Educ Psychol Rev (2016) 28:295–314
DOI 10.1007/s10648-014-9287-x

REVIEW ARTICLE

The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration

Daniel M. McNeish · Laura M. Stapleton

Published online: 19 October 2014

Abstract Multilevel models are an increasingly popular method to analyze data that originate from a clustered or hierarchical structure. To effectively utilize multilevel models, one must have an adequately large number of clusters; otherwise, some model parameters will be estimated with bias. The goals for this paper are to (1) raise awareness of the problems associated with a small number of clusters, (2) review previous studies on multilevel models with a small number of clusters, (3) to provide an illustrative simulation to demonstrate how a simple model becomes adversely affected by small numbers of clusters, (4) to provide researchers with remedies if they encounter clustered data with a small number of clusters, and (5) to outline methodological topics that have yet to be addressed in the literature.

Keywords Multilevel model · HLM · Small sample · Mixed model · Small number of clusters

Frequently in educational psychology research, observations have a hierarchical structure (Raudenbush and Bryk 2002). Students are nested within classrooms; children are nested within families, or teachers are nested within schools. When data are sampled in a multi-stage manner or if observations are clustered, modeling data by ignoring the clustering will often result in standard error estimates that are underestimated if the outcome variable demonstrates dependence based on the clustering (i.e., the intraclass correlation is greater than zero). When clustering is ignored, the residuals will not be identically and independently distributed, violating an assumption of single-level models such as ordinary least-squares regression. This dependence will ultimately result in an inflated type-I error rate for significance tests of regression coefficients. However, in the statistical literature, methods have been developed for addressing data that come from a hierarchical structure and can account for the dependence among observations. One such method has many names and acronyms but is often referred to as hierarchical linear models (HLMs), multilevel models (MLMs, used in this paper), or mixed-effects models (Raudenbush and Bryk 2002). This is the method on which this paper will focus.

To estimate MLMs without bias, adequate sample sizes must be obtained, since MLMs are often estimated with maximum likelihood (ML) methods. ML estimates are asymptomatically

D. M. McNeish (✉) · L. M. Stapleton
Measurement, Statistics, and Evaluation Program, Department of Human Development and Quantitative Methodology, University of Maryland, 1230 Benjamin Building, College Park, MD 20742-1115, USA
e-mail: DMcNeish@umd.edu

unbiased, meaning that they perform well as sample sizes approach infinity but are known to behave less desirably with smaller samples sizes, particularly when the number of clusters is small. Although a specific sample size to ensure unbiased estimates cannot been pinpointed, a few guidelines have been suggested such as 30 clusters (also deemed macrounit, level-2 sample size, site, or individuals for longitudinally clustered data) with a cluster size of 30 (microunit, level-1 sample size, or repeated measures for longitudinally clustered data) in Kreft (1996), a minimum of 20 clusters (Snijders and Bosker 2012), or 50 clusters with a cluster size of 20 for cross-level interactions or 100 clusters with 10 units each if the main interest is in the variance components (Hox 1998, 2010). From a design perspective, Snijders and Bosker (1993) also advise against MLMs if the number of clusters is below 10. However, in applied settings, the demands of these recommendations are not always realized, leading to potentially biased results. For instance, in a review by Dedrick et al. (2009), using the 30/30 guideline, of the 99 studies reviewed using MLMs between 1999 and 2003 in 13 journals from education, psychology, and sociology, 21 % had sample sizes that would not meet the recommendation. This finding suggests that researchers may not be aware of sample size recommendations or may not have the resources to adequately obtain large samples because recruiting large amounts of schools or programs can be exceedingly difficult. Even though models with small sample sizes may converge to a solution and produce parameter estimates, the estimates may contain bias which can affect inferences in applied research. It is important that researchers are aware of and acknowledge the potential untrustworthiness present with MLMs when sample size is small.

To outline the contents of this paper, first, a brief overview of MLMs and the terminology used in this paper is provided. Then, the existing literature on MLMs with a small number of clusters is reviewed in detail. Third, an illustrative simulation study is provided that depicts the potential issues in modeling data with a small number of clusters with MLMs. Fourth, alternative methods and approaches for analyses with a small number of clusters are presented. Lastly, methodological topics that have yet to appear in the literature and could serve as a basis for future research are discussed.

Brief Overview of MLMs

MLMs are conceptually similar to multiple regression in that an outcome variable is linearly predicted from multiple independent variables (a.k.a. covariates or predictors). An assumption of traditional regression models such as ordinary least squares for continuous outcomes or logistic regression for binary outcomes is that observations (or the residuals) are independent. MLMs relax the independence assumption by directly modeling the clustered structure of the data through the inclusion of random effects or by directly modeling an alternative covariance structure for the residuals that accounts for the dependence of observations.

Many different types of parameters are present within MLMs including fixed effects at each level, random effects, their variance, and possible covariance components, as well as standard error estimates of each parameter. To illustrate the different estimates in a MLM, consider the following model for continuous outcomes as outlined in Raudenbush and Bryk (2002) notation,

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}X_{1ij} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}W_{1j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_{1j} + u_{1j}, \end{aligned} \tag{1}$$

where Y_{ij} is a continuous outcome variable for the *i*th observation in the *j*th cluster, X_{1ij} is the value of the level-1 predictor for the *i*th observation in the *j*th cluster, r_{ij} is the level-1 residual for the *i*th observation in the *j*th cluster, γ are fixed effect estimates, W_{1j} is the value of the level-2 predictor for the *j*th cluster, $u_{0\mathrm{j}}$ and $u_{1\mathrm{j}}$ are the random effects whose variance and covariance are housed in the $\boldsymbol{\tau}$ matrix, and the *j* clusters are considered to be a sample from the larger population. Each individual observation will have a potentially different value for level-1 predictors as noted by the presence of both an *i* and a *j* subscript for X_1 above. On the other hand, all observations within a particular cluster will have an identical value for the level-2 predictor and random effects as noted by W_1, u_0 and u_1 having only a *j* subscript, meaning that these estimates are at level-2 and they are shared amongst all level-1 observations within a particular level-2 unit. More detail on each of these estimates and model estimation is provided shortly.

Goals and Research Questions

The first goal of this paper is to increase awareness of minimum sample size requirements for MLMs among researchers. Recently, studies have addressed how small sample sizes affect estimates in MLMs. Although this work is vitally important in applied research, the literature is highly diffuse and somewhat sparse. That is, MLMs are employed in a variety of disciplines under many different names, so the resources that do exist may be difficult to locate even when sought. Additionally, the literature is almost uniformly based either in simulation or mathematics, which may be overly technical for some audiences and thus less accessible. Our intent is to gather and summarize this information into a single location with minimal technical detail.

Second, the effect of small sample sizes on different model estimates are delineated with an emphasis on the number of clusters. Not all model estimates in MLMs are equally affected, and not all model estimates are always of interest for all research questions. The minimum recommended number of clusters for various model estimates is discussed.

Third, a simple illustrative simulation demonstrates the effects associated with small numbers of clusters. Because studies reporting the results of simulation studies are typically targeted at methodologists and statisticians and may contain many complex conditions, they can be overly technical for non-methodological researchers. A more simplified simulation example with many fixed conditions is presented to isolate how model estimates are affected with small numbers of clusters.

Fourth, barriers to obtaining the recommended number of clusters are omnipresent such as financial limitations, the use of extant data sets, or difficulties in recruiting large numbers of participants. Therefore, possible remedies are provided that may reduce the bias of estimates with small samples if increasing the sample size is not feasible or practical.

Methods

To locate the most current recommendations and findings, we reviewed methodological and statistical journal articles pertaining to sample size in MLMs. To find these studies, combinations of terms including “multilevel modeling,” “HLM,” “small sample size,” “mixed model,” “small number of clusters,” and “bias” were searched. Because MLMs are employed in a variety of disciplines, we used PsycInfo as a social science-specific database and Google Scholar and Google as broader search engines. We also searched the archives of the JISCmail Multilevel Listserv for potential resources as the topic is occasionally raised on the Listserv.

No specific inclusion criterion for the year the article was published was adhered to, but a vast majority of articles have been published in the last 10 years. We excluded articles focusing solely on single-case or “singleton” designs because they are a special, extreme case that possess unique features which may have resulted in recommendations that were not applicable to non-single case designs which are the focus for this paper. Although the literature search initially was not restricted to two-level models, only a single methodological article investigating the effects of sample size beyond two levels was found (Heo and Leon 2008). Therefore, this inquiry can only be applied to two-level models. Some studies referenced unpublished manuscripts, but these were not included if the original source itself could not be located. In total, there were seven unpublished manuscripts or conference presentations mentioned in reference sections of located resources, but the original source for five could not be located. Although the number of clusters was the primary manipulation of interest, other factors were also manipulated in many studies including sample size within each cluster (cluster size), the intraclass correlation (ICC), and the measurement scale of the outcome and predictors variables (e.g., binary, continuous).

In total, the review included 20 studies (Austin, P.C 2010; Baldwin and Fellingham 2013; Bell, B.A et al. 2014; Browne and Draper 2006; Clarke 2008; Cohen 1998; Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard 2009; Hox et al. 2012; Konstantopoulos 2010; Kreft 1996; Maas, C and Hox 2004; Maas and Hox 2005; McNeish 2014; Meuleman and Billiet 2009; Moineddin, Matheson, and Glazier 2007; Mok 1995; Paccagnella 2011; Scherbaum and Ferreter 2009; Snijders and Bosker 1993; Stegmueller 2013). Of these 20 studies, three focused solely on binary outcomes, 14 solely on continuous outcome, and three featured both binary and continuous outcomes. It is also important to mention that, although other simulation studies not included in this list manipulated the number of clusters, their motive for doing so was to investigate the number of clusters as a moderator rather than the primary focus of the study. The aforementioned articles feature the number of clusters as a primary research focus or provide substantial discussion pertaining to the number of clusters.

Within these 20 studies, four focused on mathematical issues and 16 reported on simulation study results. Although the mathematically focused studies provide core insight, the simulation studies are emphasized because they provide a wealth of illustrations and values regarding the effect of a small number of clusters on model estimates. Subsequently, an illustrative simulation using M*plus* version 7.1’s Monte Carlo feature is provided to clearly and simply provide a numerical demonstration of the isolated main effect that a small number of clusters has on MLM estimates.

Findings and Recommendations

The effect of the number of clusters on model estimates has been found to be moderated by design elements such as the ICC, the sample size within clusters, the scale of the outcome measure (binary or continuous), and the balance of the design (Browne and Draper 2006; Konstantopoulos 2010; Moineddin, Matheson, and Glazier 2007; Scherbaum and Ferreter 2009). Whether and how the number of clusters affects each of the model estimates is summarized below in the next three sections. Unless otherwise stated, it will be assumed that the outcome variable was measured on a continuous scale and that the level-1 error structure was specified as having an independent structure, meaning the variance elements on the diagonal of the level-1 covariance matrix are all equal while all covariance elements on the off-diagonal are zero. This structure is often used for cross-sectionally clustered data and is the default in popular programs for MLMs such as HLM and SAS Proc Mixed. A separate discussion for the role the number of clusters plays in estimates in models specifically with binary outcome variables is provided separately.

Table 1 shows a summary of the number of recommended clusters based on simulation studies depending on which estimates are of interest to the researcher. Table 1 is based solely on obtaining unbiased estimates and does not take statistical power into consideration. The rationale behind the recommendations in Table 1 will be covered in-depth in next two sections. These recommendations are for models estimated with maximum likelihood methods (either full maximum likelihood or restricted maximum likelihood) and without any small sample size adjustments. Although all possible values were not investigated, the findings below should generalize fairly well for ICC values between 0.10 and 0.30 and cluster sizes between 5 and 30. Different recommendations are provided for level-2 variance component estimates depending on whether full maximum likelihood (FML) or restricted maximum likelihood (REML) is used.

Both FML and REML estimation methods produce similar estimates with large numbers of clusters, but they diverge when the number of clusters is small, with REML providing less biased estimates of variance components. Some software programs allow users to employ either estimation method (e.g., HLM, SAS Proc Mixed) while other software programs offer only FML (e.g., M*plus*, SAS Proc Nlmixed).

To briefly touch on the difference between the two, FML estimates all parameters (fixed effects and variance components) in the model simultaneously whereas REML estimates the fixed effects and the variance components separately such that the estimation of the fixed effects does not interfere with variance component estimation (Raudenbush and Bryk 2002, p. 53; Searle, Casella, and McCulloch 2006). The difference in variance component estimates is analogous to calculating a sample variance: FML is similar to dividing by the sample size n whereas REML is analogous to dividing by $n-1$. When n is large, the difference is barely detectable, but, for small n, the discrepancy is much more noticeable. Thus, using FML may result in different conclusions than REML when the number of clusters is small. Snijders and Bosker (2012) offer that, when the number of clusters is greater than 50 plus the number of level-2 predictors, the difference between FML and REML is inconsequential. Raudenbush and Bryk (2002) also note that the variance components from FML can be multiplied by $(J-F)/J$ to approximate what would be obtained with REML where J is the number of clusters and F is the total number of fixed effects (the number of γ parameters). As an alternative to likelihood based methods, Bayesian Markov Chain Monte Carlo (MCMC) estimation methods and software are becoming increasingly available and useful for small sample size models (Baldwin and Fellingham 2013; Gelman et al. 2013; Hox et al. 2012; Stegmueller 2013).

Table 1 Minimum number of clusters recommended for accurate estimates by effect (assuming likelihood estimation methods and no small sample size adjustments)

Effect of interest	Continuous outcomes	Binary outcomes[a]
Level-1 fixed-effect point estimates	5	10
Level-2 fixed-effect point estimates	15	30
Fixed-effect standard errors	30	50
Level-1 variance estimate	10	30
Level-2 variance estimate	10/30 (REML/FML)	10/50 (REML/FML)
Level-2 variance standard error	50	100

[a] In addition to a minimum number of clusters, a cluster size greater than 5 is also recommended for binary outcomes

Outcome Measures

Relative bias is a simulation outcome that quantifies the difference between the estimated value for a parameter and the population value specified in the simulation design. More specifically, relative bias is calculated by

$$\text{Relative bias} = 100 \times \left(\frac{\sum \left(\widehat{\theta}\text{-}\theta \right)}{\theta} \right) / R \tag{2}$$

where $\widehat{\theta}$ is the estimate of the parameter of interest, θ is the population value for the same parameter, and R is the number of replications in the simulation.

The 95 % non-coverage rate is a simulation outcome measure that records the percentage of replications whose 95 % confidence interval does not contain the specified population value. That is, for each replication in the simulation, a confidence interval is estimated for each parameter. If the population value set in the simulation is included within the estimated confidence interval, then the replication is said to be "covered." If estimates are unbiased, the percentage of replications not covered by the 95 % confidence interval should be near 5 %. Inappropriate 95 % non-coverage rates may suggest that the point estimate is biased (i.e., the interval is centered around a biased value) or that the standard errors are estimated with bias (i.e., the interval is too narrow or too wide).

Because not all studies applied a common threshold for determining when estimates were biased, uniform criteria set forth by Bradley (1978) and Hoogland and Boomsma (1998) were imposed across all studies. Bradley (1978) suggests that values within one-half of the nominal type-I error rate be considered acceptable for 95 % non-coverage rates. This indicates that non-coverage rates less than 2.5 % or greater than 7.5 % are unacceptable. Hoogland and Boomsma (1998) suggest relative biases with absolute values lower than 10 % as acceptable for standard errors[1] and absolute values lower than 5 % as acceptable for point estimates.

Fixed-Effect Point Estimates and Associated Standard Error Estimates

Fixed-Effect Point Estimates

Fixed-effect point estimates have a very similar interpretation as regression coefficients in single-level models where a one-unit increase in the predictor yields a γ amount of change in the outcome variable (for continuous outcomes). Predictors in MLMs can enter the model either at level-1 or level-2. At level-1, each observation (e.g., a student within a classroom) will have a potentially unique value on the predictor variable. At level-2, each cluster (e.g., a classroom) will have a unique value, but observations within a cluster will all share the same value of the level-2 predictor. Level-2 predictors may be a characteristic of the cluster itself (e.g., if a school is public or private) or may be a variable aggregated over level-1 (e.g., average test score for a classroom).

[1] On a more technical note, population values for standard errors cannot be directly set within a simulation design, so other values must be used to assess the variability of estimates in the population (of which the standard error is an estimate). Although there are different values that can be used for such a purpose, the prevailing technique in the reviewed studies was to use the variability of the parameter estimates across replications. Because the same technique was implemented across studies, it is reasonable to compare these values across studies.

Fixed-Effect Point Estimate Simulation Findings The point estimates for the fixed effects were the least dependent of the model estimates on the number of clusters. Fixed-effect point estimates associated with predictors at either level are unbiased with 30 clusters and remain unbiased with as few as 15 clusters (Baldwin, S.A and Fellingham 2013; Bell et al. 2014; Maas, C and Hox 2004, 2005). Whereas the fixed effects associated with predictors at level-1 continue to be unbiased with even smaller numbers of clusters, fixed-effect estimates associated with level-2 predictors (including cross-level interactions) tend to be overestimated when the number of clusters falls below 15 (Baldwin and Fellingham 2013; Stegmueller 2013). The main effect of other factors such as ICC values and cluster size was not found to affect the bias, or lack thereof, of the fixed effect point estimates nor did their interaction with the number of clusters have an impact on bias.

Fixed-Effect Standard Errors

The standard error reflects the precision with which a fixed effect is estimated. When inferentially testing a fixed effect estimate, the standard error of the fixed effect estimate appears in the denominator of a *t*test or *Z*-test while the point estimate lies in the numerator. Because the fixed effect point estimate is largely unbiased (as reported above), if the standard error is underestimated, the denominator will be too low, resulting in spuriously high *t*values (or *Z*-values) which inflate type-I error rates. As a result, effects that are null in the population may be erroneously declared significant more often than the nominal type-I error rate (α) would stipulate. For applied researchers, the difference between significance and insignificance could affect policy recommendations that, in turn, may have social, medical, or financial consequences.

Fixed-Effect Standard Error Simulation Study Findings When the number of clusters is small, prior research has found that the resulting standard error estimates will be downwardly biased (i.e., underestimated) with standard estimation techniques. Thirty clusters have been shown to provide fixed-effect standard error estimates without bias (Maas, C and Hox 2004, 2005). Maas and Hox (2005) ran one condition with ten clusters and 5 units within each cluster to examine the effect of extremely small sample sizes. When only ten sparse clusters were simulated, the non-coverage rate of the 95 % confidence interval for fixed-effect estimates approached 10 %, far exceeding criteria in Bradley (1978). The standard errors of level-2 fixed effects required at least 30 clusters to produce unbiased estimates when estimated with standard REML in Maas and Hox (2005), and Stegmueller (2013) recommends at least 20 clusters to yield unbiased standard errors for cross-level interactions. This shows that ten clusters is inadequate with standard estimation procedures if hypothesis tests of the fixed effects are of interest to the researcher because type-I error rate is essentially twice the nominal rate. However, Baldwin and Fellingham (2013), Ferron et al. (2009), and Bell et al. (2014) found no bias for the standard error estimates of any fixed-effect estimates (level-1, level-2, within-level interactions, cross-level interactions) with less than 30 clusters and even with as few as four clusters in Ferron et al. (2009) when applying a Kenward–Roger adjustment. Kenward and Roger (1997, 2009) showed by proof, simulation, and real-data example that their denominator degrees of freedom and covariance matrix adjustment can greatly reduce the underestimation of fixed-effect standard errors when the number of clusters is small in continuous outcome models with REML. For instance, in one of their simulation studies, for a random coefficient model with an ICC of 0.20 and 24 clusters with nine repeated measures (cluster size of 9), the type-I error rate for fixed effects was reduced to 5.4 % from 12.9 % after applying their procedure.

With the unbiased fixed-effect standard errors reported in the aforementioned studies, the Kenward–Roger adjustment seems valuable for fixed-effects inference with a small number of clusters. The Kenward–Roger adjustment can be implemented in SAS Proc Mixed and with some estimation procedures in SAS Proc Glimmix and will be discussed in more detail in the concluding sections of this paper. Due to the computational complexity of the adjustment, the availability of the Kenward–Roger adjustment is mostly limited to the SAS software. The pbkrtest package in R can perform the Kenward–Roger adjustment on some models, but the package's documentation does warn users that the functionality has not been thoroughly tested (Halekoh and Højsgaard 2012).

The above studies focused mainly on continuous predictors, although binary predictors function similarly for most cases. However, when the prevalence of a binary predictor is highly discrepant (e.g., 90 % of values fall in a single category), standard errors will exhibit more bias, especially if included in an interaction. Bell et al. (2013a, b) found standard error estimates to be inflated with highly discrepant prevalence (i.e., 20 % or below for one response category) even when using the Kenward–Roger adjustment. When the highly discrepant binary predictor was part of an interaction, especially with another binary variable, standard errors did not become unbiased based on the criteria outlined in this paper until approximately 60 clusters were obtained.

Variance Component Point Estimates and Associated Standard Error Estimates

Level-1 Variance Point Estimate

Level-1 variance estimates measure the amount of variance in the outcome within clusters that is not explained by the specified model. Alternatively, the level-1 variance can be interpreted as the variance of within-cluster residuals. This value is of interest because it is directly included in the calculation of ICC values and can also be a useful metric in model building because relatively lower values indicate that the model is explaining more variance at the level of the outcome (i.e., level-1). The ICC for continuous outcomes with an independent error structure is calculated by

$$\frac{\tau_{00}}{\tau_{00}+\sigma^2} \tag{3}$$

where τ_{00} is the level-2 variance component estimate for the intercept in the unconditional model and σ^2 is the level-1 variance component estimate for the unconditional model. Therefore, bias in the level-1 variance component estimates could affect ICC calculations or perceptions of relevant predictors in model building.

Simulation Study Findings The point estimates for the level-1 variance are minimally affected by sample size at either level (Browne and Draper 2006; Maas, C and Hox 2004, 2005; Meuleman and Billiet 2009; Stegmueller 2013). Maas and Hox (2005) found the bias in the point estimates for level-1 variance to be less than 0.05 % across all sample size conditions (the smallest total sample size condition was 150), exhibiting a negligible amount of bias. Furthermore, Browne and Draper (2006) found bias less than 1 % with as few as six clusters for both FML and REML (the smallest total sample size condition was 108). Standard errors of the level-1 variance can be estimated in MLMs, but inferential tests are rarely of any practical interest, so they are often not reported in simulation or applied studies.

Level-2 Variance Estimates and Associated Standard Error Estimates

Level-2 Variance Component Point Estimates Level-2 variance components are used to assess the variability of the random effects in the model. Random effects provide an estimate for how an individual cluster deviates from the fixed effect. That is, the fixed effect captures the mean intercept or slope across all clusters while the random effect captures the deviation of a particular cluster from the mean. The random effects are often defined to have a mean of 0, so that each cluster will have a random effect (which may be estimated to be 0 by chance), and, across all the clusters, the random effects will be centered around the fixed effect. Often of interest is the amount of variance the random effects exhibit for a particular predictor variable (or the intercept) in the model. If clusters show very little variance around the fixed effect, then the particular random effect may not be necessary in the model.

Because evaluation of the ICC is a primary method for researchers to determine whether a MLM is necessary, among other uses such as in the design effect or evaluating the performance of level-2 predictors through the reduction in variance, misestimated point estimates for the level-2 variance would contribute to incorrect ICC values because the level-2 variance appears in both the numerator and denominator of the ICC formula. Bias in the level-2 variance point estimate could lead to misguided conclusions about the informativeness of the clustering or incorrect selection of models such as using a single-level model when a MLM would yield more appropriate inferences.

Level-2 Variance Component Point Estimate Simulation Study Findings Maas and Hox (2005) found that level-2 variance components were estimated with upward bias up to 25 % with ten clusters each of size 5. Browne and Draper (2006) compared REML and FML and found that REML estimates of the level-2 variance produced negligible bias with as few as six clusters with and an average of 18 units per cluster. The conflicting findings between Maas and Hox (2005) and Browne and Draper (2006) with REML may be attributable to the different cluster sizes. Clarke (2008) and McNeish (2014) have found that small cluster sizes often result in overestimated level-2 variance components, and the overestimation worsens as the number of clusters decreases with REML.

On the other hand, FML showed large amounts of downward bias with a small number of clusters (Browne and Draper 2006). When the number of clusters falls below 30 with FML, level-2 variance estimates exhibit downward bias in excess of 20 % with six clusters, resulting in ICC estimates that may be inaccurate. Similarly, Meuleman and Billiet (2009) found downward bias of 10 % with 20 clusters in a MLM estimated in the structural equation modeling (SEM) framework which uses FML.

Level-2 Variance Component Standard Errors Depending on the software program used to implement the analysis, the standard errors of the level-2 variance estimate are sometimes used in *Z*-tests to inferentially test that the level-2 variance components are null in the population. It should be noted, however, that this method does evoke some issues, since the variances are often constrained in MLMs to be non-negative and are thus bounded below by 0, putting the null value on the boundary of the parameter space (Molenberghs and Verbeke 2004; Stram and Lee 1994). Alternative methods including using a likelihood ratio test, 50:50 mixture χ^2

distribution, or information criteria may be more appropriate (Fitzmaurice, Laird, and Ware 2012; Molenberghs and Verbeke 2004). For MLMs estimated in an SEM framework, variance estimates are not always constrained to be non-negative, resulting in what are known as Heywood cases. Standard *Z*-tests may be appropriate in such situations, since the parameter space is no longer bounded and the null value is no longer on the boundary of the parameter space. Interested readers are referred to Savalei and Kolenikov (2008) for a more thorough discussion of constrained versus non-constrained variance component estimates.

Level-2 Variance Component Standard Error Simulation Study Findings As a function of the number of clusters, the standard error of the level-2 variance is the most affected of all the estimates. As a more technical note, standard errors of the level-2 variance components are a fourth-order estimator, meaning that a high volume of data is required to obtain unbiased estimates (Raudenbush and Bryk 2002).

Although a likelihood ratio test or 50:50 mixture χ^2 are preferred methods for a hypothesis test of the standard errors, a *Z*-test can also provide inferential information and is reported by some popular software programs such as M*plus* and SAS (Raudenbush and Bryk 2002). The *Z*-test divides the variance component point estimate by its standard error, so, if the standard errors are underestimated, the type-I error rate will be inflated, leading to more null hypothesis rejections than the nominal rate specifies, resulting in the retention of more variance components, and ultimately more complex models than may be necessary. For more detailed information on the hypothesis tests of variance components, the interested reader is referred to Raudenbush and Bryk (2002), pages 61 to 65.

With ten clusters, Maas and Hox (2005) found non-coverage rates of the 95 % confidence interval for the level-2 variance standard errors to approach 30 %, six times the nominal rate (the large non-coverage rate may be partially attributable to the upward bias in the point estimate rather than purely to underestimated standard errors). With 30 clusters, the level-2 variance components have been found to be estimated with a non-coverage rate around 9 % for both the level-2 variance of both slopes and the intercept, a rate that continues to exceed criteria in Bradley (1978). More disconcerting, Maas and Hox (2005) found that, even with 50 clusters and a cluster size of 30, non-coverage rates frequently exceeded 8 % for the level-2 variance with REML estimation. With FML based on the SEM framework, Meuleman and Billiet (2009) found that the non-coverage rate of the level-2 variance exceeded 9 % even with 80 clusters. This leads to the recommendation that, if hypothesis testing of the level-2 variance components is of interest via a *Z*-test, then a minimum of 50 clusters with a cluster size of 50 are suggested with 100 clusters being a more conservative figure, especially if FML estimation is utilized.

No studies we reviewed had investigated covariance between variance components at level-2. Currently, no recommendations can be made regarding how covariance estimates are affected by the number of clusters.

Because selection of level-2 variance components play a vitally important role in multilevel modeling, downwardly biased standard errors resulting in inflated type-I error rates lead to spurious significance, which may render an entire model or way of understanding a phenomenon to be inherently flawed. Researchers with a small number of clusters should therefore be particularly cautious with level-2 standard error estimates. *Z*-tests for variance components should generally be avoided, but they are especially untrustworthy with a small or even moderate number of clusters. Likelihood ratio or 50:50 mixture χ^2 tests are consequently recommended in all situations but particularly with a small number of clusters.

Binary Outcome Variables

Many of the concerns with continuous outcomes are magnified with binary outcomes. When the outcome variable was binary (i.e., requiring the use of multilevel logistic regression), in addition to small numbers of clusters being problematic, cluster size played a much larger role with binary outcomes than with continuous outcomes. Even with a very large number of clusters, conditions with a cluster size of five or fewer produced estimates that exhibited much more bias than their counterparts with continuous outcomes (Austin, P.C 2010; Clarke 2008; Moineddin et al. 2007).

Another difficulty encountered with binary outcomes is that estimation of the model is less straightforward. With binary outcomes, the likelihood does not have a closed form, since the outcome is linked to the predictor variables with a non-linear function, namely the logit function. As a result, MLMs with binary outcomes are often estimated in software with either numerical integration methods such as Adaptive Gaussian Quadrature, penalized quasi-likelihood (PQL), or with linearized pseudo-likelihood methods. The choice of estimation method is less straightforward than with continuous outcomes since Adaptive Gaussian Quadrature has been found to provide a more accurate approximation (see, e.g., Kim, Choi, and Emery 2013), but it approximates FML which is known to be biased with a small number of clusters. On the other hand, pseudo-likelihood approximations are typically less accurate but are much faster and have an analogue to REML. When selecting an estimation method with small samples, it is important to note that the Kenward–Roger adjustment can only be implemented with pseudo-likelihood estimation. Very little research has been conducted on the differences between estimation methods with small samples with binary outcomes, and future research could serve to better understand the properties of each estimation method with small samples.

Fixed-Effect Point Estimate and Associated Standard Error Estimate Simulation Findings

The point estimates for fixed effects have been found to be largely unbiased regardless of the number of clusters, as with continuous outcomes (Austin, P.C 2010; Clarke 2008; Moineddin et al. 2007; Paccagnella 2011). However, when cluster size was small (less than 10), noticeable bias was present. Moineddin et al. (2007) found relative bias between 4 % and 11 % in level-1 fixed-effect point estimates and between 6 % and 16 % for level-2 fixed-effect point estimates in their 30-cluster, 5-cluster-size condition. In their 30 cluster, 30-cluster-size condition, bias was negligible. Standard error estimates showed issues similar to those found with continuous outcome variables where fewer than 30 clusters with 30 units presented larger amounts of bias. Fixed-effect standard errors in the 50 cluster, 5-cluster-size condition in Moineddin et al. (2007) also displayed non-negligible amounts of bias whereas no such bias was found in the analogous case with continuous outcomes, showing the increased prominence of cluster size with binary outcomes.

Level-2 Variance Component Point Estimate and Associated Standard Error Simulation Findings

When estimating the level-2 variance components for binary outcomes with Adaptive Gaussian Quadrature (which approximates FML), Moineddin et al. (2007) found that the estimates were not biased when the cluster size and the number of clusters were both above 30. When cluster size was below 30, estimates were often highly positively biased with bias values ranging from 2 % to 174 % across the spectrum of simulated conditions with smaller number of cluster conditions producing larger bias.[2] Clarke (2008) found similar results using the same estimation method with bias in the level-2 variance components exceeding 100 % for conditions with 200 clusters and cluster sizes of 2 and 5.

Additionally, Austin, P.C (2010) included a wider variety of estimation methods. With FML approximation methods, estimates exhibited meaningful bias until 30 clusters were obtained. Results from REML analogs (e.g., residual pseudo-likelihood and its various forms, broadly referred to a RPL) that take the estimation of the fixed effects into account (similar to REML for continuous outcomes) performed better with a small number of clusters. RPL methods as implemented in SAS Proc Glimmix or PQL as implemented in the HLM software program consistently provided unbiased variance component estimates at about the 10 cluster, 10-cluster-size and 10 cluster, 15-cluster-size conditions, respectively, for the fairly simple model tested in Austin, P.C (2010). Interestingly, the glmmPQL function in R similarly implements PQL as is done in HLM but did not attain unbiased estimates for any condition investigated. Bayesian MCMC estimation also provided unbiased level-2 variance component estimates in the 7 cluster, 10-cluster-size condition using a diffuse inverse gamma prior[3] for the level-2 variance component.

For standard error estimates of level-2 variance components, 100 clusters with cluster size of 50, the largest sample size simulated, showed some borderline evidence of downward bias in Moineddin et al. (2007). Ninety-five percent non-coverage rates for the level-2 variance components fell between 6.7 % and 8.6 % with 100 clusters of 50 units.

For all estimates mentioned above, the prevalence rate of the outcome affected the degree of bias. For example, when the prevalence rate of the outcome was 10 % compared with 45 %, the bias was greater (Moineddin et al. 2007). As the frequency between the two levels of the outcome becomes more disparate, larger sample sizes are required to produce unbiased estimates.

Model Convergence

Another issue that is more problematic with binary outcomes is model convergence (Moineddin et al. 2007; Paccagnella 2011). That is, whether model estimates will be able to be produced. For multilevel logistic regression, a small number of clusters were found to decrease model convergence rates, although small cluster sizes seemed to be the primary force behind non-converging models (Moineddin et al. 2007). With 30 clusters, a cluster size of 5, and an ICC of 0.04, for instance, model convergence was achieved only 56 % of the time. Once 50 clusters were obtained with a cluster size of 30, model convergence rates were consistently above 87 % regardless of the ICC. Convergence rates were also affected by prevalence rates where greater discrepancies between outcome categories yielded lower convergence rates.

[2] It is important to note again that the level-2 variance component will be included in the calculation of ICC values. The level-1 variance component is calculated differently than with continuous outcomes and Goldstein, Browne, and Rasbash (2002) discuss four methods for its calculation. Most commonly $\frac{\pi^2}{3}$ or about 3.29 is substituted for the level-1 variance, since this is the variance of the logistic distribution when the scale is set to 1 with a location of 0. Other methods include simulation and Taylor series expansions. The associated problems with misestimated ICC values are the same as presented with continuous outcomes.

[3] A diffuse inverse gamma prior is common used when a researcher wants to utilize Bayesian methods but wants to limit the impact of the prior distribution on the posterior distribution. This is the default prior distribution for variances for more user-friendly Bayesian software programs such as M*plus*.

Simple Simulation

To demonstrate the effect small sample sizes have on estimates, we conducted a simple simulation for illustrative purposes. While the studies reviewed above manipulated several conditions, we intentionally kept our simulation simple by only manipulating two conditions. Although the reviewed simulation studies are thorough, the take-home message may be difficult to disentangle for some researchers not fluent with simulation studies because of the number of cells in the simulation design that result from multiple manipulated factors. We therefore only manipulated two conditions in our simulation to provide an illustration of the above synthesis that will hopefully allow for a more visual depiction of how MLM estimates are affected as the number of clusters changes. The simulation uses FML without any of the small sample adjustments mentioned previously or that are elaborated upon in the concluding section of this paper. This was done to show the "worst case" scenario of estimating MLMs without considering the effect of a small number of clusters.

Generating Model

We simulated a completely balanced random intercept model with only one level-1 predictor variable and a continuous outcome. Using notation outlined in Raudenbush and Bryk (2002), the model is as follows:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}X_{1ij} + r_{ij} \\ \beta_{0j} &= u_{0j} \\ \beta_{1j} &= \gamma_{10} \end{aligned} \tag{4}$$

The population value for γ_{10} was 0.25; level-1 (or residual) variance was 1.00; variance of the random effect (u_{0j}) was 0.20, and the population ICC was thus 0.17. The outcome measure was standardized so γ_{00} is not included, since it is equal to 0. We did not manipulate any of these design elements throughout the simulation, so it is important to note that the results of this simulation should *not* be considered on par with the studies reviewed earlier and is solely for illustrative purposes and should not replace recommendations above.

The ICC was chosen to be 0.17 based on findings from Hedges and Hedberg (2007) which reported typical ICC in applied behavioral research, with ICC values ranging from 0.045 to 0.271. Our value of 0.17 was chosen to fall about halfway between the two extremes to approximate the typical ICC one might encounter in applied behavioral research with clustered data.

We varied the number of clusters from 5 to 1,000. We had two conditions for cluster size, 20 and 50, which were fully crossed with the number of cluster conditions. We then briefly summarize our findings for the same estimates as in the synthesis above. Similar to the studies reviewed above, the percentage bias and the 95 % non-coverage rate will be tracked. To facilitate reporting of results, because the percentage bias values are expected to be negative, we report "percentage underestimated" rather than relative bias where $\text{Percentage Underestimated} = (-1 \times \text{Percentage Bias})$. This was simply done to allow for both outcome measures to be presented simultaneously, given that they would share a common scale. For percentage underestimated calculations of standard errors, the standard deviation of the parameter estimates across replications (a.k.a. the empirical standard error) was used as the population standard error for the reason that the population standard errors cannot be directly specified in a simulation design.

Fixed-Effect Point and Standard Error Estimates

The number of clusters and cluster size had no effect on the fixed-effect point estimates even with as few as five clusters in our illustrative simulation. No figure for point estimates is therefore provided.

In Fig. 1, each plot represents a different cluster size while the number of clusters is on the horizontal axis, and each simulation outcome measure is represented by a different line, although they share the vertical axis because they have a similar metric. Reference lines for the cut-off criteria (i.e., 0.075 for non-coverage rate and 0.10 for percent underestimated) are included as grey lines. The solid grey line is a reference for non-coverage rate (represented by the corresponding solid black line), and the grey dashed line is a reference for percentage underestimated (represented by the corresponding dashed black line). Biased estimates occur when a black line exceeds a grey line.

As seen in Fig. 1, the standard error estimates for the fixed effect were unbiased with ten clusters based on percentage underestimated and about 20 clusters based on the 95 % non-coverage rate. The sharp increase in percentage underestimated and increase in the 95 % non-coverage interval with fewer than 20 clusters shows the problems associated with too few clusters when estimating the standard errors of the fixed effects without an adjustment such as Kenward–Roger.

Level-1 Variance Point Estimate

The Level-1 variance point estimates did not seem to be greatly affected by a small number of clusters in our illustrative simulation. No estimates were close to being biased for this parameter as the percentage underestimated was at most 0.30 %. Consequently, no plot is provided.

Level-2 Variance Point and Associated Standard Error Estimates

In Fig. 2, the results suggest that, for fewer than about 30 clusters with FML, the variance components begin to exhibit an unacceptable amount of downward bias. Using REML should

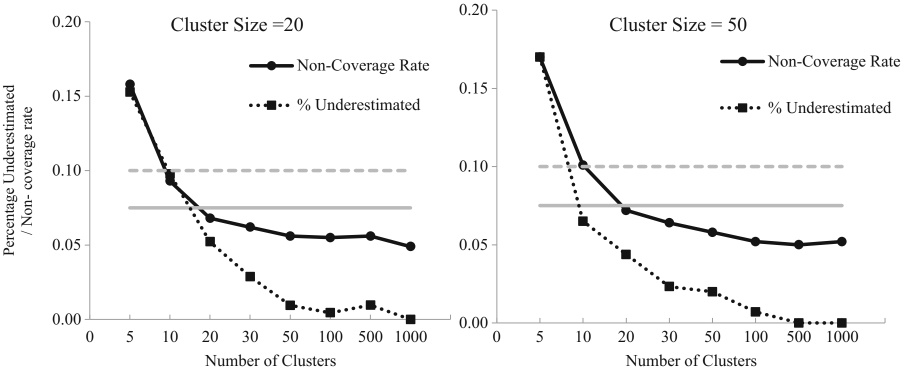

Fig. 1 Non-coverage and percentage underestimated for fixed-effect standard error estimate

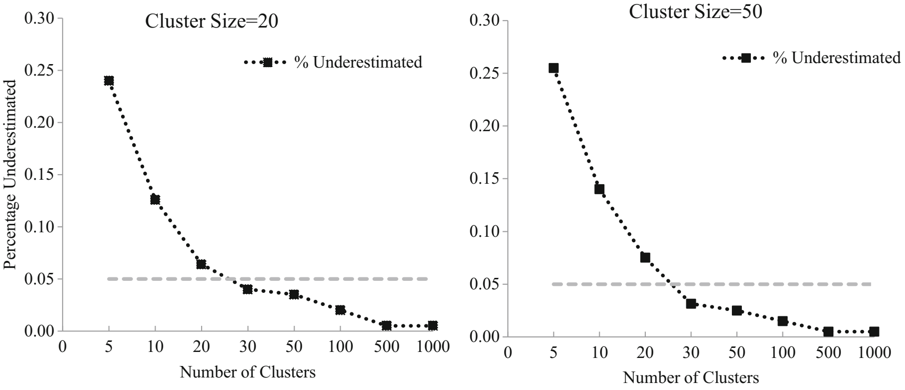

Fig. 2 Percentage underestimated for level-2 variance component point estimate

alleviate much of the concern with bias in the level-2 variance components with as few as about five clusters (Browne and Draper 2006; Ferron, J.M et al. 2009). Using FML will lead to biased level-2 variance component estimates with a small number of clusters, so researchers are advised to take note with which estimation method analyses are conducted.

As previously mentioned, standard error estimates of the level-2 variance components are the most affected model estimate when the number of clusters is small. The non-coverage rate of the 95 % interval for level-2 standard error estimate exceed criteria from Bradley (1978) even in the 100 cluster, 20-cluster-size condition, suggesting that 100 clusters may be the minimum needed to have reasonable confidence that the standard errors for the level-2 variance components are estimated without bias with FML and that inferences made from *Z*-tests are appropriate. Figure 3 below shows this information graphically.

Suggestions for Addressing Analyses with a Small Number of Clusters

For researchers who have already collected data, are using an extant dataset, or have financial or sampling limitations, there are some options that can address the bias associated with a small number of clusters. First, with continuous outcomes especially, REML estimation is

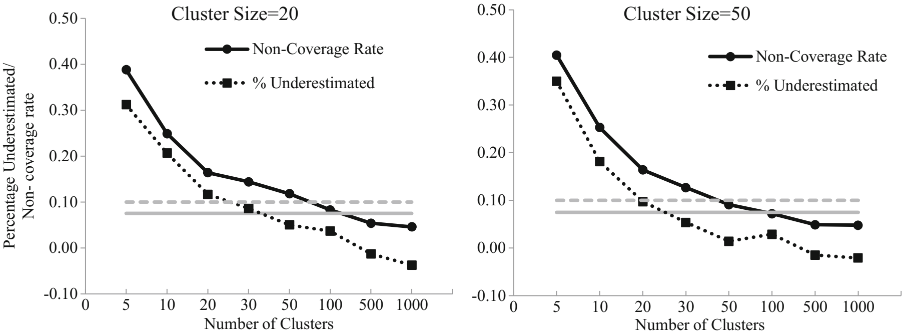

Fig. 3 Non-coverage rate and percentage underestimated for level-2 variance component standard error estimate

universally preferable to FML for unbiased variance component estimation.[4] With binary outcomes, very few studies have addressed differences in estimation methods, although those that have generally show that RPL variance component bias is less than methods that approximate FML. Second, one can use the Kenward–Roger adjustment (Kenward and Roger 1997) to guard against the inflated type-I error rate that results from underestimated fixed-effect standard errors. Although the computational formulae are beyond the scope of this paper, generally, the Kenward–Roger adjustment first inflates the covariance matrix that houses the fixed-effect standard errors to address downward bias with small sample sizes. Then, a Satterthwaite-type approximation (Satterthwaite 1946) is applied to the inflated covariance matrix to calculate more appropriate degrees of freedom for the hypothesis test that the fixed-effect estimate is equal to zero in the population (Kenward and Roger 1997). Bell et al. (2013a, b) recommended using Kenward–Roger as best practice to protect against biased estimates. Both of these options are available in Proc Mixed or Proc Glimmix with RSPL estimation as an option under the Model statement in the SAS software. As mentioned previously, software implementation of Kenward–Roger is primarily restricted to SAS. Several simulations have found the Kenward–Roger adjustment to perform well in a variety of scenarios with fewer than ten clusters. Such scenarios include a comparison to Bayesian MCMC estimation (Baldwin and Fellingham 2013), properties of inferential tests such as type-I error rates from Kenward–Roger compared with other competing methods (Kowalchuk et al. 2004) and for unbalanced data (Spilke et al. 2005).

Third, because the point estimates do not exhibit bias in most situations, one could use bootstrapping techniques to obtain a more accurate assessment of the sampling variability. Bootstrapping is a resampling procedure that calculates sampling variability based on a large number of draws from a distribution. Butar and Lahiri (2003) and Van der Leeden et al. (1997) have discussed and demonstrated the utility of bootstrapping with multilevel models with small sample sizes for continuous outcomes and González-Manteiga, Lombardía, Molina, Morales, and Santamaría (2007) demonstrated the effectiveness of bootstrap methods with binary outcomes and small sample sizes. This option is available in MLwiN for two-level analyses.

Alternatively, if the number of clusters is very low (e.g., 20 or less) and one is interested in hypothesis testing of the level-1 fixed effects but not the variance components nor any level-2 predictors, and the cluster sizes are large, researchers can dummy code each cluster in the data and include the clusters as predictors in a standard ordinary least squares regression model (Gardiner, Luo, and Roman 2009; Snijders and Bosker 2012). This approach will allow some information regarding cluster differences to be obtained or at least provide estimates that hold cluster level affiliation constant. If effect coding is used (i.e., 1 and −1 rather than 1 and 0) for cluster affiliation variables, then fixed-effect estimates similar to those obtained from MLMs can be obtained, since estimates will reflect the overall mean rather than the mean of a reference group.

Lastly, MLMs with small sample sizes could be estimated with Bayesian MCMC. MCMC does not utilize frequentist principles for estimation and, consequently, does not carry the same assumptions and properties, namely requiring large sample sizes for asymptotically unbiased estimates (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin 2013; Raudenbush and Bryk 2002). Although only a handful of studies have demonstrated this, Austin, P.C (2010), Browne and Draper (2006), and Stegmueller (2013) provided evidence that Bayesian estimates achieved unbiased estimates with lower numbers of clusters than likelihood methods even

[4] If model comparison is undertaken and models differ with respect to fixed effects, then FML must be used. Otherwise, the deviance will not be calculated appropriately with REML.

when fewer than ten clusters were present. Although not in the context of a simulation, Gelman (2006) used Bayesian MCMC to reasonably estimate a model for data with less than five clusters. One difficulty in using a Bayesian approach is the selection of prior distributions which will have more influence when the number of clusters is small (Gelman 2002, 2006). Prior distributions for variances can be most problematic due to the lower boundary at 0, although Gelman (2006) provided some evidence that uniform or half-*t* prior distributions with wide ranges for variance components might perform well with small sample sizes.

Future Research

Even though many properties of MLMs with small samples have been addressed in the literature, many important topics have yet to be investigated. First, all research up to this point has focused on two-level models. Three-level models are relatively common in educational psychology (e.g., students clustered within classrooms/schools clustered within schools/districts), and sample sizes can become increasingly small as one progresses upward through a hierarchy. For instance, if school districts are the third level of clustering, even though five or ten school districts could provide data on thousands or even tens of thousands of students, the small sample size at the third level could lead to biased estimates. In situations where the number of units is bounded by a fairly low number (e.g., counties or states), finite population correction factors may be necessary because the sampling error would otherwise be quite close to zero.

Second, the performance of different estimation methods with binary outcomes is still unclear, especially for models of the complexity used in applied research. Numerical integration methods that are widely used can require very large computational overhead to estimate, since the number of computations increases exponentially as the number of random effects increases.

Third, more studies on Bayesian MCMC estimation of MLMs are needed. In the few studies that have been conducted, MCMC methods show much promise to provide unbiased estimates with quite small sample sizes. One particular challenge that lingers is a comparison of different prior distributions for the level-2 variance components. Gelman (2006) provided some suggestions with an applied example, but a larger simulation study could more comprehensively address this issue.

Fourth, additional scales of outcome measures could be investigated. To date, only continuous and binary outcomes have been featured in simulation studies; however, count outcomes requiring multilevel Poisson or negative binomial regression have yet to be addressed in the literature. Similarly, generalizations of multilevel logistic regression for multiple nominal or ordinal outcomes (i.e., multilevel multilogit or multilevel cumulative logit models) have not been explored.

Finally, most studies reviewed in this paper considered simulation designs such that the sample size within each cluster was equal across clusters (i.e., the design was balanced; see Bell et al. 2014 and Konstantopoulos 2010 for examples of studies with unbalanced designs). Regarding the design of simulations, future studies are encouraged to include conditions where clusters have variable sizes, since data of this form are more the rule than the exception with small samples. Additionally, studies that compare balanced conditions to unbalanced conditions could also be useful, since most studies incorporate simulation designs that are balanced or unbalanced, but not both. As a result, it is difficult within the current literature to compare how the balance of the design affects the estimates.

Discussion and Conclusion

Although traditional maximum likelihood estimation methods for MLMs have been shown to provide biased estimates when the number of clusters is below 30, methods such as restricted maximum likelihood estimation, the Kenward–Roger adjustment, or Bayesian MCMC have shown potential to perform well with ten clusters or fewer in some scenarios. Researchers are encouraged to be aware of potential hazards with small samples in MLMs and to consider implementing more recently developed methods to unbiasedly estimate MLMs in such scenarios.

Researchers should be cautious not to be overly optimistic with these findings. Although methods and approaches have been developed to provide unbiased estimates with small samples, researchers must consider the implications of such analyses. For instance, unbiased estimates do not suggest any information regarding statistical power, and while standard errors estimates may be unbiasedly estimated, they may be too large to be informative for inferential purposes. Additionally, assumptions can be difficult to verify with small samples. As an example, to assess whether the random effects are reasonably normal as is often checked through Q–Q plots or a histogram, a level-2 sample of 10 or 20 would feature plots with a sparsity of data points, and definitive conclusions would be difficult to draw. Put more succinctly, even though estimates can be reasonably estimated with small samples, larger samples are still preferable when possible.

References

References marked by an (*) indicate they were included in the review

* Austin, P.C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *The International Journal of Biostatistics*, *6, Article 16*.

*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.

Bell, B., Ene, M., Smiley, W., & Schoeneberger, J. (2013). *A multilevel primer using SAS Proc Mixed, SAS Global Forum*.

Bell, Schoeneberger, Smiley, Ene, and Leighton (2013). *Doubly diminishing returns: an empirical investigation on the impact of sample size and predictor prevalence on point and interval estimates in two-level linear models.* Paper presented at the Modern Modeling Methods Conference (M3*)*. Storrs.

*Bell, B.A., Morgan, G.B., Schoeneberger, J.A., Kromrey, J.D., & Ferron, J.M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *10*, 1–11.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.

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

Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference, 112*, 63–76.

*Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single level models with sparse data. *Journal of Epidemiology and Community Health*, *62*, 752–758.

*Cohen, J. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics*, 14, 267–275.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., & Lee, R. (2009). Multilevel modeling: a review of methodological issues and applications. *Review of Educational Research, 79*, 69–102.

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

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*. Hoboken: Wiley.

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: what are the differences? *Statistics in Medicine, 28*, 221–239.

Gelman, A. (2002). Prior distribution. *Encyclopedia of Environmetrics.*

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton: CRC press.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, 1*, 223–231.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis, 51*, 2720–2733.

Halekoh, U., & Højsgaard, S. (2012). pbkrtest: parametric bootstrap and Kenward Roger based methods for mixed model comparison. URL http://cran.r-project.org/web/packages/pbkrtest/pbkrtest.pdf [accessed on 14 March 2014].

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

Heo, M., & Leon, A. C. (2008). Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics, 64*, 1256–1262.

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

Hox, J. J. (1998). Multilevel modeling: when and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Berlin: Springer.

Hox, J. (2010). *Multilevel analyses: techniques and applications* (2nd ed.). Mahwah, NJ: Erlbaum.

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

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

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

Kim, Y., Choi, Y. K., & Emery, S. (2013). Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages. *The American Statistician, 67*, 171–182.

*Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education,* 78, 291–317.

Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement, 64*, 224–242.

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

*Maas, C., & Hox, J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica.,*58,127-137.

*Maas, C.J., & Hox, J.J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 86–92.

*McNeish, D.M. (2014). Modeling sparsely clustered data: design-based, model based, and single-level methods. *Psychological Methods.* DOI: 10.1037/met0000024.

*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.

*Moineddin, R., Matheson, F.I., & Glazier, R.H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, 34.

*Mok, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter*, *7*, 11–15.

Molenberghs, G., & Verbeke, G. (2004). Meaningful statistical model formulations for repeated measures. *Statistica Sinica, 14*, 989–1020.

*Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *7*, 111–120.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Satterthwaite, F. E. (1946). An approximate distribution of the estimates of variance components. *Biometrics, 2*, 110–114.

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13*, 150–170.

*Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample size for organizational research using multilevel modeling. *Organizational Research Methods,* 12, 347–367.

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. Hoboken: Wiley.

*Snijders, T., & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics,* 18, 237–259.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

Spilke, J., Piepho, H. P., & Hu, X. (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics, 10*, 374–389.

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

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50*, 1171–1177.

Van der Leeden, R., Busing, F., & Meijer, E. (1997, April). *Applications of bootstrap methods for two-level models*. Paper presented at the Multilevel Conference. Amsterdam.