Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: www.tandfonline.com/journals/hmbr20

Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction

Daniel McNeish

To cite this article: Daniel McNeish (2017) Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction, Multivariate Behavioral Research, 52:5, 661-670, DOI: 10.1080/00273171.2017.1344538

To link to this article: https://doi.org/10.1080/00273171.2017.1344538

Published online: 17 Jul 2017.

MULTIVARIATE BEHAVIORAL RESEARCH
2017, VOL. 52, NO. 5, 661–670
https://doi.org/10.1080/00273171.2017.1344538

QUANTITATIVE METHODS IN PRACTICE: TUTORIAL

Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction

Daniel McNeish

University of North Carolina, Chapel Hill; Arizona State University

ABSTRACT

Studies on small sample properties of multilevel models have become increasingly prominent in the methodological literature in response to the frequency with which small sample data appear in empirical studies. Simulation results generally recommend that empirical researchers employ restricted maximum likelihood estimation (REML) with a Kenward-Roger correction with small samples in frequentist contexts to minimize small sample bias in estimation and to prevent inflation of Type-I error rates. However , simulation studies focus on recommendations for best practice, and there is little to no explanation of why traditional maximum likelihood (ML) breaks down with smaller samples, what differentiates REML from ML, or how the Kenward-Roger correction remedies lingering small sample issues. Due to the complexity of these methods, most extant descriptions are highly mathematical and are intended to prove that the methods improve small sample performance as intended. Thus, empirical researchers have documentation that these methods are advantageous but still lack resources to help understand what the methods actually do and why they are needed. This tutorial explains why ML falters with small samples, how REML circumvents some issues, and how Kenward-Roger works. We do so without equations or derivations to support more widespread understanding and use of these valuable methods.

KEYWORDS
Restricted maximum likelihood; Kenward-Roger; tutorial; explanation; mixed model

In behavioral science studies, clustered data arise quite frequently in the form of either cross-sectional clustering or longitudinal clustering (Raudenbush & Bryk, 2002). Cross-sectional clustering occurs when people belong to higher-level units, such as students clustered within schools in education or employees clustered within companies in business. Longitudinal clustering occurs when the same individuals are measured repeatedly over time such that observations are clustered within people. The clustered nature of such data violates the independence assumption posited by the general linear model; therefore, specialized statistical techniques are required to appropriately model such data. In behavioral sciences, the most common technique is *multilevel modeling*, which is also commonly referred to as *mixed effects modeling* or *hierarchical linear modeling*.[1]

In multilevel models (MLMs), sample size issues tend to be more prominent because the data feature multiple levels. For instance, if students are cross sectionally clustered within schools, there is a Level-1 sample size that refers to the number of students in the data and a Level-2 sample size that refers to the number of schools in the data. In MLMs, the Level-2 sample size is the most important factor for determining whether the model will be afflicted by small sample issues (Snijders & Bosker, 1993). In empirical studies, this tends to be problematic because the Level-2 sample size is the most difficult and financially costly to increase. Using the students within schools as example, the number of students may be over 1,000 but if they are clustered within 20 schools, the data (perhaps unintuitively) fall under the "small sample" umbrella.

With clustered data, small samples are quite common. An informal review of behavioral science meta-analyses by McNeish (2016) reported that about 33% of growth models, 20% of multilevel models, 40% of meta-analyses, and 30% of cluster randomized trials would be classified as small sample problems based on cutoffs suggested in the literature. Small sample issues for MLMs have not gone unnoticed, however. In fact, there has been a recent increase in methodological studies exploring small sample data and MLMs in the last decade beginning with the seminal study by Maas and Hox (2005), which was one of the first to explore sufficient sample sizes to use multilevel models. Recent studies have been conducted to explore how small researchers could take sample sizes and

CONTACT Daniel McNeish ✉ dmcneish@asu.edu Department of Psychology, Arizona State University, P.O. Box 871104, Tempe, AZ 85287, USA.

[1] There are some minor differences between these types of models but, for the purpose of this manuscript, we consider the models to be more or less interchangeable.

still trust their estimates (Bell, Morgan, Schoeneberger, Kromrey, & Ferron, 2014), a comparison of small sample corrections (McNeish & Stapleton, 2016a), a comparison of multilevel bootstrapping and small sample corrections (Huang, 2017), comparisons of Bayesian and frequentist methods (Hox, van de Schoot, & Matthijsse, 2012), and a review of small sample methods for MLMs (McNeish & Stapleton, 2016b).

These studies generally find that Level-2 sample sizes below 50 are susceptible to small sample biases, which include downwardly biased estimates of both the variance components and the fixed effect standard errors, resulting in inflated Type-I error rates for inference about fixed effects. Results show that data with Level-2 sample sizes below 25 will almost certainly encounter these issues if precautions are not taken. In the frequentist framework, restricted maximum likelihood (REML) estimation has been shown to have much improved small sample properties and continues to perform well with Level-2 sample sizes into the single digits (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; McNeish & Stapleton, 2016a). REML does not completely solve issues related to inflated Type-I error rates for fixed effects, so the Kenward-Roger correction (Kenward & Roger, 1997) has been shown to maintain nominal Type-I error rates and therefore its use has been recommended as best practice (e.g., McNeish & Stapleton, 2016b). This correction has also been recently made available in SAS (PROC MIXED and PROC GLIMMIX), Stata, and in the pbkrtest and lmerTest R packages.

These studies, along with popular textbook treatments of the topic, are undoubtedly informative and they tend to focus on *how* to model small sample data by providing *best practice* recommendations. This goal is helpful for empirical researchers who possess small sample data because making informed modeling decisions is vital to improving small sample analyses. However, these sources allocate little to no space to explain *why* small sample issues are a problem in the first place. For researchers looking to understand the mechanism underlying why MLMs fail with smaller sample sizes with traditional methods and why alternative methods for small samples are necessary, the existing methodological literature does little to delineate the issue beyond mentioning a few platitudes such as noting that maximum likelihood (ML) is asymptotic and therefore has finite sample bias or presenting the REML formulas—the intuition behind the source of the bias is often left unaddressed. This is not intended to reflect negatively on these sources as explaining the underlying mechanism is admittedly not their focus. Currently, if researchers wish to understand the root cause of small sample issues, they must go deep into the mathematical statistics literature where equations and Greek notation outnumbers English text. Again, this is not a criticism as mathematics is the language of newly proposed statistical methods. However, a treatment of the tenets of small sample issues that are accessible to empirical researchers encountering these issues and deciding on how to handle them has not yet appeared in the literature. Explaining this mechanism is the focus of the current manuscript.

To outline the remainder of this manuscript, we first introduce MLMs and the basic notation that we use. We assume that readers have some baseline familiarity with MLMs and why they are used. We transition to standard full ML for MLMs and explain why issues arise with small Level-2 sample sizes. Then, we show how REML differs from ML and why REML is less susceptible to small Level-2 sample sizes. We continue by noting that REML does not completely solve problems when the Level-2 sample size is small and we discuss the logic of the Kenward-Roger correction and its antecedent from Kackar and Harville (1984). An example analysis is provided to demonstrate how each successive small sample method reduces the reliance on asymptotic assumptions and, in the process, improves estimation and inference for small sample analyses.

Overview of multilevel models

Though we assume some familiarity with MLMs, they are extensively used across a wide span of disciplines, so many different sets of notation and terminology exist. In this paper, we use the set provided by Raudenbush and Bryk (2002) that is common in psychology and education. To clarify this notation, an MLM for a continuous outcome variable and one predictor at each level in this notation is written as

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

where

- Y_{ij} is the outcome variable for the ith person in the jth cluster;
- β_{0j} is the cluster-specific intercept for the jth cluster;
- β_{1j} is the cluster-specific slope for the jth cluster;
- X_{1ij} is the Level-1 predictor variable value for the ith person in the jth cluster;
- r_{ij} is the Level-1 residual for the ith person in the jth cluster;
- γ_{00} is the fixed effect for the intercept;
- W_{1j} is the Level-2 predictor variable value for the jth cluster;
- γ_{01} is the fixed effect coefficient for the effect of W_{1j} on β_{0j};
- γ_{10} is the fixed effect for the slope of X_{1ij};

- γ_{11} is the fixed effect coefficient for the effect of W_{1j} on β_{1j};
- u_{0j} is the random effect for the intercept for the jth cluster;
- u_{1j} is the random effect for the slope for the jth cluster.

The general idea of an MLM is that there is a regression line for the entire sample comprised by estimates of the fixed effects (γ_{00} and γ_{10} in Equation [1]) but that each cluster has a unique regression line. The random effects (the u terms) capture the difference between aspects of the cluster-specific intercept and slopes and the overall regression line formed by the fixed effects. For this reason, the random effect terms u are sometimes called "Level-2 residuals." Though not shown explicitly in Equation (1), a main interest of MLMs is in assessing how variable the cluster-specific regression lines are (i.e., what is the variance of the u parameters across clusters around the fixed effects). These are referred to as variance components; for the model in Equation (1), the variance components would be written as $\mathbf{u} \sim MVN(\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}\tau_{00} & \tau_{01}\\ \tau_{10} & \tau_{11}\end{smallmatrix}\right])$. That is, the vector of random effects $\mathbf{u}$ has a multivariate normal distribution with a mean vector of $\mathbf{0}$ (because, across clusters, the best regression line is captured by the fixed effects) and a covariance matrix comprising tau estimates. The diagonals of the tau-matrix are the variance components; the off-diagonals are the covariance components that capture the relation between the random intercept and the random slope (e.g., a positive covariance indicates that clusters with higher cluster-specific intercepts also tend to have higher cluster-specific slopes). Level-2 predictor variables can be included in the model to explain why cluster-specific regression lines are different from the fixed effect regression line.

Maximum likelihood and associated problems

When estimating an MLM with ML, variance components and fixed effects are estimated simultaneously. Estimating the parameters in this simultaneous fashion raises an important issue—namely that the variance components and the fixed effects, to a degree, are dependent upon one another. The variance components assess the amount of variation in the cluster-specific regression lines. To assess variation, it is necessary to have a location around which observations vary. Consider the basic formula for calculating the variance of a single variable X, $Var(X) = [\sigma(X_i - \bar{X})^2]/N$. In this formula, the reference location is the mean such that the variance is the average squared deviation of each observed value of X (noted by the i subscript) from the mean. In an MLM, the reference location for a variance component is the respective fixed effect. However, the fixed effects are unknown *a priori* and must be estimated so that the variance components can be calculated. Otherwise, it would not be known around what the cluster-specific lines are varying.

When estimating an MLM with ML, this dependency of the parameters means that a closed-form solution is not typically possible and iterative approaches such as the Expectation-Maximization (EM) algorithm or iterative generalized least squares are necessary (e.g., Raudenbush & Bryk, 2002, p. 52). The process is iterative such that the fixed effects are estimated initially, with the random effect being considered missing for all observations in the EM algorithm, for example. The variance components are then estimated based on the fixed effect values in the first iteration and these variance components are then used to update the fixed effect estimates in the next iteration. The process continues until there is essentially no change between the iterations. Thus, even though the process is iterative in nature, the fixed effects and variance components are estimated within the same algorithm in sequential steps.

By first estimating the fixed effects, these estimates end up being treated as known when calculating the variance components because the reference point in variance computations is a fixed point. This is problematic for two reasons. First, all the variability in the fixed effect estimates is ignored. Second, the degrees of freedom consumed to estimate the fixed effects is not accounted for. In larger samples, these issues do not have much of an effect: sampling variability of fixed effects decreases with sample size and changes in degrees of freedom above about 50 have only trivial effects. However, in smaller samples, sampling variability of fixed effects tends to be larger and small changes in degrees of freedom have a noticeable impact.

As a simple analogy, consider again the simple variance formula for a single variable. Readers may have noted that we presented the population version of this formula with N in the denominator rather than the sample version with $N-1$. The sample version of the variance formula reduces the denominator by 1 as a penalty for needing to estimate the sample mean in the numerator. With asymptotically large samples, the two formulas will yield essentially equal results. However, with smaller samples, the population formula will yield estimates that are too small because the population formula overestimates the precision in the data and does not account for the fact that the sample mean is estimated and not known.

In the variance formula analogy, ML is related to the population variance in that it does not account for the fact that fixed effects need to be estimated when estimating the variance components. Just like the population variance formula, the ML variance components are underestimated with smaller samples (Browne & Draper, 2006; McNeish, 2016). In MLMs, this is an issue because the variance components are prominently featured in the formula for calculating the standard errors for the fixed effects (e.g., Raudenbush & Bryk, 2002). Therefore, if the variance components are too small, the standard errors for the fixed effects will also be too small. If the standard error estimate is too small, the inferential *t* or *Z* test statistic will be too large meaning that *p*-values will be too small. This ends up inflating the Type-I error rates for the fixed effects. Thus, estimating the fixed effects and variance components simultaneously in ML leads to issues when the Level-2 sample size is small with the result being underestimated variance components and inflated Type-I error rates.

Restricted maximum likelihood

REML is the "$N-1$" version of ML and has been known to perform better than ML when the Level-2 sample size is small (Browne & Draper, 2006; Maas & Hox, 2005; McNeish & Stapleton, 2016b). As mentioned in the previous section, many of the issues that ML possesses with small Level-2 sample sizes stem from the fact that fixed effects and variance components are estimated simultaneously (albeit in iterative steps). REML addresses this issue by separating the estimation of the fixed effects from the variance components. This leads to improved estimates of variance components with smaller samples which, in turn, can improve the fixed effects standard error estimates. REML has the added benefit of being asymptotically equivalent to ML and it is not computationally intensive above and beyond ML.

To demonstrate how REML separates the fixed effect from the variance components in the estimation process, imagine a research question revolving around modeling *Math Scores* based on *Hours Studied* when students are nested within classrooms. A model with no Level-2 predictors and random effects for the intercept and slope would be written out as

$$\begin{aligned} \mathit{Math\ Score}_{ij} &= \beta_{0j} + \beta_{1j}\ \mathit{Hours\ Studied}_{ij} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{2}$$

The first step of REML is to obtain the ordinary least square (OLS) residuals via a single-level model, ignoring clustering.[2] Using the *Math Score* example,

$$\mathit{Math\ Score}_i = \beta_0 + \beta_1\ \mathit{Hours\ Studied}_i + e_i \tag{3}$$

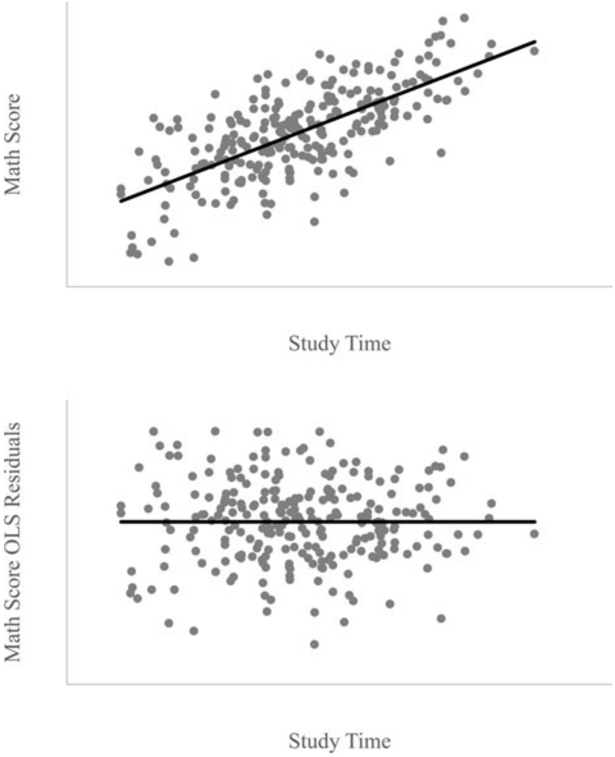

Figure 1. Comparison of the relation between the outcome and the predictor (top panel) and between the OLS residuals and the predictor (bottom panel).

The residuals from the OLS fit (e_i) are then saved. By definition, the OLS residuals (e_i) are independent of the predictor variables. This means that the correlation between *e* and *Hours Studied* is necessarily 0. The MLM is then fit using ML with the *OLS residuals* as the outcome rather than the original outcome variable. This may seem odd, but there is a rationale. The OLS residuals serve as a linear transformation of the original data that now are independent of the predictor variables (*Hours Studied* in this example) but have the same amount of variance (conditional on the predictors).

This is demonstrated in Figure 1 using a hypothetical simulated data set for this example. The top panel shows the relation for the original *Math Score* outcome on *Hours Studied* and the bottom panel shows the relation for the OLS residuals and *Hours Studied.* When using the OLS residuals, the fixed effects no longer need to be estimated because they are known to be zero by definition (via the properties of OLS) as illustrated by the horizontal line in the bottom panel of Figure 1 (i.e., the mean of the OLS residuals is 0 by definition and the correlation between *e* and *Hours Studied* is also 0 by definition). This essentially has the effect of rotating the relation of *Math* Score and *Hours Studied* so the fixed effects for the intercept and slope are necessarily equal to 0. That is, if Figure 1 is rotated about 60° clockwise, the points in the top panel essentially map onto the points in the bottom

[2] Technically, when implemented in software, this step is conducted with an error contrast of the outcome variable and the projection matrix rather than actually fitting a single-level model with OLS and saving the residuals. The result of the process has the same effect, so we describe the process using OLS to increase the intuition of the process and minimize the reliance on mathematical terminology.

panel—the data points are essentially in the same location and have the same amount of variability, conditional on *Hours Studied*. In fact, the conditional variance remains identical whether the outcome is the OLS residuals or the original outcome variable. So, REML estimates the variance components of a model with no fixed effects (i.e., the fixed effects are constrained to 0), which serves simply to partition the variance into Level-1 and Level-2 components—there is no simultaneous estimation because the focus of the first stage of REML lies solely on the variance components. Because REML does not have to deal with issues related to simultaneous estimation, the variance component estimates are (appropriately) higher than those estimated by ML with small samples. When samples sizes are large, the REML and ML variance components will essentially be equal.

The first stage of REML leaves the fixed effects out, so the next question naturally is: how does one get fixed effects estimates with REML? Once the variance components are estimated, REML then estimates the fixed effects in the second stage through *generalized least squares* (GLS). GLS is similar to OLS in that the fixed effect estimates are calculated with matrix multiplication without the need for iterative methods. Unlike OLS, GLS is capable of accounting for clustered data. The GLS fixed effect estimates are identical, on average, to ML estimates provided that the covariance structure of observations within clusters is known (e.g., Goldstein, 1986). That is, GLS does not estimate variance components but rather requires that researchers know these values ahead of time. In REML, the variance components are estimated first, so the Level-1 and Level-2 variance component estimates from the first stage are used to compose the covariance structure for GLS.[3] That is, rather than knowing the variance components ahead of time, the REML variance components are substituted into the GLS estimating equation as if they were known ahead of time and the GLS estimates are used for the fixed effects. Thus, REML achieves improved estimates of the variance components by separating the estimation and also yields nearly equivalent fixed effect estimates as ML.[4] Figure 2 conceptually shows the difference between how ML and REML iterate to a solution.

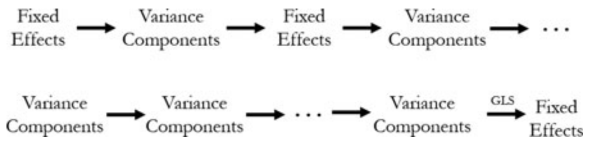

Figure 2. Comparison of iterative process between ML (top) and REML (bottom). ML uses simultaneous estimation and bounces between fixed effects and variance components until convergence is met. REML iterates to estimate the variance components until convergence, then estimates the fixed effects as the final step with GLS.

Fixed effect standard errors

Though the use of REML improves fixed effect standard error estimates because the variance components are more accurately estimated, standard error estimates continue to be problematic with smaller Level-2 sample sizes, even with REML. The issues related to standard error estimation with small Level-2 sample sizes stem from the frequentist definition of the standard error itself rather than the estimation issues used.

Frequentist inferential p-values essentially ask, if the study were conducted infinitely many times by repeatedly drawing random samples of identical size from the target population, what is the probability that one would obtain the estimated value of the coefficient if the population value were 0? In order to be able to discern information about probabilities, it is important to know the *sampling distribution* of the estimates. However, because studies are only conducted once, the distribution is unknown as researchers end up with only a single estimate that they want to compare to 0 in the population.

So then, how are p-values calculated? Using mathematics and by relying on asymptotics, it can be shown via the *central limit theorem* that the sampling distribution approaches a normal distribution asymptotically. The null sampling distribution should be centered around 0, but, to calculate probabilities, the variance of the distribution is needed. Fortunately, again using asymptotics and mathematics, it can be shown that *Fisher information* can approximate the variance of the sampling distribution, provided that it is sufficiently normal. Fisher information is taken from the inverse of the Hessian matrix of the likelihood function where the Hessian matrix consists of the second partial derivatives of the likelihood function with respect to each parameter in the model. Essentially, the curvature of the likelihood function at the ML estimate is used to inform the variability of the sampling distribution. That is, if the curvature of the likelihood function is very sharp at the ML estimates, then changes to the parameter estimates in any direction greatly reduce the likelihood, indicating that there is more certainty that the ML estimates represent the best solution. If the curvature of the likelihood is gradual at the ML estimates,

[3] In MLMs, the total variance **V** is equal to $\mathbf{V} = \mathbf{Z}\tau\mathbf{Z}^{\mathrm{T}} + \mathbf{R}$ where τ is the covariance matrix of the Level-2 random effects, **Z** is the random effect design matrix, and **R** is the covariance matrix of the Level-1 residuals. Using the REML estimates of τ and **R** to compute **V**, the GLS fixed effects are computed by $(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$ where **X** is the fixed effect design matrix and **y** is a vector of the outcome variable values. The addition of the **V** term accounts for the dependence of observations within clusters which is ignored by OLS. The OLS fixed effect estimates are calculated by $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

[4] The GLS and ML estimates are equal on average but are not perfectly identical. That is, in the long run, GLS and ML are expected to yield the same values but there may be small differences when estimates are compared for a single data set.

then changes in the parameter estimates in any direction change the likelihood very little, indicating that there are alternative solutions that are essentially as plausible as the ML estimates. Put another way, if another sample were taken, it would not be surprising to obtain different ML estimates when the likelihood curvature is gradual. The square root of the diagonal elements of Fisher information provide standard error estimates with likelihood estimation. Using asymptotics and advanced mathematics, the sampling distribution for parameters in the model can be determined even from just a single sample of data and properties of the likelihood function. For a more expansive and highly readable introduction to ML estimation, readers are referred to Chapter 3 of Enders (2010).

However, note that the previous paragraph used the word "asymptotic" on a number of occasions. This is because these relations only hold when the sample size approaches infinity or is otherwise sufficiently large. In the context of clustered data with small Level-2 sample sizes, this condition is highly unlikely to be met. In fact, with smaller samples, the central limit theorem often fails to take effect, which makes the shape of the sampling distribution more ambiguous (van de Schoot et al., 2014). Furthermore, Fisher information is a good approximation of the sampling variability asymptotically, but the approximation is much poorer with smaller samples, often being underestimated (Efron & Hinkley, 1978).

Thus, with smaller samples, the statistical machinery upon which frequentist inference is based breaks down and p-values are no longer trustworthy as Type-I error rates become highly inflated. Though some recent studies use failings of the frequentist statistical machinery as motivation to explore Bayesian methods with smaller samples (e.g., Muthén & Asparouhov, 2012; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015), there are frequentist corrections that are also available that preclude researchers from necessarily having to resort to a Bayesian framework. McNeish (2016) argued that the use of these small-sample-corrected frequentist methods can sometimes be preferable to Bayesian methods with small samples, especially when researchers lack strongly informative priors.

Small sample standard error corrections

Kackar-Harville

In the 1980s, statistical research attempted to determine the effect that violating asymptotic assumptions had on standard error estimates. Kackar and Harville (1984) posed the general conceptualization of the problem as (taken from Equation [2.1] in the original paper),

$$Var(\gamma) = Var^{REML}(\hat{\gamma}) + \text{Small Sample Bias} \qquad (4)$$

or, in English, that the sampling variability of the fixed effects is equal to the REML-estimated sampling variability estimate plus some amount of small sample bias incurred by violating asymptotic assumptions. It is known that $Var^{REML}(\hat{\gamma}) < Var(\gamma)$ with smaller samples, so the "Small Sample Bias" term is a corrective inflation factor. Thus, the goal of this line of research was to discern a mathematical function for the "Small Sample Bias" portion of the function so that proper standard error estimates could be obtained with smaller samples.

The mathematical details are intense to say the least as this mechanism is quite complex; however, the basic finding is that the Small Sample Bias portion does have a mathematical form but it requires that one have the population values for the variance components (see, the left-hand column of p. 854 in Kackar & Harville, 1984 for full details). To work around this issue Kackar and Harville (1984) use a *Taylor series expansion* whose basic premise is to approximate a complex nonlinear function (with potentially unknown values) with a simpler polynomial function (consisting solely of known values). The function to represent the Small Sample Bias portion of Equation (4) is one such complex nonlinear function, and Kackar and Harville (1984) discuss that its computation is infeasible without population values (p. 854) but it can be approximated as a function of model estimates (Kackar & Harville, 1984, Equation [2.2] or Equation [1] in Kenward & Roger, 1997).

To elucidate the idea of Taylor series expansion in a simpler context, take the nonlinear function $y = e^x$ where e is base of the natural logarithm (≈2.718). Perhaps the value of e is unknown (like the population values in a MLM) but that x is known (like estimated values in a MLM). Though we will not show the mathematics, a (second-order) Taylor series expansion of $y = e^x$ is $y = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$: notice that the Taylor series expansion is polynomial function only of x and that e is not featured in the function. Figure 3 compares the original function $y = e^x$ in black with the Taylor series expansion $y = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$ in gray. Notice that the two functions are almost on top of one another, showing that $y = e^x$ can be approximated solely by a polynomial function of x.

The Kackar-Harville correction works much in the same way except that the original function (the equivalent of "e" in Figure 3) is more complex and in multivariate space, while the equivalent of the "x" input is a vector of variance components estimates and their

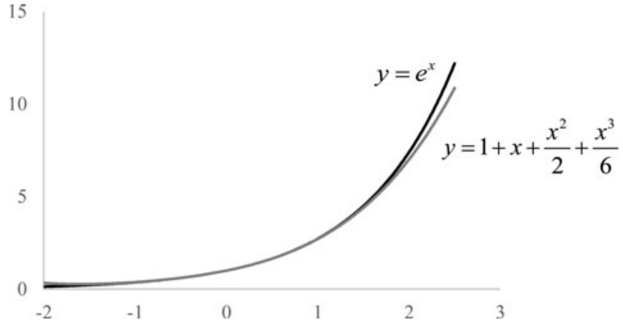

Figure 3. Comparison of $y = e^x$ (black) and the Taylor series expansion (gray).

REML-estimated covariance matrix.[5] The original idea of Kackar and Harville (1984) has been refined and extended by Prasad and Rao (1990) and Harville and Jeske (1992), so the correction is sometimes referred to as the Prasad-Rao-Jeske-Kackar-Harville correction. This is how the correction is listed in the SAS software, for instance, and this is how we refer to the correction in later sections in the interest of technical accuracy.

Kenward-Roger

Kenward and Roger (1997) uses Kackar and Harville (1984) as a starting point but note the $Var^{REML}(\hat{\gamma})$ term in Equation (4) similarly has issues. Namely, REML uses GLS to estimate the fixed effects after the variance components are estimated. However, the variance component estimates are not known and have some associated sampling variability, which is not accounted for because the *estimates* of the variance components are substituted into the GLS estimating equation where values are assumed to be known. This is the same issue that arises within the “Small Sample Bias” approximation of Kackar and Harville (1984) in Equation (4).

Therefore, in addition to the expansions in the Prasad-Rao-Jeske-Kackar-Harville family of corrections, Kenward and Roger (1997) perform a second Taylor series expansion to $Var^{REML}(\hat{\gamma})$ to account for the fact that variance components are estimated and not known when estimating fixed effects and their standard errors with REML (Equation [2] in Kenward & Roger, 1997). These two expansions together form the first step of the Kenward-Roger correction and provide more accurate fixed effect standard error estimates that are robust to violations of asymptotic assumptions (see Equation [3] in Kenward & Roger, 1997 for the full formulaic form of the correction).

With smaller samples, inference should be conducted with t-statistics rather than Z-statistics as with any other statistical model.[6] t-statistics take the ratio of the estimated value to its standard error ($t = \hat{\gamma}/SE_{\hat{\gamma}}$), so the first step of the Kenward-Roger correction ensures that the t-statistic is calculated accurately—Kackar and Harville (1981) show that the fixed effect point estimates, $\hat{\gamma}$, are estimated without bias in small samples and the Kenward-Roger correction ensures that $SE_{\hat{\gamma}}$ is estimated without small sample bias. Therefore, t-statistics should be accurate after employing appropriate corrections. To ensure proper p-values with smaller samples, accurate t-statistics are not sufficient, however. Additionally, one must have accurate degrees of freedom. With small samples, changes in degrees of freedom can lead to notable differences in p-values because the weight of the tails of the null t-distribution will change. In MLMs, degrees of freedom are notoriously difficult to calculate due to the sample sizes at multiple levels and no fixed formulas exist except for the most ideal situations (Schallje, McBride, & Fellingham, 2002). Software defaults for calculating degrees of freedom for fixed effects vary from program to program[7] and typically rely on approximations that tend to perform undesirably with smaller samples (Keselman, Algina, Kowalchuk, & Wolfinger, 1999). Therefore, to obtain more accurate p-values for inferential purposes, the second step of the Kenward-Roger correction is to approximate the degrees of freedom with a method of moments matching procedure, a procedure that is essentially a scaled version of the Satterthwaite (1946) procedure. This procedure often results in fractional degrees of freedom in order to give the most precise results.

In summary, the goal of the Kenward-Roger correction is to first correct the fixed effect standard error estimates in order to reduce the reliance on asymptotics. This results in standard error estimates that are larger than the asymptotic REML estimates. The first step leads to more accurate t-statistics, but degrees of freedom are still problematic. The second step of Kenward-Roger therefore provides an alternative procedure for degree of freedom calculation in order to refine p-values to improve inferential decisions. Note that Equation (4) relies on the REML estimate of the fixed effect sampling variability. Although traditional degree of freedom methods such as containment, between-within, or residual are available

[5] The full mathematical expression in Equation 4 is much more complex than the function $y = e^x$ used in Figure 3, so we do not wish to imply that the a Taylor series expansion of Equation 4 will necessarily result in the remarkably close approximation seen in Figure 3. The example in Figure 3 was chosen because it is a textbook example of how a Taylor series expansion operates. As a result, the Taylor series expansion happens to be quite good.

[6] We refer to univariate tests here, such as testing a single coefficient. If multiparameter tests are desired, the statement should be revised to say that researchers should use F tests rather than χ^2 tests because the F distribution is a finite sample version of the χ^2 distribution (i.e., an F distribution with infinite denominator degrees of freedom is equal to a χ^2 distribution if the numerator F degrees of freedom match the χ^2 degrees of freedom).

[7] Some software programs such as Stata and M*plus* report Z-tests for fixed effects which assume infinitely large samples and therefore do not have degrees of freedom. With smaller samples, Z-tests are not appropriate, generally speaking.

with ML or REML estimation, the derivations of the Kenward-Roger correction (and its antecedents) rely on REML estimation. That is, the Kenward-Roger is not applicable to models estimated with ML. If ML estimation is attempted along with a Kenward-Roger correction in software such as SAS PROC MIXED, an error message is returned.

Empirical example

To demonstrate the effect that different estimation and correction methods have on MLMs with small Level-2 sample sizes, consider data presented in Stapleton, Pituch, and Dion (2015). The data come from a cluster randomized trial (a common source of small Level-2 sample sizes) interested in whether a new socioemotional curriculum affects behavior for children at Head Start sites (higher behavior scores indicate better behaved students). The data feature 14 sites (7 in the treatment condition, 7 in the control condition) each with 6 students (84 students total; Level-2 sample size is 14). Each child's *Socioemotional Knowledge* is included as a Level-1 predictor and is group-mean centered. Level-2 predictors include treatment group status (variable label: *Treatment*) and the site average for *Socioemotional Knowledge* (i.e., the model uses a between-within specification; Bell & Jones, 2015). Random effects are included for the intercept and Level-1 slope. The model can be written as

$$\begin{aligned} Behavior_{ij} &= \beta_{0j} + \beta_{1j}(SEK_{ij} - \overline{SEK}_j) + r_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} Treat_j + \gamma_{02}\overline{SEK}_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{5a}$$

$$\mathbf{u} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{10} \\ \tau_{01} & \tau_{11} \end{bmatrix}\right) \tag{5b}$$

The model is fit five different ways with each relying on asymptotic assumptions to varying degrees. The five methods are (in order of decreasing reliance on asymptotic assumptions):

1. ML with inferential Z-tests, asymptotic standard errors;
2. ML with inferential t tests, containment degrees of freedom, and asymptotic standard errors;
3. REML with inferential t tests, containment degrees of freedom, and asymptotic standard errors;
4. REML with inferential t tests, containment degrees of freedom and Prasad-Rao-Jeske-Kackar-Harville standard errors[8];
5. REML with inferential t tests, Kenward-Roger degrees of freedom, and Kenward-Roger standard errors.

All models are estimated in SAS 9.3 using PROC MIXED. Containment degrees of freedom were selected for Model 1 through Model 4 because this is the SAS default for the model.[9]

Table 1 shows the estimates for the fixed effect coefficients, fixed effect p-values, and variance component estimates. Of particular note, focus on how the p-value for the treatment effect changes across the table from left to right (and, to a lesser extent, *Socioemotional Knowledge*). Using all asymptotic methods in Model 1, *Treatment* is significant where $p = .019$. Changing the inferential test from a Z-test to a t-test with containment degrees of freedom in Model 2, the p-value increases slightly to $p = .022$. If the estimation method is changed from ML to REML in Model 3, *Treatment* is now on the border of significance because REML variance components have (appropriately) inflated the variance component estimates (the intercept variance increases from 2.57 to 3.78, an increase of about 35%). When using Prasad-Rao-Jeske-Kackar-Harville standard errors (with containment degrees of freedom) rather than asymptotic standard errors in Model 4, *Treatment* is nonsignificant with $p = .068$. Finally, using the Kenward-Roger correction with associated degrees of freedom (relying on no asymptotic information) in Model 5, *Treatment* has a p-value of .106. As can be seen, even though the coefficient estimates are about equal across all five methods, the inferential interpretation of the results varies widely depending how strongly one (inappropriately) relies on asymptotic assumptions.

Discussion and concluding remarks

Though many resources exist to elucidate *when* REML and Kenward-Roger should be employed and when ML

[8] The Prasad-Rao-Jeske-Kackar-Harville is not available as a formal option in SAS though the relevant information can be pieced together. Using the DDFM=KR(FIRSTORDER) option in the MODEL statement uses the Prasad-Rao-Jeske-Kackar-Harville method for the standard error correction rather than the Kenward-Roger correction. The output still contains the Kenward-Roger degrees of freedom, so we manually calculated the p-values using the containment degrees of freedom.

[9] Containment degrees of freedom are calculated by $N - \text{rank}(\mathbf{X}, \mathbf{Z})$ where N is the overall sample size, $\mathbf{X}$ is a design matrix for the fixed effects and $\mathbf{Z}$ is a design matrix for the random effects. In linear algebra, the rank of a matrix is the dimension of the vector space spanned by the columns. The computation can be simplified with a random intercepts model for balanced data (as is the case in the Stapleton et al. data). For a fixed effect that does not have a random effect, the containment degrees of freedom are equal to (*total sample size*) minus (*number of predictors*) minus (*number of clusters minus Level-2 predictors*). This model has 84 total people, 3 predictors, 14 clusters, and 2 Level-2 predictors, so the containment degrees of freedom for a nonvarying fixed effect is $84 - 3 - (14 - 2) = 69$. For effects that do have random effects (the intercept in this model), the containment degrees of freedom is equal to (*number of clusters*) minus (*number of Level-2 random effects*) minus (*number of Level-2 predictors*). In this data, there are 14 clusters, 1 Level-2 random effect, and 2 Level-2 predictors, so the degrees of freedom for the intercept is $14 - 1 - 2 = 11$. These formulas are not as straightforward for unbalanced data or models with random slopes because the output from the rank operator will be dependent on aspects like the degree of unbalancedness and the random effect covariance structure.

Table 1. Comparison of estimates using methods that differ by their reliance on asymptotic assumptions.

	ML, Z		ML, t		REML		PRJKH		KR	
	Est	*p*	Est	*p*	Est	*p*	Est	*p*	Est	*p*
Treatment	2.83	.019	2.83	.022	2.77	.052	2.77	.068	2.77	.106
SEK	0.57	<.001	0.57	.012	0.57	.011	0.57	.013	0.57	.016
Level-2 SEK	– 0.17	.562	– 0.17	.565	– 0.14	.670	– 0.14	.700	– 0.14	.700
Var (Int)	2.57		2.57		3.78		3.78		3.78	
Var(SEK)	0.15		0.15		0.13		0.13		0.13	
Cov (I,S)	– 0.39		– 0.39		– 0.38		– 0.38		– 0.38	

Note: SEK, socioemotional knowledge; ML, maximum likelihood; Z, inference conducted with *Z*-test; *t*, inference conducted with *t*-test; REML, restricted maximum likelihood; PRJKH, REML with Prasad-Rao-Jeske-Kackar-Harville correction to standard error estimates; KR, REML with Kenward-Roger correction to standard error estimates and degrees of freedom based on a scaled method of moments.

should be avoided with multilevel models, there are far fewer resources to elucidate *what* these methods do and *how* they improve model estimates and inferences for multilevel models with smaller samples. We hope this paper helps to provide some insight as to what these methods are doing to circumvent small sample issues.

It is important to note one primary drawback of REML which lies in model comparisons. Because REML removes the fixed effects from the estimation, REML uses the restricted likelihood function that does not contain any information about the fixed effects (Raudenbush & Bryk, 2002). Therefore, if comparing the fit of different models using REML, the fixed effects must be identical for the comparisons of the restricted likelihood to be meaningful. Otherwise, ML must be used for model comparisons because the full likelihood includes information about both the fixed effects and the variance components. After competing models are compared using full ML, the final model can be estimated with REML (and associated corrections, if necessary) to protect against bias in variance component estimates and inflated Type-I error rates.

In terms of software implementation, there are some important differences that make some software programs more or less attractive for modeling multilevel data with small Level-2 sample sizes. SAS PROC MIXED and SAS PROC GLIMMIX have included the Kenward-Roger for some time and it is quite easy to implement the correction: one only needs to specify DDFM = KR as an option in the MODEL statement. SAS also offers the Kenward-Roger 2 correction via the DDFM = KR2 option. However, this version of the correction is only required when the covariance structure contains nonlinearities (Kenward & Roger, 2009) which is not commonplace in behavioral science research. Stata uses asymptotic *Z*-tests by default but beginning in version 14 (released April 2015) Stata now allows users to implement Kenward-Roger with the option dfmethod(kroger) in the mixed procedure. The pbkrtest R package does allow for Kenward-Roger to be implemented but does so for model comparisons rather than for the results of a single model.

SPSS does not allow users to implement Kenward-Roger, but it does allow for the Satterthwaite degrees of freedom to be used with *t*-statistics for testing fixed effects. The Satterthwaite procedure does not include the standard error correction from either Kackar and Harville (1984) or Kenward and Roger (1997), but it does provide a degree of freedom adjustment similar to the second step of the Kenward-Roger correction. The HLM software uses *t*-statistics to test fixed effects but does not currently offer Kenward-Roger as of this writing. Finally, frequentist estimation in M*plus* offers little assistance for handling small sample issues. REML is not available as an estimation method, there are no available small sample standard error corrections, and all effects are tested with *Z*-statistics with no option to elicit *t*-statistics instead. On the other hand, to their credit, M*plus* does have a user-friendly interface for using Bayesian methods which can be useful with smaller samples (e.g., van de Schoot et al., 2015). Researchers should note that the M*plus* default prior distributions in the Bayes module can produce some issues with small sample analyses and they should be changed prior to running a small sample analysis (McNeish, 2016).

Article information

Conflict of Interest Disclosures: The author signed a form for disclosure of potential conflicts of interest. The author did not report any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The author affirms having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by any external funding.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The author would like to thank Dan Bauer, Patrick Curran, and Dave Thissen for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the author's institution is not intended and should not be inferred.

References

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, *3*, 133–153. doi:10.1017/psrm.2014.7

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go?. *Methodology*, *10*, 1–11. doi:10.1027/1614-2241/a000062

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

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*, 457–482. doi:10.1093/biomet/65.3.457

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

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

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*, 43–56. doi:10.1093/biomet/73.1.43

Harville, D. A., & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, *87*, 724–731. doi:10.1080/01621459.1992.10475274

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

Huang, F. L. (2017). Using cluster bootstrapping to analyze nested data with a few clusters. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164416678980.

Kackar, R. N., & Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics-Theory and Methods*, *10*, 1249–1261. doi:10.1080/03610928108828108

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, *79*, 853–862. doi:10.2307/2288715

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. doi:10.1016/j.csda.2008.12.013

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

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite F tests and a nonpooled adjusted degrees of freedom multivariate test. *Communications in Statistics-Theory and Methods*, *28*, 2967–2999. doi:10.1080/03610929908832460

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

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

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

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

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

Prasad, N. G. N., & Rao, J. N. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*, 163–171. doi:10.1080/01621459.1990.10475320

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

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114. doi:10.2307/3002019

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 512–524. doi:10.1198/108571102726

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237–259. doi:10.2307/1165134

Stapleton, L. M., Pituch, K. A., & Dion, E. (2015). Standardized effect size measures for mediation analysis in cluster-randomized trials. *The Journal of Experimental Education*, *83*, 547–582. doi:10.1080/00220973.2014.919569

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

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