Psychological Test and Assessment Modeling, Volume 60, 2018 (2), 265-285

How many classes are needed to assess effects of instructional quality? A Monte Carlo simulation of the performance of frequentist and Bayesian multilevel latent contextual models

Christoph Helm[1]

Abstract

This study addresses the sample size question for multilevel latent contextual models (MLCM), which are commonly used in educational science to assess the effects of instructional quality. In terms of MLCM, only few studies have investigated whether the Bayesian toolbox helps to overcome small-sample issues. The main goal was to investigate the performance of maximum likelihood versus non, weakly, and highly informative Bayesian estimation techniques under small-sample conditions. We assumed that incorporation of prior information derived from TIMSS data would help to produce reasonable results with small samples. As expected, our results showed that the Bayesian approaches outperformed ML estimation under all conditions when informative priors were used, as these yield almost unbiased and highly accurate estimates even under unfavourable conditions (small number of level-2 groups and small group size). The study results are discussed in the light of published findings. Implications for applied educational research are derived.

Keywords: Bayesian statistics, multilevel modeling, small sample, TIMSS, informative prior distributions

[1] *Correspondence concerning this article should be addressed to:* Christoph Helm, Linz School of Education, Johannes Kepler University of Linz, Altenberger Straße 69, 4040 Linz-Auhof, Austria. email: christoph.helm@jku.at

Aside from ability/competence testing, the assessment of instructional features (e.g., instructional quality) and their effects on student outcomes represent a central challenge to educational researchers. Modelling instructional features (using student data) and contextual effects is not only theoretically and methodologically complex (e.g., Marsh, Lüdtke, Robitzsch, Trautwein, Asparouhov, Muthén, & Nagengast, 2009; Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, & Muthén, 2008), but also requires a relatively large number of clusters/groups (i.e., school classes to be observed). The latter is particularly true if the estimation approaches used are based on asymptotic theory (e.g., full maximum likelihood, FML). Kreft (1996) suggested using 30 clusters (of size 30 each). Maas and Hox (2005) recommend at least 50. However, a review by Dedrick, Ferron, ... Lee (2009) showed that only 21 out of 99 multilevel studies (identified in 13 journals from the fields of education, psychology, and sociology) met the 30/30 Kreft guideline. According to McNeish and Stapleton (2016a), this finding suggests that researchers may not have the resources to obtain adequately large samples. Clearly, small samples not only cause biased estimates (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011), but also lead to serious convergence problems (see Zitzmann et al., 2015). Estimation approaches are therefore needed that allow applied researchers to overcome this unsatisfactory situation. A review by McNeish and Stapleton (2016a) identified growth in the body of literature on this issue in the last 10 years. In addition to the use of small-sample corrections to the ML estimator (e.g., Kenward-Roger correction), the Bayesian approach has recently been discussed more extensively (Depaoli & Clifton, 2015; Hox et al., 2012; McNeish, 2016; Stegmueller, 2013; Zitzmann et al., 2015, 2016).

While published Bayesian studies have produced promising results, providing unbiased estimates for relatively small sample sizes, the following deficiencies remain to be overcome: (1) Extant studies often start with relatively large numbers of groups, for instance, 25, 40 or 50 (Depaoli & Clifton, 2015; Zitzmann et al., 2015, 2016). Only Hox et al. (2012) and McNeish (2016; McNeish & Stapleton, 2016b; Gelman, 2006) carried out simulations with fewer clusters. (2) McNeish and Stapleton (2016a) pointed out that more studies on Bayesian multilevel models are needed – especially with regard to the effect of various priors. In a similar vein, Zitzmann et al. (2015, p. 702) argued that it would be interesting to "incorporate previous findings from related studies by specifying an informative prior for the group-level effect". (3) Few studies advise researchers on what to do in small-sample situations (McNeish & Stapleton, 2016b, p. 496).

The present study seeks to address these shortcomings on the basis of a Monte Carlo simulation of the performance of ML and Bayesian estimation for very small samples (as few as 10 clusters). Further, historical data from Trends in Mathematics and Science Studies (TIMSS) was used to derive informative priors that enable researchers to obtain more stable estimates of group-level effects. Exploiting historical data to generate informative priors that facilitate Bayesian parameter estimation is an approach that has hitherto been rarely applied (König & Van de Schoot, 2017). Thus, in the present study we show how applied educational researchers can use existing knowledge to obtain reliable results in small-sample situations. Furthermore, a comparison of non, weakly, and highly informative priors adds important knowledge to the above-mentioned question of how different prior specifications influence Bayesian estimation. In summary, this study constitutes one of the first examples that show how this statistical approach can be applied successfully in educational science. This is of practical importance, since Bayesian analysis is becoming (or will soon be) increasingly popular in educational research and is now available in most software packages (even in SPSS 25, IBM 2017).

The structure of this paper is guided by the question "How many classes are needed to accurately assess instructional quality and its contextual effects on student outcomes (e.g., abilities)?" First, doubly latent multilevel models as state-of-the-art specifications of instructional quality in educational research are introduced. Second, the state of research regarding the sample-size question is presented. Third, in the main part, the study design of the Monte Carlo simulation is explained, and it is shown how prior distributions for Bayesian analysis are derived from historical large-scale data sets. The simulation results in terms of convergence rate, relative bias and relative root mean squared error (RMSE) are then presented. Finally, these results are discussed in the context of previous studies of sample-size requirements in SEM and multilevel frameworks. Implications for measuring educational instruction using student data are given.

The multilevel latent contextual model

Theoretical assumptions and prerequisites

Marsh et al. (2009) introduced the multilevel latent contextual model (MLCM) to accurately assess instructional quality and its effects. Like all structural equation models, it consists of a measurement part and a structural part. In relation to the former, Marsh, Lüdtke, and colleagues (e.g., Marsh et al., 2009) developed doubly latent multilevel models (DLMMs) against the background of *climate studies* that are interested in the effects of classroom or teacher characteristics on students' learning (Morin, Marsh, Nagengast, & Scalas, 2014, p. 144). Thus, these models are mainly referred to – and applied – in connection with *climate constructs* (e.g., classroom climate, classroom competitiveness and teacher's autonomy support). Climate constructs are of a *reflective* nature, where the term "reflective" refers to the assumption that the reports of students within a class reflect the same (latent) classroom climate variable and are thus aggregated at the class-level. Divergences between student perceptions are considered to be a measurement error and therefore a source of unreliability of the assessment of the classroom climate construct via student data (ibid., p. 146). Furthermore, students' deviations from the class mean have "no substantive meaning in relation to the interpretation of the L2 climate effects" (ibid., p. 147). The reason for this is that climate constructs refer to class-level features (e.g.: "Our math teacher cares about how we feel"), whereas individual constructs refer to individual students (e.g.: "My math teacher cares about how I feel"). Assessing the class-level referent calls for multilevel models (ibid.). With regard to the structural part, the modelling of contextual effects was also discussed in Marsh et al. (2009). As in Zitzmann et al. (2016), predicting an observed dependent variable (that is assumed to be measured without measurement error) is of central interest in the present study.

Statistical assumptions and notation

Ignoring the multilevel structure of the data would yield confounded effects that represent a mixture of class-level and student-level characteristics. Other substantive reasons for multilevel modelling (in particular with regard to climate constructs) relate to appropriately considering measurement AND sampling error (e.g., Lüdtke et al., 2008). While measurement error refers to unreliability that results from sampling only a finite number of items, sampling error refers to unreliability that results from sampling only a finite number of persons. Considering both types of error leads to climate constructs that are *doubly* latent (with regard to both items and persons) – which gives DLMM its name. Equation (1) shows the algebraic notation of DLMM (Marsh et al., 2009, p. 776), the measurement model (see Table 1):

Table 1:

Description of Parameters in Equation 1

Parameters	Comments
Y	student answer
$l \dots L$	indicators of the latent construct of interest
i	student
j	school class
y	latent construct of interest
$\lambda_{ly,w}$	within-factor loadings
$\lambda_{ly,b}$	between-factor loadings
U_{yij}	unobserved true score of construct y at level-1
U_{yi}	unobserved true score of construct y at level-2
R_{lyij}	level-1 residuals
R_{lyj}	level-2 residuals

$$Y_{lij} = \mu_{ly} + \lambda_{ly,w} * U_{yij} + R_{lyij} + \lambda_{ly,b} * U_{yj} + R_{lyj}\ . \tag{1}$$

The variance of student (i in school class j) answers (Y) to an indicator item (l) of the latent construct (y) is decomposed into a within and a between part as well as a factor and an error component (Muthén, 1991, p. 345; Equation (2)):

$$\sigma^2_{Ylij} = \lambda^2_{ly,w} * \sigma^2_{Uyij} + \sigma^2_{Rlyij} + \lambda^2_{ly,b} * \sigma^2_{Uyj} + \sigma^2_{Rlyj}\ . \tag{2}$$

The structural part of the multilevel latent contextual model is:

Micro Model (level-1): $$DV_{ij} = \beta_w * U_{ij} + \zeta_{ij}\ ; \tag{3a}$$

Macro Model (level-2): $$DV_j = \alpha + \beta_b * U_j + \zeta_j; \tag{3b}$$

Combined Model: $$DV_{ij} = \alpha + \beta_w * U_{ij} + \beta_b * U_j + \zeta_{ij} + \zeta_j\ , \tag{3c}$$

where βw denotes the within-group (level-1) regression coefficient and βb the between-group (level-2) regression coefficient. In this study, βb is of central interest, as it represents the group-level effect (i.e., climate effect). ζw and ζb refer to the dependent variable's residual variances at level-1 and level-2.

Estimation of these models requires large samples, in particular when ML estimation strategies are used that are based on asymptotic theory (e.g., Stegmueller, 2013). Hence, ML estimates and confidence intervals may be downwardly biased; in other words, significance testing is overstated in this case (ibid., p. 749). To address the small-sample issue, alternative statistics – Bayesian analysis, which is not based on the assumptions of asymptotic theory – are discussed in the literature (see Depaoli & Van de Schoot, 2017, for an application-oriented discussion).

Bayesian analysis

As previously mentioned, Bayesian analyses are not based on asymptotic theory, but for each parameter of interest, existing knowledge (prior distribution) is used, which is combined with the data (likelihood distribution) to find a distribution of most likely estimates (the posterior distribution). In other words, Bayesian approaches combine the information reflected by the prior distributions with the information contained in the collected data (for each parameter of interest). The prior multiplied with the data likelihood yields the full posterior distribution (for the parameter of interest). Thus, when working with large samples, priors have little impact. In contrast, model estimates are more sensitive to priors when small samples are analysed. Priors are usually categorized according to the degree of information they incorporate into the estimation process:

> Weakly informative priors contain more information compared to diffuse, but use less information than is available as to exhibit some degree of uncertainty (Gelman, 2006). Finally, an informative prior incorporates a great deal of certainty about the value of the model parameter (Depaoli and Clifton, 2015, p. 331).

With regard to this categorisation, two things should be noted: First, the idea of non-informative priors is "more a myth than reality" (McNeish, 2016) if the information of the data is limited (i.e., the likelihood of the data is small) (see McNeish & Stapleton, 2016b for a discussion). Second, Depaoli and Clifton (2015, p. 331) pointed out that "the extent to which parameters are recovered accurately in a Bayesian analysis depends in large part on the quality and amount of information modeled in the prior." In the same vein, scholars have argued that Bayesian methods offer a promising estimation approach for group-level effects in small samples especially if "strong and defensible prior knowledge is available" (Depaoli & Van de Schoot, 2017, p. 240; Zitzmann et al., 2015, 2016).

In conclusion, with the integration of prior knowledge via prior distributions, Bayesian estimation allows more complex models to be estimated with small data sets (e.g., Asparouhov & Muthén, 2010a; Lee & Song, 2004; Stegmueller, 2013; Van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015). Thus, the specification of priors is crucial when doing Bayesian analysis. In order to harness the potential of Bayesian analysis, we therefore specify highly *informative* priors by means of a historical-data-driven approach in the present study (see Section "Informative priors"). This data-driven approach is one of the present study's novelty, since published studies have usually used either default or theoretical priors, as can be seen from the following review of existing research on the sample size issue.

The performance of Bayesian and ML estimation in small samples

McNeish and Stapleton (2016, p. 298) concluded their review of two-level models with small samples as follows: "The effect of the number of clusters on model estimates has been found to be moderated by design elements such as the [intraclass correlation] ICC, the sample size within clusters, the scale of the outcome measure (binary or continuous), and the balance of the design". Further, the choice of estimation approach matters. While ML often leads to estimation problems under small-sample conditions (e.g., Hox et al., 2012; Zitzmann et al., 2016), Bayesian estimation seems more promising: In the study by Depaoli and Clifton (2015, p. 336)

> parameters were recovered most accurately under Bayesian estimation with informative priors, followed by Bayesian estimation with weakly informative priors, frequentist estimation, and Bayesian estimation with diffuse priors. In general, each estimator recovered the cluster-level covariate effect more accurately as the amount of information provided by the data increased (i.e., as the number of clusters, average cluster size, and ICC increased).

Bias. Extant research collected by McNeish and Stapleton (2016) shows that ML estimation of fixed-effects at either level are unbiased for as few as 15 clusters (Baldwin & Fellingham, 2013; Maas & Hox, 2004, 2005; Stegmueller, 2013). However, Lüdtke et al. (2011; see also Meuleman & Billet, 2009) showed – using a doubly latent approach – that the contextual effect is positively biased under extreme conditions (small ICC and low reliability; the lowest number of groups was 50) when ML estimation is used. For this reason, Depaoli and Clifton (2015), Hox et al. (2012), and Zitzmann et al. (2016) investigated whether (doubly) latent models and their contextual effects can be estimated without bias using Bayesian approaches. Hox et al. (2012) showed that, when Bayesian estimation with non-informative priors is used, negligibly biased group-level effects can also be estimated in small samples (level-2 units = 20). In contrast, Lüdtke et al. (2011) and Zitzmann et al. (2015) found for the "manifest-measurement/latent aggregation model" that both ML and Bayesian estimation are biased in small samples (level-2 units = 50) when ICC and group size are small. Under all other conditions, both estimation approaches provided almost unbiased estimates of the group-level effect. These results hold if a doubly latent model is used as predictor variable: Zitzmann et al. (2016) showed that under conditions with a low number of groups and low ICC, the Bayesian approach resulted in negatively biased group-level effects. However, if informative priors were used, context effects gave rise to negligible bias (Depaoli and Clifton, 2015; the lowest number of groups investigated was 40). In contrast, the use of diffuse priors resulted in biased estimates.

RMSE. In the doubly latent approach, ICC has a strong effect on the RMSE when ML estimation is used (Lüdtke et al. 2011). Zitzmann et al. (2016) and Depaoli and Clifton (2015) concluded that Bayesian estimation (in particular, informative priors) leads to more accurate (in terms of RMSE) estimates of the group-level effect under conditions with a low number of groups and a low ICC. Furthermore, Depaoli and Clifton (2015) found that for ML and Bayesian estimation with diffuse priors, the RMSE was negatively related to the number of groups, group size, and ICC.

These findings suggest that the advantages of the Bayesian toolbox come into effect primarily (a) for small sample sizes, (b) when weakly informative or informative priors are used, and (c) when ICC is high. The present study extends these findings to the use of highly informative priors and very low numbers of groups. To explore the potential benefits of the Bayesian toolbox, we specified conditions typically encountered in psychological and educational research (e.g., Zitzmann et al. 2015, 2016; McNeish, 2016), which are presented in the next section.

Method

In order to investigate how the various estimation approaches perform on small samples, a simulation study was set up and conducted as follows.

Population and analysis models

Multiple data sets were produced using a data generation model (population values also called true values) and then analysed based on analysis models that differed in the estimation approach used. The specification of the data generation model and the analysis models was identical: A four-indicator doubly latent model with cross-level constraints (i.e., metric invariance across school classes) was specified as independent variable. The first indicator was fixed to 1. This independent variable is assumed to predict a dependent variable that is measured without measurement error. Thus, a group-level effect was specified as structural path at the between level. With regard to the population model one TIMSS 2015 country was chosen randomly, and its empirical parameter values regarding the within and between factor loadings, the within variance components, and the group-level effect were used as population values. The level-2 variance components were varied conditional on the desired ICC values, which represented one factor of the simulated conditions (see below). Additionally, one level-2 error variance was fixed to zero because negative residual variances are often encountered in real data analyses (see also Maas & Hox, 2005, p. 89). Regarding the *Bayesian analyses* models the priors used are described in detail below.

Factorial design

In order to investigate the performance of various estimation approaches, the following full factorial 6 x 4 x 3 x 4 design (288 conditions) was set up. The values for Factor 1 – number of groups (school classes) – were 10, 15, 20, 25, 30, and 100, with an equal number of level-1units each (balanced design). The lower limit of the range was chosen against the background of Snijders and Bosker's (1999) statement "that multilevel modeling becomes attractive when the number of groups is larger than 10" (cited from Maas & Hox, 2005, p. 90). Further, the values between 10 and 30 were chosen because data collection in 10 to 30 school classes is manageable for a single researcher, whereas a larger number of classes might only be possible with institutional or even national support. In addition, a sample size of 100 was chosen to study the asymptotic behaviour of the estimators. The values for Factor 2 – intraclass correlation (ICC) were .05, .10, .15, and .20. Lacking a systematic review of typical ICC values in instructional science, we investigated the latest TIMSS data set (2015, grade 8). This data shows that the ICC for the latent construct used in the present study varied significantly between countries in the range from 6% (Ireland, Korea) to 21% (Dubai). A brief glance at the literature indicates that – depending on the nature of the construct assessed – ICC values typically lie at the upper bound of this range, around 15% to 20% (Willems, 2011; Helm, 2016) or even higher (Warwas & Helm, 2017; Fauth, Decristan, Rieser, Klieme, & Büttner, 2014), and low values seem to be rare. For instance, Kunter, Baumert, and Köller (2007) found that "perceived rule clarity" as a classroom management measure has only limited class-level variation (8%). Factor 3 represents the precision of the DLMM (sampling error, Lüdtke, Trautwein, Kunter, & Baumert, 2006) and was analysed at three different levels: Since sampling error is a function of ICC and *k* (i.e., average number of raters in the group/cluster; e.g., students in classes), *k* had a value of 5, 15, or 25 for each ICC condition. This approach resulted in data sets with increasing level-2 reliability (.21 to .86), and decreasing sampling error. According to Lüdtke, Trautwein, Kunter, and Baumert (2006), a level-2 reliability greater than .70 is acceptable. In the present study, this requirement was met for conditions with either k = 25 if ICC ≥ 10% or k = 15 if ICC ≥ 15%. Factor 4 – estimator – refers to 4 different estimation approaches: ML estimation, Bayesian estimation with non-informative priors (Mplus default options were used), Bayesian estimation with weakly informative priors for level-2 variance components as recommended by Depaoli and Clifton (2015) as well as Zitzmann et al. (2015) (low to moderate ICC >> IG (0.1, 0.1), large ICC >> IG (-1, 0)), and Bayesian estimation using highly informative priors based on historical TIMSS data. In the following section, detailed information is given on how the highly informative priors were derived.

Informative priors

Since in educational research extensive information of high (and internationally accepted) quality is available in the form of large-scale data sets, it makes sense to use this existing information for Bayesian analysis. Thus, we exploited student questionnaires from TIMSS 1999, 2003, 2007, 2011, and 2015 by estimating MSEM as described in Section "Multilevel latent contextual model" (see also Zitzmann et al., 2016).

In the first step, we specified an MSEM for each partaking country. The independent variable was represented by a four-indicator doubly latent model[2] with cross-level constraints (i.e., metric invariance across school classes). The indicators were items relating to mathematics instruction: I am interested in what my teacher says. My teacher gives me interesting things to do. My teacher has clear answers to my questions. My teacher is good at explaining mathematics. Each item had a response pattern ranging from 1 (agree strongly) to 4 (disagree strongly). With regard to the structural part of the MSEM, the independent variable was assumed to predict a dependent variable that was measured without measurement error. For some TIMSS assessments, a significant proportion of countries showed ill-fitting models or models that – due to very small negative residual variances of a level-2 indicator – did not converge. In such cases the error variances of two level-1 indicators were allowed to co-vary and/or the negative residual variance of a level-2 indicator was fixed to zero. Models that still did not fit (CFI < .90) or did not converge were excluded. This approach led to 365 estimated models.

In the next step, the distribution of each model parameter at level-2 across all estimated models was used to calculate the hyperparameters of the priors (see Table 2). Among these hyperparameters, those for the level-2 variances are of special interest, as variance estimates are most sensitive to priors (Asparouhov & Muthén, 2010a, p. 6; Lee & Song, 2004; Stegmueller, 2013; Baldwin & Fellingham, 2013; Browne & Draper, 2006; Gelman, 2006). Although for variance components the inverse gamma (IG) distribution is most commonly used in research practice, discussion is ongoing about which priors perform best under what conditions (e.g., Depaoli & Clifton, 2015; Gelman, 2006; Zitzmann et al., 2015). Zitzmann et al. (2015) recommended using larger shape/scale values (0.1, 0.1) because they pool group-level variance estimates away from zero when group-level variance is small, which is often the case in DLMM. Here, we use and compare non-informative IG priors (Mplus default), weakly informative IG priors as recommended by Depaoli and Clifton (2015) as well as Zitzmann et al. (2015) – that is, IG(-1, 0) for conditions with large ICC and IG(0.1, 0.1) for conditions with small to moderate ICC – and the priors derived from TIMSS data. For the last-mentioned, Table 2 provides information about the prior specification for the level-2 part. R squared values in Table 2 indicate how well the theoretical parameter distribution (based on the hyperparameters) works as a "proxy" for the empirical parameter distribution (based on TIMSS data). Note the mismatch regarding

[2] The "V" parameterization (Asparouhov & Muthén, 2010a) was used where the first factor loading is fixed to 1 and the variance of the factor is estimated.

the inverse gamma distribution: Very small values of variances (near zero) are extremely unlikely, but they often occur with empirical data and can lead to convergence problems.

Table 2:
Information on Informative Priors at Level 2

Parameter	Prior	Type	Source of information	Hyperparameters	R^2: Mean (SD)
Group-level effect (βb)	normal	informative	TIMSS 1999 TIMSS 2003 TIMSS 2007_G4 TIMSS 2007_G8 TIMSS 2011_G4 TIMSS 2011_G8 TIMSS 2015_G4 TIMSS 2015_G8	(m = -27.076, sd = 189.59)	.89 (.01)
Outcome variance	inverse gamma	informative		(shape = 4.771, scale = 8659.99)	.85 (.09)
Latent variance (U_{yj})	inverse gamma	informative		(shape = 3.889, scale = 0.115)	.80 (.11)
Residual variances (ε_B)	inverse gamma	informative		(shape = 2.479, scale = 0.040)	.90 (.06)
Factor loadings (λ_B)	normal	informative		(m = 1.211, sd = 0.760)	.91 (.01)
Indicator intercepts (μ_B)	normal	informative		(m = 1.814, sd = 0.415)	.96 (.01)

Note. The hyperparameters for the normal priors are simply the mean (m) and standard deviation (sd) of the empirical distribution. The hyperparameters for the inverse gamma priors are derived by the following equations: shape (α) = $2 + m^2/v$; scale (β) = $m + m^3/v$; where m and v denote the mean and variance of the empirical distribution. R^2 refers to the amount of variation in the observed distribution that is explained by the theoretical distribution using the hyperparameters.

Technical information

TIMSS data analyses, data generation, and model estimation were carried out in Mplus 8 (Muthén & Muthén, 1998-2017) using MplusAutomation (Hallquist & Wiley, 2016). The simulation results were analysed in R (R Development Core Team, 2016) using the packages ggplot2 (Wickham, 2009), MCMCpack (Martin, Quinn & Park, 2011), and stringr (Wickham, 2017).

For each of the 288 conditions, 1000 data sets were generated. Additional analyses (i.e., cumulative averages plots) not reported here showed that 1000 sets were sufficient, since estimates stabilized around this value. For Bayesian analyses, Gibbs sampling with two chains (which is the default in Mplus 8; Asparouhov & Muthén, 2010b) for 10,000 iterations (i.e., default MDITERATIONS option in Mplus 8) was used. To reduce computational time, we used a supercomputer with 2048 CPU cores and 16 terabyte memory. Since in the Bayesian approach parameter estimates are taken from a posterior distribution, summary statistics (mean, median, mode) represent the Bayesian parameter estimate. As recommended by Zitzmann et al. (2015, 2016), we used the mode of the posterior distribution.

Reported quantities

To compare the performance of the estimators under various simulated conditions, five central and common quantities were evaluated: convergence rate, bias of (point) estimates, and accuracy in terms of the root mean squared error (RMSE).

Convergence rate. This rate refers to the ratio of completed/converged to requested Monte Carlo replications. Models that did not converge were excluded from the analyses. In the case of the Bayesian approach, the convergence diagnostic PSR (Potential Scale Reduction, Gelman & Rubin, 1992; Asparouhov & Muthén, 2010b) was used to assess chain convergence.[3] PSR is implemented in Mplus and has also been used in other simulation studies (e.g., Hoofs, Van de Schoot, Jansen, & Kant, 2015). The basic idea of PSR is to relate the within-chain variance to the between-chain variance of a parameter over a certain number of iterations. Low between-chain variance – that is, a PSR below 1.05 (for models with one parameter) or 1.10 (for models with a large number of parameters) – indicates convergence (Asparouhov & Muthén, 2010b).

Bias of (point) estimates. This central quantity reflects the deviation of the parameter estimates (averaged over all replications) from the true population value. The percent relative bias "is simply the difference between estimated and true value expressed as a proportion of the true value" (Stegmueller, 2013, p. 752): $\frac{\hat{\theta}-\theta}{\theta} * 100$. According to Muthén and Muthén (2002), the relative bias should not be greater than 10% (see also Hoogland & Boomsma, 1998).

Root mean squared error (RMSE). Zitzmann et al. (2015, p. 694) argued that "when the estimator is unbiased, a single estimate might still not be close to the true value." Thus, this evaluation criterion indicates the overall accuracy of the average parameter estimate. The RMSE was calculated by the square root of the expectation of the squared deviation of the estimate from the population value divided by the population value (ibid.). Since RMSE combines relative bias and variability of estimates, they argued that a more accurate estimator that might produce slightly more biased estimates is preferable to a less biased estimator that produces estimates that are slightly more variable.

[3] Although Depaoli and Van de Schoot (2017) recommend visual inspection of convergence for each parameter and each model in Bayesian analysis, this is simply not possible in simulation studies with thousands of estimated parameters.

Results

Convergence

Under all conditions, the convergence rate was high (at least 99%). Only under conditions with small numbers of groups (< 15) and small k (= 5) values the Mplus default priors yielded convergence rates that were slightly lower. Although the average convergence rates of the ML approach seem acceptably high at first glance, all ML models for which the number of groups was smaller than the number of parameters (i.e., 21) resulted in warning messages. These warning messages indicated that, due to the low number of clusters, the models might not be identified. Moreover, for 2.8% of all ML models, warnings indicated that a saddle point was reached. Other warning messages involved less than 1% of all replications. Non-converged solutions were not included in the final analyses. At this point, we want to refer to the valuable comment of one of our blind reviewers, who argued that the reported differences between ML and Bayesian estimations might not reflect a problem of the ML method per se, but might rather be a result of the unconstrained variance estimation that is implemented in Mplus.[4]

Relative bias

In this section, the evaluation criteria are inspected for the group-level effect only (see Table 3), since this parameter is of principal interest to applied researchers. For all other parameters, information is available from the author. Columns 3-6 of Table 3 show the relative bias for the group-level estimate. If informative priors or weakly informative priors are used, Bayesian estimation clearly outperforms ML estimation. Interestingly, the two Bayesian approaches performed equally well and provided almost unbiased estimates even under the most unfavourable conditions (small number of groups, low ICC, low group size). In contrast, the bias for the ML parameters can be considered acceptable only under favourable conditions (ICC ≥ 20% and number of groups > 20), as otherwise estimates are positively biased. Mplus default priors, however, lead to negatively biased average estimates when the ICC is low (≤ 10%) and the number of groups is small (≤ 15). With a number of groups as large as 100 (and ICC ≥ 10%) only the uninformative (Mplus default) priors produced downwardly biased estimates that are slightly outside the 10% limit.

In terms of accuracy of the average group-level parameter estimate, the two informative Bayesian approaches, again, clearly performed better than the non-informative Bayesian and the ML estimation approach. Columns 7-10 of Table 3 show very low values for the weakly informative and informative Bayesian approaches under all conditions, whereas

[4] Note that in our simulation study we do not evaluate the ML estimator per se but the way in which it is implemented in Mplus.

for the other approaches comparably high RMSE values are found, especially when ICC and the number of groups are low. With 100 groups and an ICC ≥ 10% all four estimation approaches yielded RMSE values ≤ 10.

Table 3:
Relative Bias and RMSE for the Group-level Effect of the Latent Predictor Variable with Various Estimation Approaches – Conditional on ICC

		Relative Bias				RMSE			
N	ICC	ML	B(non.)	B(weak)	B(inf.)	ML	B(non.)	B(weak)	B(inf.)
10	5	79.43	-22.37	1.53	-0.33	24.1	40.92	0.96	0.46
	10	67.54	-38.92	-3.93	0.34	16.26	30.31	0.96	0.46
	15	49.09	-7.03	-1.98	-1.00	12.04	33.15	0.95	0.46
	20	44.59	235.48	-1.01	-0.64	9.84	26.94	0.95	0.46
15	5	51.72	-68.18	-3.77	-2.21	18.52	20.65	0.90	0.45
	10	55.45	-40.03	-2.94	-1.13	11.62	24.67	0.89	0.44
	15	43.09	-11.84	-4.57	-0.14	8.3	19.65	0.92	0.46
	20	31.30	-10.17	-3.38	-0.06	6.2	11.06	0.89	0.44
20	5	64.73	-37.02	-2.84	1.36	17.36	10.68	0.88	0.44
	10	45.67	-2.94	-2.81	-1.59	10.24	9.76	0.85	0.44
	15	24.27	-2.40	-0.67	0.39	6.68	5.97	0.87	0.43
	20	10.73	-11.47	-4.34	0.24	5.16	4.88	0.85	0.43
25	5	61.14	-21.62	-2.94	-0.37	14.3	9.69	0.83	0.44
	10	25.10	-7.82	-4.02	-0.42	8.74	6.59	0.84	0.43
	15	18.47	-14.56	-2.55	-1.86	5.74	4.59	0.81	0.42
	20	15.91	-12.55	-4.64	-1.53	4.28	4.07	0.83	0.42
30	5	46.03	-35.39	-4.8	0.34	12.93	8.55	0.79	0.43
	10	36.26	-20.62	-6.97	-1.69	7.66	5.76	0.83	0.42
	15	15.08	-5.08	-3.29	-0.42	4.62	4.06	0.81	0.41
	20	5.27	-11.64	-3.81	-0.83	3.53	2.93	0.78	0.41
100	5	20.18	-77.65	-7.10	-1.42	6.16	18.07	0.69	0.36
	10	-1.89	-16.94	-7.19	-1.53	2.76	3.27	0.73	0.37
	15	-8.06	-13.95	-7.50	-2.03	1.67	1.81	0.71	0.36
	20	-9.60	-15.31	-9.25	-2.18	1.28	1.41	0.71	0.37

Note. N = number of groups, ICC = intraclass correlation, RMSE = root mean squared error, ML = maximum likelihood, B = Bayes, non. = non-informative, inf. = informative, weak = priors for variance components as recommended by Depaoli and Clifton (2015) and Zitzmann et al. (2015)

All analyses presented were also carried out conditional on k (group size) instead of ICC (see Table 4). With regard to the relative bias, group size (k) has an effect only if k ≤ 15 and ML or Bayesian non-informative estimation is used. Under all other conditions (except for some non-informative Bayesian conditions) the relative bias of the group-level effect is within acceptable limits. Similar to the findings above, the RMSE is comparably higher for ML and non-informative Bayesian estimation under all conditions – especially if k is low (= 5).

Table 4:
Relative Bias and RMSE for the Group-level Effect of the Latent Predictor Variable with Various Estimation Approaches – Conditional on Group Size

		Relative Bias				RMSE			
N	k	ML	B(non.)	B(weak)	B(inf.)	ML	B(non.)	B(weak)	B(inf.)
10	5	170.09	149.62	-0.61	0.51	20.48	79.02	0.96	0.46
	15	-3.06	-21.84	-2.14	-0.53	13.49	11.13	0.96	0.47
	25	13.46	-2.41	-1.30	-1.20	12.71	8.34	0.95	0.45
15	5	117.70	-79.99	-3.33	-1.33	14.76	43.2	0.91	0.45
	15	11.70	-9.07	-3.25	-1.56	10.09	7.32	0.88	0.45
	25	6.77	-8.61	-4.42	0.23	8.64	6.50	0.90	0.44
20	5	106.82	-24.18	-1.32	0.75	13.7	11.74	0.86	0.44
	15	-2.52	-15.65	-3.61	-0.48	9.48	6.29	0.88	0.43
	25	4.75	-0.55	-3.06	0.04	6.39	5.44	0.85	0.43
25	5	89.78	-31.62	-4.84	-0.97	12.31	8.43	0.81	0.43
	15	-6.95	3.10	-2.43	-1.21	7.08	5.74	0.84	0.43
	25	7.64	-13.90	-3.36	-0.96	5.4	4.54	0.84	0.42
30	5	86.28	-23.50	-5.98	-0.38	10.98	7.24	0.80	0.42
	15	-9.38	-11.46	-2.79	-0.71	5.86	4.80	0.81	0.42
	25	0.09	-19.58	-5.39	-0.86	4.71	3.94	0.80	0.41
100	5	16.39	-57.14	-7.35	-1.85	4.83	14.41	0.70	0.36
	15	-12.31	-25.37	-9.84	-2.30	2.32	2.08	0.70	0.36
	25	-3.60	-10.39	-6.09	-1.22	1.75	1.93	0.73	0.37

Note. N = number of groups, k = group size, RMSE = root mean squared error, ML = maximum likelihood, B = Bayes, non. = non-informative, inf. = informative, weak = priors for variance components as recommended by Depaoli and Clifton (2015) and Zitzmann et al. (2015)

Furthermore, all analyses were conducted using the mean as summary statistic of the posterior distribution. The results in terms of group-level effect are only different for the non-informative (Mplus) priors with regard to relative bias and RMSE. Both are substantially higher than the values based on the mode statistics. However, these differences do not affect the general conclusions.

Sensitivity analysis of prior specifications

In Bayesian analysis, a central question is to what extent the specification of the hyperparameters of the priors influences the study results. Hence, usually robustness is checked by varying the choice of hyperparameters systematically. Since the behaviours of non, weakly, and highly informative priors were investigated in the results, a robustness check was already done. As can be seen from the results, the choice of hyperparameters for the level-2 priors strongly influences the results in terms of relative bias and RMSE. In particular, using default options can lead to biased and inaccurate average Bayesian parameter estimates.

In summary, we regressed the evaluation criteria on the simulated conditions. The conditions were dummy coded. To facilitate interpretation, the most favourable conditions (N groups = 100, ICC = 25%, k = 25) and informative Bayesian estimation were chosen as reference categories, since they led to the least biased and most accurate parameter estimates. Table 5 confirms the results presented above: The relative bias of parameter estimates significantly increases with ML estimation, low number of groups (= 10) and low group size (= 5). The same is true for the accuracy of the estimates of the group-level effects. Furthermore, using Mplus default priors increases the RMSE significantly. Note that, as can be seen in Tables 2 and 3, each of the reported main effects is likely to be moderated by the other conditions (especially by low ICC and k values).

Table 5:
Effects of the Simulated Conditions on Relative Bias and RMSE for the Group-level Effect of the Latent Predictor Variable

	Relative Bias		RMSE	
	b	*p*	b	*p*
Intercept	-15.82	.181	-7.57	.000
N groups = 10	**35.14**	.001	**9.91**	.000
N groups = 15	12.16	.267	**5.33**	.004
N groups = 20	15.17	.166	2.20	.237
N groups = 25	12.95	.237	1.39	.453
N groups = 30	10.62	.332	0.89	.633
ML	**33.76**	.000	**8.74**	.000
Bayesian (non.)	-10.47	.241	**12.47**	.000
Bayesian (weak)	-3.17	.723	0.42	.783
ICC = 5	-8.44	.345	**5.02**	.001
ICC = 10	-7.26	.417	2.34	.123
ICC = 15	-7.48	.402	1.12	.462
k = 5	**22.81**	.003	**7.12**	.000
k = 15	-3.32	.667	0.64	.625
adj. R^2	*0.13*	*.000*	*0.38*	*.000*

Note. Statistically significant coefficients are printed in bold, b = unstandardized regression coefficient, p = p value, N = number of groups, ICC = intraclass correlation, k = group size, RMSE = root mean squared error, ML = maximum likelihood, non. = non-informative, weak = priors for variance components as recommended by Depaoli and Clifton (2015) and Zitzmann et al. (2015), adj. R^2 = adjusted R squared

Discussion

This study contributes to the debate around the sample-size question for multilevel structural equation models (MSEMs). The present simulations were based on doubly latent models with cross-level constraints, and allowed investigation of the performance of ML versus non, weakly, and highly informative Bayesian estimation techniques. With regard to the highly informative Bayesian approach, this study extends the knowledge on how existing information can be used to obtain more accurate parameter estimates for small samples. More specifically, from a large set of international large-scale student surveys, hyperparameters for priors were derived that can be used in further studies. Only few studies have investigated whether the Bayesian toolbox helps to overcome small-sample issues in the case of MSEMs (e.g., Hox et al., 2012; Zitzmann et al., 2016, 2015). Of these, most focus either on multilevel models with manifest variables or on latent variable models without multiple levels. Nonetheless, these studies illustrate the potential of Bayesian estimation. Thus, we assumed that incorporation of prior information using Bayesian analysis would help to produce reasonable results with small samples. Additionally, we assumed that intraclass correlation and group-level reliability of the independent variable affect estimation of the group-level effect. Using Monte Carlo simulation, we compared the performance of the various estimation techniques under several conditions typically encountered in educational research (e.g., small samples). As expected, the Bayesian approach outperformed ML estimation under all simulated conditions if weakly informative or informative priors were used. Weakly informative and informative priors yielded almost unbiased and highly accurate estimates even under unfavourable conditions (small number of level-2 groups and small group size).

The reported findings are in line with those of previous studies. Depaoli and Clifton (2015) found that group-level effects are most accurate when informative priors are used (followed by weakly informative priors, ML estimation, and diffuse priors). The present study revealed the same ranking of the performance of the various estimation approaches (see Column 4 in Table 5 for the additional RMSE in comparison to B(inf.): ΔB(weak): .42, ML: 8.47, ΔB(non.): 12.47). Further, our findings show – in line with Zitzmann et al. (2016) and Depaoli and Clifton (2015) – that group-level effects of MSEMs with DLMM can be estimated more accurately with Bayesian approaches when weakly informative or informative priors are used. Summarizing published studies, McNeish and Stapleton (2016) argued that the effect of the number of groups is moderated by ICC and the group size. This corresponds to the findings of this study, especially for ML estimation and non-informative Bayesian estimation results (see Table 3). Scholars often argue that small sample sizes lead to biased estimates. However, several studies – including this one – have shown that the impact of the number of groups on various evaluation criteria (such as relative bias) is (a) relatively low when compared to the effect of estimation approaches and (b) is substantial only when ICC and group size are also low. Even in these unfavourable situations, Bayesian estimation with weakly informative or informative priors could help to accurately estimate MSEM group-level effects. Thus, careful choice of priors may help to save data collection resources and to limit perceived burdens of external evaluations in schools.

From the findings of this study, recommendations can be derived for assessing instructional quality and its effects in situations where only a small number of school classes is available:

1. When designing a study of the effects of instructional features, one should reflect upon the expected intraclass correlation of the climate construct to be assessed. Some constructs may vary more substantially at the class level than others. Low ICC is supposed to lead to greater bias of class-level effects.

2. When collecting data, one should keep in mind that, the higher the number of students within each school class, the less biased the group-level effects. The total number of school classes seems less relevant, since even a number of 100 could lead to bias; however, Bayesian estimation using weakly informative or informative priors works well even for very small samples.
3. When using Bayesian estimation with informative priors, one should carefully choose the hyperparameters. For instance, Zitzmann et al. (2015) showed that the choice of shape and scale parameters of the inverse gamma prior also affects the magnitude of group-level effect. Thus, sensitivity analyses should be performed. Depaoli and Van de Schoot (2017) provided comprehensive guidance for selecting priors and investigating their effects on study results. For applied instructional research, one might use the hyperparameters presented in Table 2 or those suggested by Depaoli and Clifton (2015). Both sets of hyperparameters led to similarly unbiased estimates. However, note that the priors in Table 2 are based on specific TIMSS items relating to cognitive activation in mathematics instruction. Further research is necessary to determine how they perform in other domains and in the context of other climate constructs.
4. Finally, when specifying the MSEM cross-level, constraints as recommended by Jak and Jorgensen (2017) should be used. Firstly, this ensures that the "same" construct is measured in each school class (i.e., metric invariance). Secondly, the present study is also based on DLMM with cross-level constraints, so it is unclear how ML and Bayesian estimation perform for small samples when factor loadings are not constrained across levels.

Additionally, applied researchers should keep in mind that Monte Carlo studies are performed under optimal conditions; this means that several assumptions required by MLM hold true in simulation studies, but may not be fulfilled in studies using empirical data (see also McNeish & Stapleton, 2016b). For instance, data was generated under the normality assumption, which produced real numbers. However, in practice, categorical data is collected by means of questionnaires with a 1 to 5 response pattern that violates the normality assumption by definition. Thus, convergence and bias are likely to be worse for empirical data (Stegmueller, 2013).

Furthermore, more solutions to the small-sample problem exist than we have presented. Other simulation studies showed that

- other estimation approaches, such as the restricted maximum likelihood or the ML with Kenward-Roger adjustment, also work well under unfavourable conditions (McNeish & Stapleton, 2016b);
- the doubly *manifest* approach leads to more accurate group-level effects with small groups and low ICC than the doubly latent approach (Lüdtke et al., 2011), although the doubly manifest approach is more biased; this counterintuitive finding is known as bias-accuracy trade-off (ibid.);

- other priors (not implemented in Mplus, such as the half-Cauchy prior) for group-level variances might work even better than those used in this study (McNeish & Stapleton, 2016b);
- "the number of items and the size of the loadings had a strong effect on the magnitude of the estimated relative percentage bias" (Lüdtke et al., 2011, p. 454).

These and other conceivable conditions fell outside the scope of the investigations presented here. Nonetheless, this simulation study demonstrates once more the potential of the Bayesian toolbox.

References

Asparouhov, T., & Muthén, B. (2010a). *Bayesian Analysis of Latent Variable Models using Mplus.* Available from: https://www.statmodel.com/download/BayesAdvantages18.pdf (23.04.2018)

Asparouhov, T., & Muthén, B. (2010b). *Bayesian Analysis Using Mplus: Technical Implementation.* Available from: https://www.statmodel.com/download/Bayes3.pdf (23.04.2018)

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

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

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., . . . Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69–102. doi:10.3102/0034654308325581

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(3), 327–351. doi:10.1080/10705511.2014.937849

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22*(2), 240–261. doi:10.1037/met0000065

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29,* 1–9. doi:10.1016/j.learninstruc.2013.07.001

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–511.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515–533. doi:10.1214/06-BA117A

Hallquist, M., & Wiley, J. (2016). *MplusAutomation: Automating Mplus Model Estimation and Interpretation.* R package version 0.6-4. https://CRAN.R-project.org/package=MplusAutomation (23.04.2018)

Helm, C. (2016). Basic dimensions of instructional quality and their effects on student outcomes in accounting. *Zeitschrift für Bildungsforschung, 6*(2), 101–119. doi:10.1007/s35834-016-0154-3

Hoofs, H., Van de Schoot, R., Jansen, N. W. H., & Kant, I. (2017). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement, 34*(6), 1–32. doi:10.1177/0013164417709314

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., Van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6*(2), 87–93. doi:10.18148/srm/2012.v6i2.5033

IBM (2017). *SPSS Inc. Released 2017. SPSS for Windows, Version 25.0.* Chicago, SPSS Inc.

Jak, S., & Jorgensen, T. D. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology, 8*(1640), 1–9. doi:10.3389/fpsyg.2017.01640

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies.* Unpublished manuscript, California State University, Los Angeles.

König, C., & Van de Schoot, R. (2017). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review,* 1–24. doi:10.1080/00131911.2017.1350636

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*(5), 494–509. doi:10.1016/j.learninstruc.2007.09.002

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*(4), 653–686. doi:10.1207/s15327906mbr3904_4

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467. doi:10.1037/a0024376

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

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9,* 215–230. doi:10.1007/s10984-006-9014-8

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. doi:10.1027/1614-2241.1.3.86

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764–802. doi:10.1080/00273170903333665

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software, 42*(9), 1–21. doi:10.18637/jss.v042.i09

McNeish, D. M. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750–773. doi:10.1080/10705511.2016.1186549

McNeish, D. M., & Stapleton, L. M. (2016a). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*(2), 295–314. doi:10.1007/s10648-014-9287-x

McNeish, D. M., & Stapleton, L. M. (2016b). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51*(4), 495–518. doi:10.1080/00273171.2016.1167008

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3,* 45–58.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly-latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education,* 82(2), 143–167. doi:10.1080/00220973.2013.769412

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338–354. doi:10.1111/j.1745-3984.1991.tb00363.x

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599–620. doi:10.1207/S15328007SEM0904_8

R Development Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science, 57*(3), 748–761. doi:10.1111/ajps.12001

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*(1), 25216. doi:10.3402/ejpt.v6.25216

Warwas, J., & Helm, C. (2017). Enjoying working and learning in vocational education – A multilevel investigation of emotional crossover and contextual moderators. *Empirical Research in Vocational Education and Training, 9:11.* doi:10.1186/s40461-017-0055-2

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag. doi:10.1007/978-0-387-98141-3

Wickham, H. (2017). *stringr: Simple, Consistent Wrappers for Common String Operations.* R package version 1.2.0. https://CRAN.R-project.org/package=stringr (16.09.2017)

Willems, A. S. (2010). *Bedingungen des situationalen Interesses im Mathematikunterricht: Eine mehrebenenanalytische Perspektive.* Münster: Waxmann.

Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research, 50*(6), 688–705. doi:10.1080/00273171.2015.1090899

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 661–679. doi:10.1080/10705511.2016.1207179