the multilevel literature, such estimators have been suggested by Chung et al. (2013), Greenland (2000), Grilli and Rampichini (2011), and Zitzmann et al. (2015), for example. Next, we develop the direct strategy, and recap the indirect strategy of specifying the prior.

2. THE DIRECT STRATEGY

We refer to the first strategy as the *direct strategy* because the prior is specified directly for the between-group slope (β_b). To illustrate, we assume a normal prior, which can be formalized as:

$$\beta_b \sim \mathrm{N}\left(a, b\right) \tag{8}$$

which should be read as "β_b is normally distributed with mean a and variance b." However, for better interpretability, we employ another, more convenient parameterization. Instead of a and b, we use the terms β_0 and $\frac{\tau_Y^2}{\nu_0\tau_X^2}$:

$$\beta_b \sim \mathrm{N}\left(\beta_0, \frac{\tau_Y^2}{\nu_0\tau_X^2}\right) \tag{9}$$

As we will show, β_0 and ν_0 can be meaningfully interpreted.

One way of expressing the likelihood for the slope is:

$$\beta_b \sim \mathrm{N}\left(\hat{\beta}_b, \frac{\hat{\tau}_Y^2}{J\hat{\tau}_X^2}\right) \tag{10}$$

where $\hat{\tau}_Y^2$ and $\hat{\tau}_X^2$ are the sampling variances of τ_Y^2 and τ_X^2, respectively. If we combine the prior in Equation (9) with the likelihood, we obtain the following posterior:

$$\beta_b \sim \mathrm{N}\left(\frac{\nu_0}{\nu_0+J}\beta_0 + \frac{J}{\nu_0+J}\hat{\beta}_b, \frac{J}{\nu_0+J}\frac{\hat{\tau}_Y^2}{J\hat{\tau}_X^2}\right) \tag{11}$$

which is also a normal distribution. The mean of this distribution defines the Bayesian Expected A Posteriori (EAP) estimator, which is the standard choice for a point estimator in Bayesian estimation (Note that the Bayes module in M*plus* uses the median of the posterior). With $w = \frac{J}{\nu_0+J}$, this Bayesian estimator can also be expressed as:

$$\bar{\beta}_b = (1-w)\,\beta_0 + w\hat{\beta}_b \tag{12}$$

As can be seen from the equation, the estimator is simply the weighted average of the mean of the prior (β_0) and $\hat{\beta}_b$, which suggests straightforward interpretations for the parameters of the prior. One may think of β_0 as the *prior guess* for the between-group slope and ν_0 as the *prior sample size* (see also Hoff, 2009). These interpretations are substantiated by the observation that the larger ν_0, the smaller w, and the more the estimates shrink toward β_0. Less technically speaking, when we are more confident in β_0, the prior will gain more weight, and the posterior will shift to the mean of the prior. However, when we choose ν_0 to be very small, w will be close to 1, and $\bar{\beta}_b$ will be similar to $\hat{\beta}_b$, which justifies the view that the modified estimator includes the original ML estimator as a limiting case. Notice that the prior guess does not need to represent previous knowledge about β_b. Rather, it could be set to a value that is much smaller than what previous studies have suggested and also much smaller than the parameter in the population. However, such an "incorrect" prior guess might still be beneficial, particularly when the sample size is small.

To be able to compare the properties of the Bayesian estimator with the ML estimator and with the Bayesian estimator from the second strategy of specifying the prior, we again use the Taylor expansion, and we ignore terms involving higher order factors. A rough approximation of the bias of $\bar{\beta}_b$ is then given by:

$$\mathrm{E}\left(\bar{\beta}_b\right) - \beta_b \approx (1-w)\,\beta_0 + \left\{-(1-w) - \frac{2w}{J-1}\left[-\frac{2\,(1-\rho_X)}{n\rho_X} + \frac{1-\rho_X}{n\rho_X}\left(1 + \frac{\beta_w}{\beta_b}\right)\right]\right\}\beta_b \tag{13}$$

[1] The ICC quantifies the amount of the total variance that can be attributed to differences between the groups (e.g., Snijders and Bosker, 2012).

Similar to the ML estimator, $\bar{\beta}_b$ is generally biased when the sample size is small. However, the bias vanishes when J approaches infinity because w approaches 1, and $\frac{1}{J-1}$ approaches 0 (asymptotic unbiasedness). Moreover, if ν_0 is set to a value close to 0, the bias will become similar to the bias of $\hat{\beta}_b$.

The variability of $\bar{\beta}_b$ can be approximated as:

$$\mathrm{Var}\left(\bar{\beta}_b\right) \approx \frac{w^2}{J-1}\left\{\left[\frac{\rho_Y}{\rho_X}+\frac{1-\rho_X}{n\rho_X}\left(\frac{\rho_Y}{\rho_X}+\frac{1-\rho_Y}{1-\rho_X}\right)\right]\frac{\tau_Y^2+\sigma_Y^2}{\tau_X^2+\sigma_X^2} + \left[-1-\frac{2\left(1-\rho_X\right)}{n\rho_X}\frac{\beta_w}{\beta_b}\right]\beta_b^2\right\} \tag{14}$$

With a very large J, the variability becomes negligibly small, and the same holds for the MSE. However, the more interesting questions are: How does the MSE of $\bar{\beta}_b$ depend on the prior parameters (β_0, ν_0) when the sample size is small, and how must they be chosen such that the MSE will be smaller than the MSE of ML? Before we compare the different choices for (β_0, ν_0), we present another strategy for specifying the prior. Alternatively to specifying the prior directly for the between-group slope, one can also specify a prior for the group-level variance of the predictor, thereby also modifying the estimator of the slope. We call this strategy the *indirect strategy*.

3. THE INDIRECT STRATEGY

The principle that underlies the indirect strategy was discovered in the early years of Structural Equation Modeling (SEM), where models were fit on the basis of the variances and covariances of variables. One observation was that when the sample size was small, covariance matrices tended to be on the border of positive definiteness (e.g., a variance estimate close to 0, correlations close to −1 or 1; e.g., van Driel, 1978; Dijkstra, 1992; Kolenikov and Bollen, 2012). Hence, estimators of slope parameters tended to have high variability and thus also a large MSE. This led Yuan and Chan (2008) to develop the ridge technique to mitigate such problems as it modifies the estimator of the covariance matrix by adding a small value to the main diagonal (see also Yuan and Chan, 2016; Yang and Yuan, 2019). The main idea behind this technique can also be adapted for Bayesian estimation. Papers by Chung et al. (2013), Chung et al. (2015), or Zitzmann et al. (2015) are good examples of this. By means of simulation, Zitzmann et al. (2015) verified that specifying a slightly informative prior for the group-level variance of the predictor that pulls estimates of this variance slightly away from zero can increase the accuracy of the estimator of the between-group slope by reducing its MSE. Note that pulling the variance estimates away from zero corresponds to adding a value to these estimates. A formal argument for why such a prior reduces the MSE was only recently presented by Zitzmann et al. (2020). For reasons of completeness and comparability with the two previously presented estimators, we illustrate the strategy here once more, using the example model from above.

Rather than beginning with the assumption of a normal prior for the between-group slope, we begin with a gamma prior for the inverse of the group-level variance of the predictor variable (τ_X^2):

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(a, b\right) \tag{15}$$

where a and b are the parameters of the gamma distribution.[2] Equation (15) reads "τ_X^2 is inverse-gamma distributed." As for the normal prior in the previous section, we employ a reparameterization for better interpretability later on. If we set a to $\frac{\nu_0}{2}$ and b to $\frac{\nu_0\tau_0^2}{2}$, the prior reads:

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\tau_0^2}{2}\right) \tag{16}$$

where, as we will show, τ_0^2 and ν_0 have interpretations similar to those of the parameters of the (reparameterized) normal prior.

The likelihood for the inverse of the group-level variance can be written as:

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(\frac{J}{2}, \frac{J\hat{\tau}_X^2}{2}\right) \tag{17}$$

where $\hat{\tau}_X^2$ is the sample variance. Combined with the prior in Equation (16), we yield the inverse gamma posterior:

$$\frac{1}{\tau_X^2} \sim \mathrm{Gamma}\left(\frac{\nu_0+J}{2}, \frac{\nu_0\tau_0^2+J\hat{\tau}_X^2}{2}\right) \tag{18}$$

As Zitzmann et al. (2020) showed in their Appendix C, the mean of this distribution can be approximated as:

$$\bar{\tau}_X^2 \approx (1-w)\,\tau_0^2 + w\hat{\tau}_X^2 \tag{19}$$

where $w = \frac{J}{\nu_0+J}$. This equation defines the Bayesian EAP estimator of τ_X^2. It is interesting to note that the equation resembles Equation (12). The right-hand side of the equation is also a weighted average, and τ_0^2 and ν_0 can be thought of as the prior guess and the prior sample size, respectively (see Hoff, 2009; Lüdtke et al., 2018; Zitzmann et al., 2020).

Adding a prior for τ_X^2 also has consequences for the estimator of the between-group slope (β_b). Replacing the denominator in Equation (4) ($\hat{\tau}_X^2$) with $\bar{\tau}_X^2$ results in:

$$\tilde{\beta}_b = \frac{\hat{\tau}_{YX}}{(1-w)\,\tau_0^2 + w\hat{\tau}_X^2} \tag{20}$$

This new estimator is indicated by a tilde (∼) in order to better differentiate it from the ML estimator and from the Bayesian estimator that results from the direct strategy of specifying the prior (Equation 12).

[2]The inverse of a variance is sometimes also referred to as the precision in the statistical literature (e.g., Hoff, 2009).

To derive some properties of $\tilde{\beta}_b$, we apply exactly the same reasoning that led to the respective properties of $\hat{\beta}_b$ and $\bar{\beta}_b$. Accordingly, the bias of $\tilde{\beta}_b$ is roughly:

$$\mathrm{E}\left(\tilde{\beta}_b\right) - \beta_b \approx (f-1)\,\beta_b - \frac{2wf^2}{J-1}\left\{-wf\left[1+\frac{2\,(1-\rho_X)}{n\rho_X}\right]+1 + \frac{1-\rho_X}{n\rho_X}\left(1+\frac{\beta_w}{\beta_b}\right)\right\}\beta_b \tag{21}$$

where f is used as an abbreviation for the ratio $\frac{\tau_X^2}{(1-w)\tau_0^2+w\tau_X^2}$.[3] Notice that the equation implies that $\tilde{\beta}_b$ is generally biased when the sample size is finite, whereas the bias diminishes when J approaches infinity (asymptotic unbiasedness). Moreover, the bias becomes similar to the biases of $\bar{\beta}_b$ and $\hat{\beta}_b$ when we let ν_0 become very small.

The variability of $\tilde{\beta}_b$ is:

$$\mathrm{Var}\left(\tilde{\beta}_b\right) \approx \frac{f^2}{J-1}\left\{\left[\frac{\rho_Y}{\rho_X}+\frac{1-\rho_X}{n\rho_X}\left(\frac{\rho_Y}{\rho_X}+\frac{1-\rho_Y}{1-\rho_X}\right)\right]\frac{\tau_Y^2+\sigma_Y^2}{\tau_X^2+\sigma_X^2} + \left[2wf\left(wf\left(1+\frac{2\,(1-\rho_X)}{n\rho_X}\right) - 2\left(1+\frac{1-\rho_X}{n\rho_X}\left(1+\frac{\beta_w}{\beta_b}\right)\right)\right)+1 + \frac{2\,(1-\rho_X)}{n\rho_X}\frac{\beta_w}{\beta_b}\right]\beta_b^2\right\} \tag{22}$$

Similar to the previous equation, we can easily infer that when J is large, the variability will be small and, thus, the MSE, which combines bias and variability, will be small as well—an observation that once again demonstrates the role of the sample size in determining the accuracy of estimations. However, as mentioned above, it is much more interesting to ask how the prior parameters τ_0^2 and ν_0 must be chosen such that the MSE will be reduced in comparison with ML in small samples.

4. COMPARING THE MSEs IN SMALL SAMPLES

In this section, we investigate the MSE of the different strategies for specifying priors in small samples for different choices of the prior parameters, using the example model from above to simulate data that are typical in psychology. Because it is difficult to infer from the equations how the MSEs compare with each other, they were plotted against the sample size to allow for graphical comparisons.

In accordance with Lüdtke et al. (2008), we considered the case of standardized variables (standardized at Level 1), and we assumed that the between-group slope (β_b) was 0.7 in the population. Moreover, we set the number of persons per group (n) to 5, which is not uncommon in many subdisciplines of psychology, including organizational, personality, and social psychology. The ICC of the predictor was 0.1, which could be considered small- to medium-sized compared with typical ICCs (Snijders and Bosker, 2012; Zitzmann et al., 2015). The sample size at the group level (J) was varied from 20 to 60 groups because these numbers represent small sample sizes (e.g., Hox et al., 2012; see also Hox et al., 2010) and the aim was to compare the estimators in these situations.

Figure 1 depicts a normalized version of the MSE, the Root Mean Squared Error (RMSE), for five different estimators of the slope. The first estimator in the figure is the ML estimator (solid black line). The second estimator (blue dashed line) is the Bayesian estimator that results when the direct strategy is combined with a correct prior for β_b (i.e., the prior guess, β_0, equals the parameter in the population). Because β_b was 0.7 in the population, a correct prior for β_b was specified by setting β_0 equal to this value. The third estimator (blue dotted line) also resulted from the direct strategy. However, β_0 was set to 0 (and thus well below 0.7) in order to shrink estimates that were too large toward zero. The fourth estimator (red dashed line) resulted from the indirect strategy with a correct prior for the group-level variance of the predictor (τ_X^2). The prior guess (τ_0^2) was set to 0.1, which was the value of τ_X^2 in the population.[4] The fifth estimator (red dotted line) resulted from the indirect strategy as well. However, β_0 was set to 1, which was above the parameter in the population. Thus, estimates of the variance were pulled away from zero, and, therefore, the estimates of the slope were shrunken. The three different panels of **Figure 1** show the RMSEs for different values of ν_0: 0.1 (upper left), 1.0 (upper right), and 5.0 (lower left). The first two values can be considered choices that are only slightly informative, whereas the latter is more informative and was used here to illustrate what happens to the RMSE when the priors become more informative.

As can be seen in the **Figure 1**, the different estimators tended to provide different RMSEs. The RMSE was largest for the ML estimator, whereas the RMSE was reduced when a Bayesian estimator was used. The reduction was particularly pronounced when J was very small. In addition and more important, the extent of the reduction also depended on the strategy for specifying the prior and the choices for the prior parameters. Although the direct strategy reduced the RMSE overall, the RMSE was slightly smaller when this strategy was combined with an incorrect prior (i.e., $\beta_0 = 0$) than with a correct prior (i.e., $\beta_0 = 0.7$). Moreover, the choice of a larger ν_0 was associated with a smaller RMSE. However, the smallest RMSEs emerged when the indirect strategy was used with an incorrect prior (i.e., $\tau_0^2 = 1$, which was also the upper bound of τ_X^2 due to standardization). With a larger value of $\nu_0 = 1$, the RMSE was reduced relative to a ν_0 of 0.1. However, setting ν_0 to 5 did not yield an RMSE that was even smaller. Rather, the RMSE was slightly larger than with a ν_0 of 1 because the bias induced by the prior outweighed the variability in the computation of the RMSE. Additional results

[3] We would like to state that we recognized a typo in the bias formula of Zitzmann et al.'s (2020) original publication. There should be a minus sign in front of the first term of the curly-bracketed expression. Equation (21) presents the corrected formula. However, the numerical results on which Figure 3 in Zitzmann et al. (2020) was based were not affected by the typo because these results were generated from formulas that were correct and also provided even more precise approximations than the ones presented in the article (because terms with higher order factors were not omitted).

[4] Because of the standardization, τ_X^2 is equal to the value of the ICC.

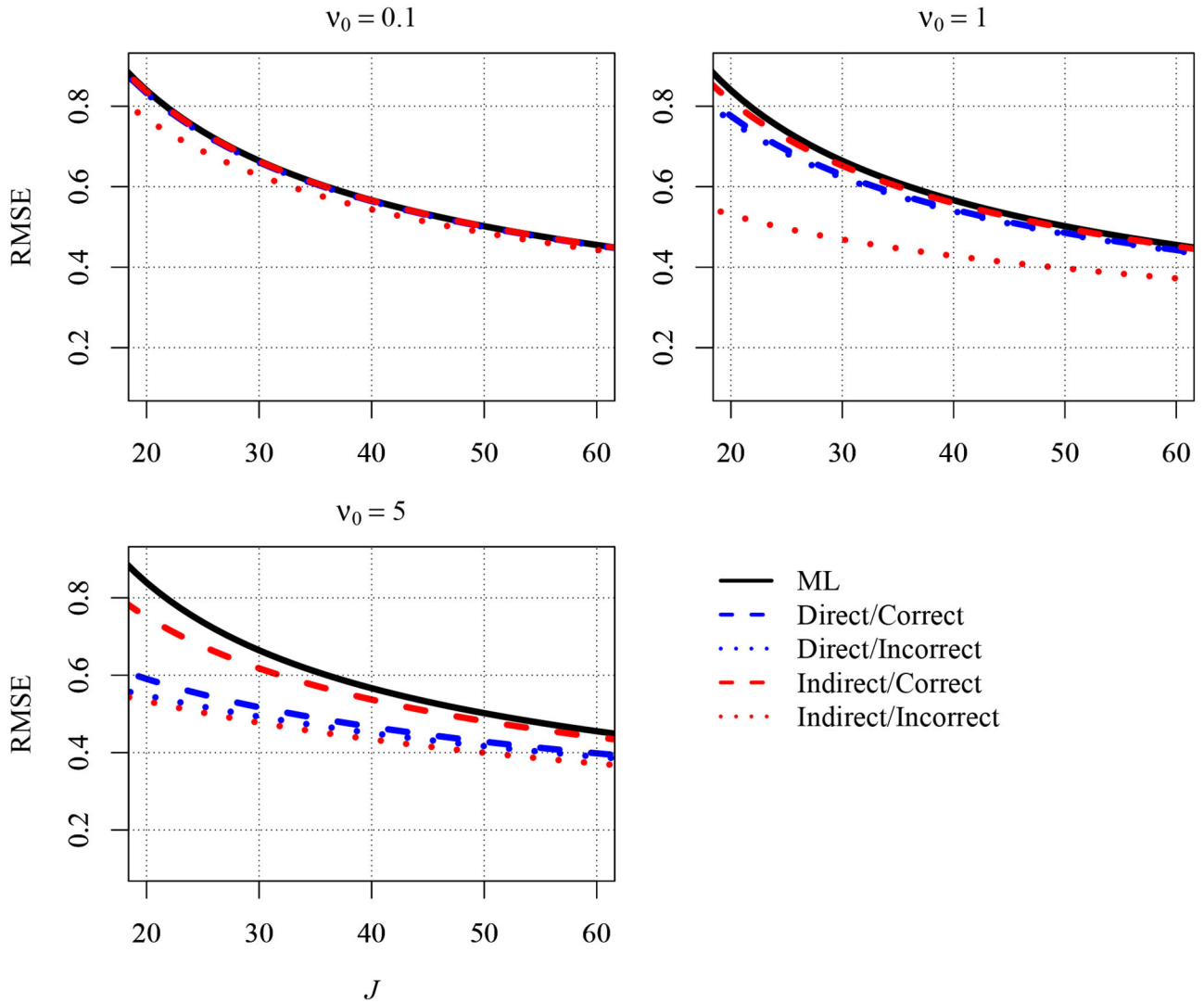

FIGURE 1 | The analytically derived Root Mean Squared Error (RMSE) in estimating the between-group slope for the direct and the indirect approach as a function of the sample size at the group level (J) and the prior distribution. Results are shown for $n = 5$ persons per group and an intraclass correlation of ICC = 0.1. ML, maximum likelihood; direct, the prior was specified directly for the between-group slope; indirect, the prior was specified for the group-level variance of the predictor variable; correct, correct prior (i.e., the prior guess equaled the value of the parameter in the population); incorrect, incorrect prior (i.e., the prior guess deviated from the parameter in the population); ν_0, prior sample size.

are presented in the **Appendix**. **Figure A1** shows the RMSEs of the different estimators for a larger number of 10 persons per group, whereas **Figure A2** shows the RMSEs for a higher ICC of .2. Although the RMSEs were smaller in **Figures A1**, **A2** compared with **Figure 1**, the big picture was similar overall: The different estimators provided different RMSEs. All Bayesian estimators provided smaller RMSEs than the ML estimator in very small samples except the indirect strategy with an incorrect informative prior.

To sum up, both strategies for specifying the prior offer attractive ways to obtain more accurate estimators of the between-group slope in small samples when used with slightly informative priors. Especially when no previous knowledge exists about the parameters, the choice of a relatively small prior guess for the between-group slope or a relatively large prior guess for the group-level variance of the predictor could be useful when these choices are combined with a small ν_0 in the low one-digit range. Although somewhat biased, the resulting Bayesian estimators of the slope were found to be more accurate than ML when the sample size was small.

5. DISCUSSION

It has been argued that Bayesian approaches can be beneficial when the sample size is small because prior distributions can be used to increase estimation accuracy. In the present article,

we focused on the between-group slope because this parameter is often of interest in multilevel latent variable modeling. Two approaches for specifying priors can be distinguished, both of which are aimed at reducing the MSE of the estimator of the between-group slope: In the first approach, a slightly informative prior is specified directly for the slope, whereas in the indirect approach, the MSE is reduced by using a slightly informative prior for the group-level variance of the predictor variable. In the present article, we worked out the former approach mathematically and compared it with the indirect approach and with ML. Graphical inspections suggested that both approaches can be very effective in reducing the MSE compared with ML in small samples, rendering them attractive for researchers. We would like to add that these approaches are not mutually exclusive and that researchers can also apply them simultaneously by specifying slightly informative priors for the slope as well as for the group-level variance of the predictor variable. To provide initial information about how such a simultaneous application of the two approaches performs, we conducted an additional simulation study with 20 to 60 groups, 5 persons per group, and an ICC of the predictor variable of 0.1. **Figure 2** depicts the RMSE for five different estimators of the slope. The first estimator is the ML estimator (solid black line). The second estimator (blue dashed line) is the Bayesian estimator that resulted when the direct strategy and the indirect strategy were simultaneously applied and combined with correct priors for the between-group slope and the group-level variance of the predictor, respectively. The third estimator (blue dotted

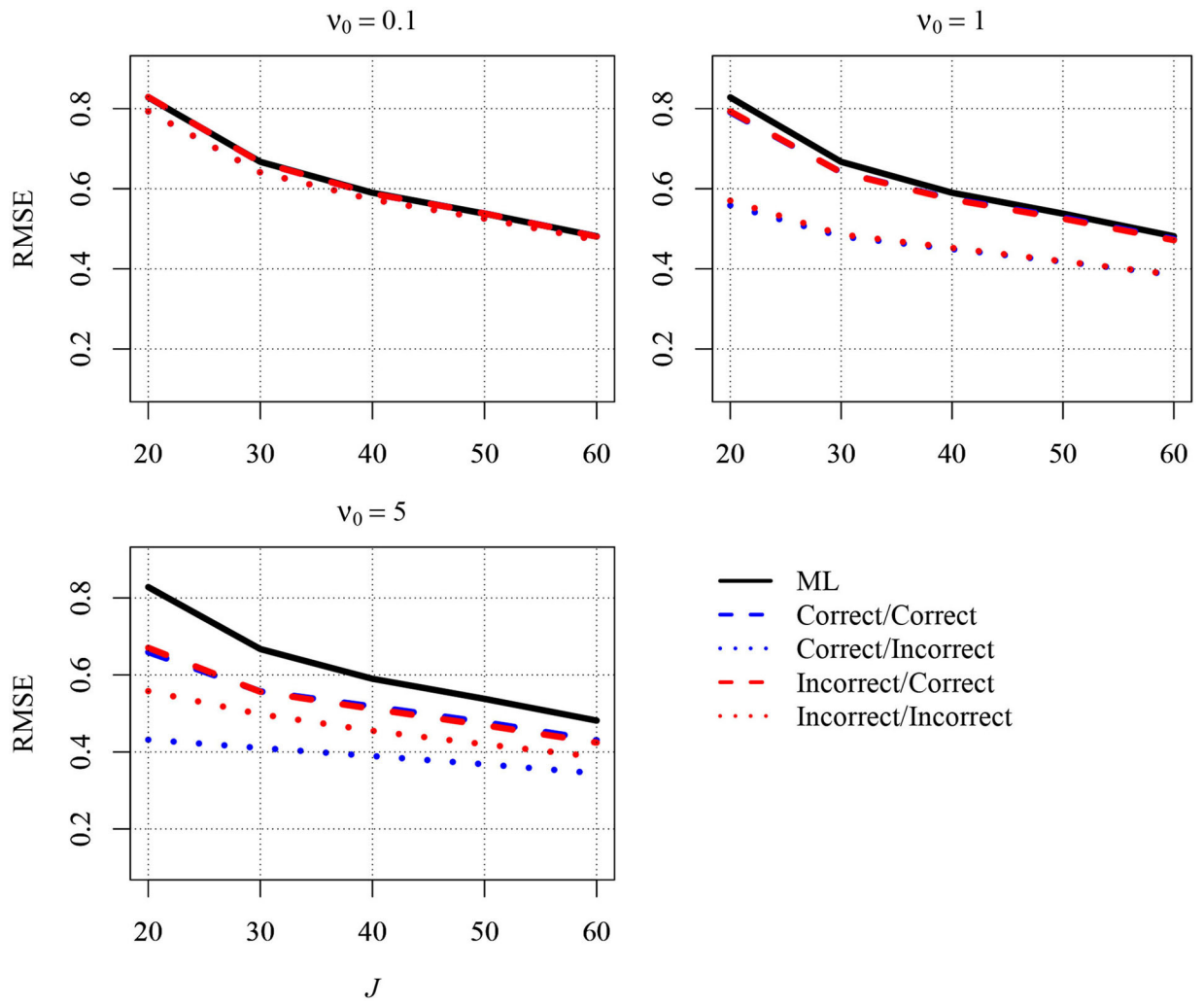

FIGURE 2 | The simulated Root Mean Squared Error (RMSE) in estimating the between-group slope for the combined approach as a function of the sample size at the group level (*J*) and the prior distribution. Results are shown for $n = 5$ persons per group and an intraclass correlation of ICC = 0.1. ML, maximum likelihood; correct/correct, correct priors (i.e., the prior guesses equaled the values of the parameters in the population) were specified for the between-group slope and the group-level variance of the predictor variable; correct/incorrect, a correct prior was specified for the between-group slope, and an incorrect prior (i.e., the prior guess deviated from the parameter in the population) was specified for the group-level variance of the predictor variable; incorrect/correct, an incorrect prior was specified for the between-group slope, and a correct prior was specified for the group-level variance of the predictor variable; incorrect/incorrect, incorrect priors were specified for the between-group slope and the group-level variance of the predictor variable; ν_0, prior sample size.

line) also resulted from combining the two strategies. However, whereas the direct strategy was combined with a correct prior, the indirect strategy was combined with an incorrect prior. The fourth estimator (red dashed line) resulted from simultaneously applying the direct strategy with an incorrect prior and the indirect strategy with a correct prior. The fifth estimator (red dotted line) resulted from the simultaneous application of the two strategies as well. However, both strategies were combined with incorrect priors. The three different panels of the figure show the RMSEs for different values of the prior sample size. Again, the RMSE was largest for the ML estimator, whereas the RMSE was reduced when a Bayesian estimator was used, particularly when this estimator was combined with slightly informative priors and the sample size was small. Thus, the overall finding from this simulation confirmed that a simultaneous application of the two approaches (i.e., specifying slightly informative priors for the slope as well as for the group-level variance of the predictor variable) can also be beneficial. However, because the consequences of such a use could not be studied exhaustively here, it would be interesting to conduct a more thorough simulation on this topic in future research.

Although our findings were generally favorable and could be considered a successful "proof of concept," a word of caution is nevertheless needed. Our demonstrations were very limited. For example, the specific conditions we studied do not completely reflect real data. Future research should consider a wider range of conditions for more conclusive findings. Moreover, the example model we used was overly simple. Realistic models typically involve more than one predictor and also multiple indicators per construct. However, one can derive the Bayesian estimators analogously in this more general multivariate case. Zitzmann (2018) even showed that in a multilevel SEM with two latent predictors with three indicators each, a slightly informative inverse Wishart prior for the covariance matrix of the predictors led to more accurate estimators of the between-group slopes, particularly when the samples size was small. Finally, the MSEs of the estimators we derived were only rough approximations. These approximations can nevertheless be useful for deriving hypotheses about which prior works well under which condition.

Before we come to M*plus*, we wish to acknowledge that parameter stabilization does not require Bayesian estimation. In fact, the idea of using slightly informative priors is similar to using techniques from the frequentist framework (Hastie et al., 2009). For example, the weighting parameter (w) of the Bayesian estimator in Equation (12) has an effect similar to that achieved by the penalty in regularized SEM (e.g., Jacobucci et al., 2016), and the weighting parameter in Equation (19) corresponds with the tuning parameter in ridge generalized least squares (e.g., Yuan and Chan, 2016) and regularized consistent partial least squares estimation (e.g., Jung and Park, 2018). Despite the existence of these methods, we employed Bayesian estimation here for reasons of convenience and because this type of estimation is an option in M*plus*, which is the software that many researchers use to fit multilevel latent variable models.

M*plus* does not use Bayesian estimation as the default, and users must request it by setting ESTIMATOR to BAYES. Next, to yield a more accurate estimator of the between-group slope by using a slightly informative prior for this parameter, users must specify such a prior manually. In M*plus*, normal priors are parameterized as in Equation (8), where a is the mean and b is the variance. Thus, to specify a normal prior with the prior guess (β_0) and the prior sample size (ν_0) equaling 0 and 1, respectively, users must compute a and b first. Given $a = \beta_0$ and $b = \frac{\tau_Y^2}{\nu_0 \tau_X^2}$, we yield an a of 0 and a b of $\frac{\tau_Y^2}{\tau_X^2}$. Because τ_Y^2 and τ_X^2 are unknown, they need to be replaced with, for example, their sample estimates. Assuming that these estimates are $\hat{\tau}_Y^2 = 0.15$ and $\hat{\tau}_X^2 = 0.1$, then the prior is specified by the following line of code:

```
MODEL PRIORS:
   Name of slope ~ N(0, 1.5);
```

Our findings suggest that this prior increases the accuracy of estimation in small samples. Choosing an even smaller value for b can also be useful in these situations. Alternatively, one could also specify a slightly informative prior for the group-level variance of the predictor. To be able to do this, users must compute the parameters a and b in Equation (15) because M*plus* uses this parameterization of the inverse gamma prior. Setting both τ_0^2 and ν_0 to 1 results in $a = b = \frac{1}{2}$, using $a = \frac{\nu_0}{2}$ and $b = \frac{\nu_0 \tau_0^2}{2}$. The following code line implements the prior with M*plus*:

```
MODEL PRIORS:
   Name of variance ~ IG(0.5, 0.5);
```

For a standardized predictor, this prior is quite effective when the sample size is small. Specifying somewhat larger values (e.g., by setting $\nu_0 = 2$) might increase estimation accuracy even further (Depaoli and Clifton, 2015).

To conclude, we worked out and discussed Bayesian approaches that perform better than ML in small samples, and we offered some practical guidance on how to implement these approaches with M*plus*. We hope that this article will help researchers in the field of psychology move beyond using Bayesian estimation as "just another estimator" and will help them make choices that are beneficial when their aim is to fit multilevel latent variable models and the sample size is small.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SZ: writing, mathematical derivations, and graphic design. CH: writing. MH: writing and lead. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We acknowledge support by Open Access Publishing Fund of University of Tübingen.

REFERENCES

Asparouhov, T., and Muthén, B. O. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Struct. Equat. Model.* 26, 119–142. doi: 10.1080/10705511.2018.1511375

Casella, G., and Berger, R. L. (2001). *Statistical Inference, 2nd Edn*. Pacific Grove, CA: Duxbury Press.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *J. Educ. Behav. Stat*. 40, 136–157. doi: 10.3102/1076998615570945

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* 78, 685–709. doi: 10.1007/s11336-013-9328-2

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

Dijkstra, T. K. (1992). On statistical inference with parameter estimates on the boundary of the parameter space. *Brit. J. Math. Stat. Psychol.* 45, 289–309. doi: 10.1111/j.2044-8317.1992.tb00994.x

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

Greenland, S. (2000). Principles of multilevel modelling. *Int. J. Epidemiol.* 29, 158–167. doi: 10.1093/ije/29.1.158

Grilli, L., and Rampichini, C. (2011). The role of sample cluster means in multilevel models. *Methodology* 7, 121–133. doi: 10.1027/1614-2241/a000030

Hamaker, E. L., and Klugkist, I. (2011). "Bayesian estimation of multilevel models," in *Handbook of Advanced Multilevel Analysis*, eds J. J. Hox and J. K. Roberts (New York, NY: Routledge), 137–161.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edn*. New York, NY: Springer.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. New York, NY: Springer. doi: 10.1007/978-0-387-92407-6

Hox, J. J., Maas, C. J. M., and Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Stat. Neerland*. 64, 157–170. doi: 10.1111/j.1467-9574.2009.00445.x

Hox, J. J., van de Schoot, R., and Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Surv. Res. Methods* 6, 87–93. doi: 10.18148/srm/2012.v6i2.5033

Jacobucci, R., Grimm, K. J., and McArdle, J. J. (2016). Regularized structural equation modeling. *Struct. Equat. Model.* 23, 555–566. doi: 10.1080/10705511.2016.1154793

Jung, S., and Park, J. (2018). Consistent partial least squares path modeling via regularization. *Front. Psychol.* 9:174. doi: 10.3389/fpsyg.2018.00174

Kolenikov, S., and Bollen, K. A. (2012). Testing negative error variances: is a Heywood case a symptom of misspecification? *Sociol. Methods Res.* 41, 124–167. doi: 10.1177/0049124112442138

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., and Muthén, B. O. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychol. Methods* 13, 203–229. doi: 10.1037/a0012869

Lüdtke, O., Robitzsch, A., Kenny, A., and Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychol. Methods* 18, 101–119. doi: 10.1037/a0029252

Lüdtke, O., Robitzsch, A., and Wagner, J. (2018). More stable estimation of the STARTS model: a Bayesian approach using Markov chain Monte Carlo techniques. *Psychol. Methods* 23, 570–593. doi: 10.1037/met0000155

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Struct. Equat. Model.* 23, 750–773. doi: 10.1080/10705511.2016.1186549

McNeish, D. (2017). Small sample methods for multilevel modeling: a colloquial elucidation of REML and the Kenward-Roger correction. *Multivar. Behav. Res.* 5, 661–670. doi: 10.1080/00273171.2017.1344538

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

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén Muthén.

Natarajan, R., and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *J. Am. Stat. Assoc.* 95, 227–237. doi: 10.1080/01621459.2000.10473916

Preacher, K. J., Zhang, Z., and Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychol. Methods* 21, 189–205. doi: 10.1037/met0000052

Preacher, K. J., Zyphur, M. J., and Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychol. Methods* 15, 209–233. doi: 10.1037/a0020141

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods. Advanced Quantitative Techniques in the Social Sciences*, 2nd Edn. Thousand Oaks, CA: Sage.

Smid, S. C., McNeish, D., Miočević, M., and van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equat. Model.* 27, 131–161. doi: 10.1080/10705511.2019.1577140

Snijders, T. A. B., and Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd Edn. Los Angeles, CA: Sage.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychol. Methods* 22, 217–239. doi: 10.1037/met0000100

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika* 43, 225–243. doi: 10.1007/BF02293865

Yang, M., and Yuan, K.-H. (2019). Optimizing ridge generalized least squares for structural equation modeling. *Struct. Equat. Model.* 26, 24–38. doi: 10.1080/10705511.2018.1479853

Yuan, K.-H., and Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Comput. Stat. Data Anal.* 52, 4842–4858. doi: 10.1016/j.csda.2008.03.030

Yuan, K.-H., and Chan, W. (2016). Structural equation modeling with unknown population distributions: ridge generalized least squares. *Struct. Equat. Model.* 23, 163–179,. doi: 10.1080/10705511.2015.1077335

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivar. Behav. Res.* 53, 612–632. doi: 10.1080/00273171.2018.1469086

Zitzmann, S., Lüdtke, O., and Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivar. Behav. Res.* 50, 688–705. doi: 10.1080/00273171.2015.1090899

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Hecht, M. (2020). On the performance of Bayesian approaches in small samples: a comment on Smid, McNeish, Miočević, and van de Schoot (2020). *Struct. Equat. Model.* doi: 10.1080/10705511.2020.1752216. [Epub ahead of print].

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Struct. Equat. Model.* 23, 661–679. doi: 10.1080/10705511.2016.1207179

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Zitzmann, Helm and Hecht. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

FURTHER RESULTS

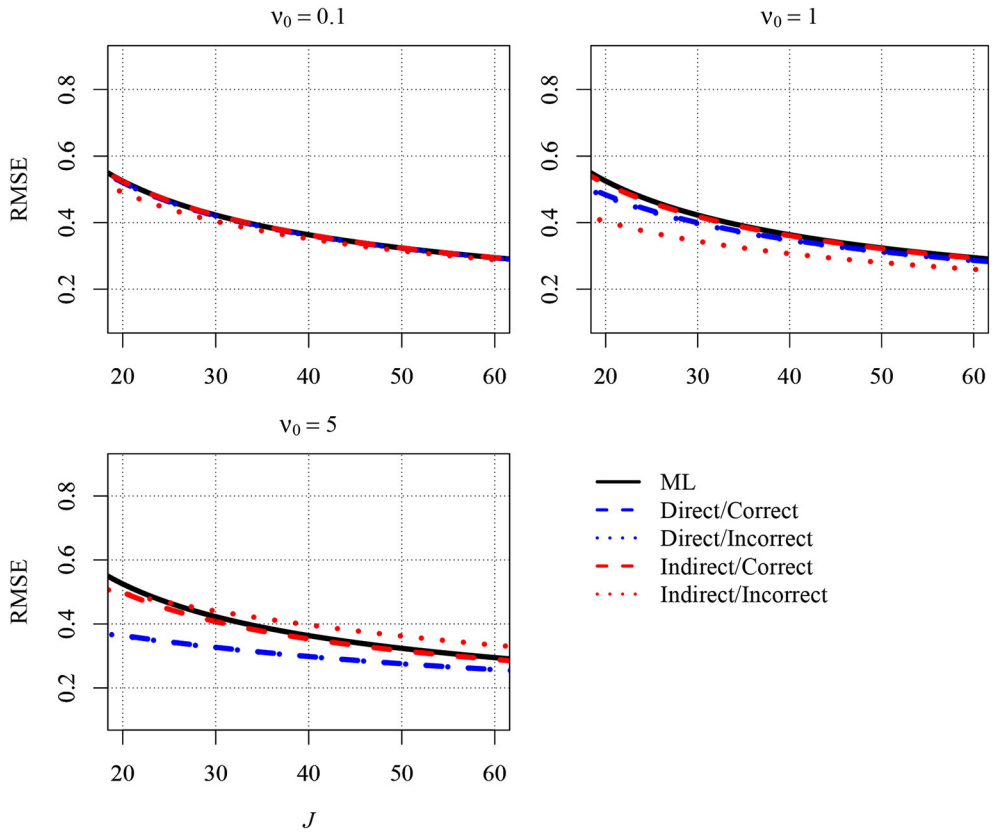

FIGURE A1 | The analytically derived Root Mean Squared Error (RMSE) in estimating the between-group slope for the direct and the indirect approach as a function of the sample size at the group level (J) and the prior distribution. Results are shown for $n = 10$ persons per group and an intraclass correlation of ICC = 0.1. ML, maximum likelihood; direct, the prior was specified directly for the between-group slope; indirect, the prior was specified for the group-level variance of the predictor variable; correct, correct prior (i.e., the prior guess equaled the value of the parameter in the population); incorrect, incorrect prior (i.e., the prior guess deviated from the parameter in the population); ν_0, prior sample size.

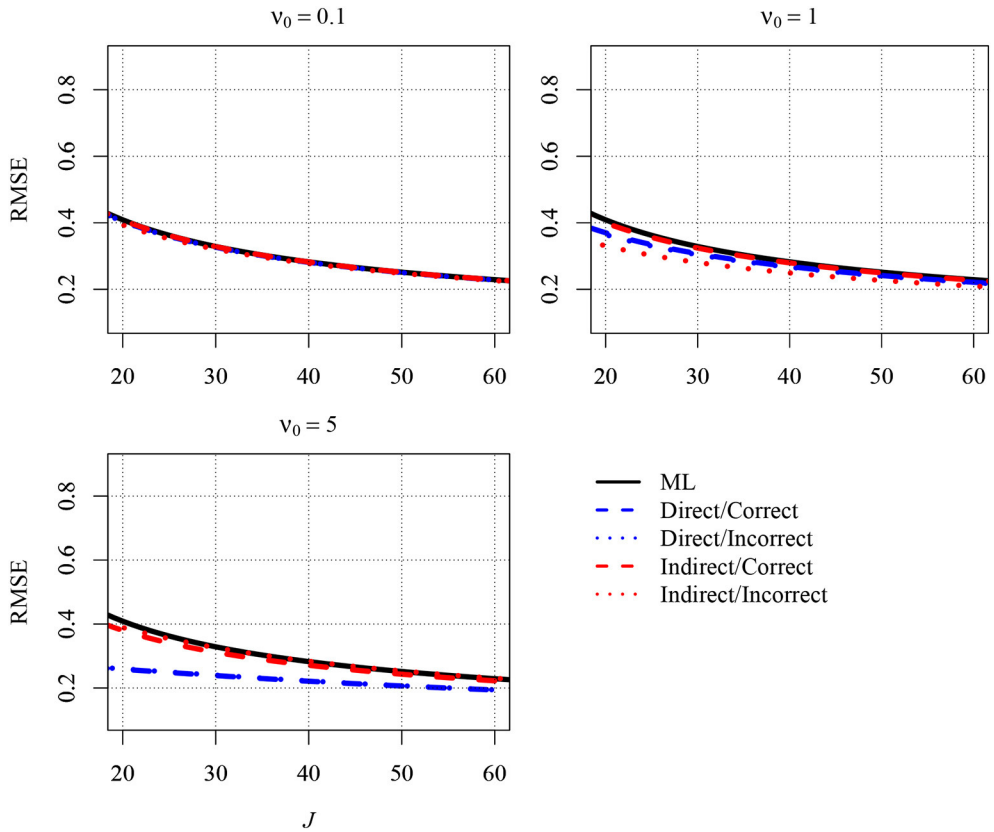

FIGURE A2 | The analytically derived Root Mean Squared Error (RMSE) in estimating the between-group slope for the direct and the indirect approach as a function of the sample size at the group level (J) and the prior distribution. Results are shown for a small number of $n = 5$ persons per group and an intraclass correlation of ICC = 0.2. ML, maximum likelihood; direct, the prior was specified directly for the between-group slope; indirect, the prior was specified for the group-level variance of the predictor variable; correct, correct prior (i.e., the prior guess equaled the value of the parameter in the population); incorrect, incorrect prior (i.e., the prior guess deviated from the parameter in the population); ν_0, prior sample size.