12-22-2018

Model misspecification and assumption violations with the linear mixed model: A meta-analysis

Brandon LeBeau
University of Iowa

Yoon Ah Song
University of Iowa

Wei Cheng Liu
University of Iowa

DOI: https://doi.org/10.1177/2158244018820380

LeBeau, B., Song, Y. A., & Liu, W. C. (2018). Model Misspecification and Assumption Violations With the Linear Mixed Model: A Meta-Analysis. *SAGE Open*. https://doi.org/10.1177/2158244018820380

Original Research

Model Misspecification and Assumption Violations With the Linear Mixed Model: A Meta-Analysis

SAGE Open
October-December 2018: 1–16

DOI: 10.1177/2158244018820380
journals.sagepub.com/home/sgo
SAGE

Brandon LeBeau[1], Yoon Ah Song[1], and Wei Cheng Liu[1]

Abstract

This meta-analysis attempts to synthesize the Monte Carlo (MC) literature for the linear mixed model under a longitudinal framework. The meta-analysis aims to inform researchers about conditions that are important to consider when evaluating model assumptions and adequacy. In addition, the meta-analysis may be helpful to those wishing to design future MC simulations in identifying simulation conditions. The current meta-analysis will use the empirical type I error rate as the effect size and MC simulation conditions will be coded to serve as moderator variables. The type I error rate for the fixed and random effects will be explored as the primary dependent variable. Effect sizes were coded from 13 studies, resulting in a total of 4,002 and 621 effect sizes for fixed and random effects respectively. Meta-regression and proportional odds models were used to explore variation in the empirical type I error rate effect sizes. Implications for applied researchers and researchers planning new MC studies will be explored.

Keywords

linear mixed model, longitudinal data, type I error rate, meta-analysis

Introduction

The linear mixed model (LMM), also commonly referred to as a multilevel model (Goldstein, 2010) or hierarchical linear model (Raudenbush & Bryk, 2002), is an extension of the multiple regression model to account for cluster dependency arising from nested designs. Included within nested designs are longitudinal designs where repeated measurements are nested within an individual. This data setup will serve as the primary focus of this article. These series of models were first introduced in the early 1980s (Laird & Ware, 1982), and the rapid improvement in computational power has helped these models become a popular data analysis method for researchers.

The LMM takes the following general matrix form:

$$Y_j = X_j\beta + Z_j b_j + e_j. \quad (1)$$

This is very similar to the multiple regression model, except now there are additional terms, $Z_j b_j$. b_j are random effects, which serve as additional residual terms and represent cluster-specific deviations from the average growth curve, and Z_j represents the design matrix for the random effects. The rest of the terms in the model are identical to a multiple regression, where Y_j represents the dependent variable for cluster j. X_j is the design matrix of covariates, β is a vector of fixed effects, and e_j is a vector of within-cluster residuals (i.e., residuals for every observation).

The random components of the LMM (i.e., b_j and e_j) are commonly assumed to be identically and independently normally distributed with means of zero and a specified variance matrix. These common assumptions can be summed up as follows: $b_j \sim iid \quad N(0,\mathbf{G})$ and $e_j \sim iid\ N(0,\sigma^2)$ (Raudenbush & Bryk, 2002). In addition, the independence assumption of the within-cluster residuals (i.e., e_j) is conditional on the random effects specified in the model (Browne & Goldstein, 2010). These random effects are what account for the dependency due to repeated measures, although there is the ability to allow the within-cluster residuals to be correlated due to the time lag in repeated measurements, a phenomenon called serial correlation (Diggle, Heagerty, Liang, & Zeger, 2002). This serial correlation may be especially important if the time lag between measurement occasions is short (Browne & Goldstein, 2010).

[1]The University of Iowa, Iowa City, USA

Corresponding Author:
Brandon LeBeau, The University of Iowa, 311 Lindquist Center, Iowa City, IA 52242, USA.
Email: brandon-lebeau@uiowa.edu

The statistical assumptions for the LMM are difficult to assess analytically due to the computationally intensive and iterative procedure of obtaining model estimates. In addition, when normality of the random components is not assumed, the mathematics becomes increasingly more difficult and intractable. As a result, Monte Carlo (MC) methods have been used to explore the relationship between assumption violations and model performance (Skrondal, 2000). MC studies have the advantage of strong internal validity due to the researcher directly manipulating the conditions of interest. The direct manipulation is not unlike true experiments, where the researcher can isolate the source of problems when estimating parameters (Skrondal, 2000).

The major drawback in MC studies is the potential lack of external validity (Skrondal, 2000). This weakness stems from the MC results being conditional on the conditions chosen to study. For example, if researchers conducting an MC study only simulate the random effects coming from a normal or chi-square(1) distributions, the question must be asked, can the study results be generalized beyond those two distributions. Hoaglin and Andrews (1975) started a discussion of best practices when reporting and conducting MC studies. More recently, Paxton, Curran, Bollen, Kirby, and Chen (2001) and Skrondal (2000) have offered design considerations to improve external validity. Much of the recommendations surround reducing the number of replications to increase the coverage of the simulation conditions (Skrondal, 2000).

Although the papers by Paxton et al. (2001) and Skrondal (2000) offer suggestions for improving new MC studies, the design considerations from past MC studies can not be altered to improve the external validity. As such, the current study aims to leverage prior MC studies to help improve the external validity of these studies, better understand gaps in simulation conditions, and succinctly inform applied researchers of assumption violations that can greatly affect the study results. This article aims to accomplish these three goals by quantitatively synthesizing the MC longitudinal LMM literature with a meta-analysis. The meta-analysis allows for the pooling of study conditions across numerous MC studies to increase sample size and depth of coverage of simulation conditions. In addition, many MC studies only report descriptive statistics for the study results and may miss complex interaction effects found through inferential modeling. A meta-regression was performed to overcome this limitation of some of the MC literature. A similar study for the one- and two-factor ANOVA models was done by Harwell, Rubinstein, Hayes, and Olds (1992).

Common Simulation Conditions With the LMM

MC studies exploring assumption violations with the LMM have focused primarily on the impact of nonnormal random effect distributions (LeBeau, 2012, 2013; Maas & Hox, 2004, 2005), the effect of serial correlation (Browne, Draper, Goldstein, & Rasbash, 2002; Ferron, Dailey, & Yi, 2002; Kwok, West, & Green, 2007; LeBeau, 2012; Murphy & Pituch, 2009), missing data (Black, Harel, & McCoach, 2011; Kwon, 2011; Mallinckrodt, Clark, & David, 2001), and estimation method (Delpish, 2006; Overall & Tonidandel, 2010). The random effect distributions simulated tend to be normally distributed, skewed (such as a chi-square distribution), or have heavy tails (such as a *t* or Laplace distribution). The serial correlation structures also tend to fall into three categories, independent structures, autoregressive (AR) type structures, or banded structures (such as the moving average models).

Most MC studies include sample size as a simulation condition, where, surprisingly, there is little variation in sample sizes used. The number of repeated measures is commonly less than 10 and the number of clusters rarely is larger than 100. The choice of sample size conditions is likely a function of using maximum likelihood for estimation. Maximum likelihood is an asymptotic estimation method, as such, understanding how the estimation method behaves for small samples is informative, and increasing sample size would likely improve estimation.

The number of fixed and random effects is another aspect of the simulation design that is chosen by the researcher. Unfortunately, these are two design choices that are commonly not manipulated directly. Instead, the number of fixed effects and random effects are held constant across studies. In addition, there is even less variation in the number of fixed and random effects chosen by researchers. It is uncommon for the number of fixed effects to be larger than six and the number of random effects to be larger than two.

The advantage of the current meta-analysis is to attempt to combine unique design choices made by independent researchers. This can offer some insight into data conditions that were not manipulated directly in a single MC study (such as the number of fixed effects), but do vary across studies. In addition, the slightly distinctive design choices made for a single manipulated condition (such as sample size or random effect distribution) can again be combined to explore whether one condition has a larger effect than others. This can aid applied researchers in their ability to understand implications when certain model assumptions are not met in practice. In addition, it can help inform researchers designing MC studies about gaps in the literature that would be worthy of further study.

Assumption Violations With the LMM

Results of the MC studies with assumption violations can be grouped into three categories. Estimation of the fixed effects tends to be unbiased regardless of the assumption violations (e.g., Kwok et al., 2007; LeBeau, 2013; Maas & Hox, 2004;

Murphy & Pituch, 2009). This has been shown with nonnormal random effect distributions (LeBeau, 2013; Maas & Hox, 2004), different sample sizes (LeBeau, 2013; Maas & Hox, 2004, 2005), with the presence of serial correlation (Kwok et al., 2007; LeBeau, 2012; Murphy & Pituch, 2009), and with different estimation algorithms (Delpish, 2006). Therefore, if the researcher is solely interested in estimates of the fixed effects, then little care to assumption violations is needed.

However, if the researcher is interested in estimates of the random effects or inference with the LMM, then researchers need to pay specific attention to assumption violations. Nonnormal random effects can inflate estimates of the random effects, especially in small sample size conditions (LeBeau, 2013; Maas & Hox, 2004). In addition, not modeling serial correlation when present can also cause an inflation in the random effects and underestimate the standard errors for the fixed effects (Kwok et al., 2007; LeBeau, 2012; Murphy & Pituch, 2009). This can lead to an inflation in the empirical type I error rate. Finally, misspecifying, specifically underspecifying, the random effect structure can also lead to severe inflation of the empirical type I error rates for fixed effects (LeBeau, 2012).

These results suggest that checking of model assumptions is important when researchers are interested in conducting inference, which likely encompasses most applied researchers. In addition, inflated estimates of the variance of the random components can lead researchers to include predictors to explain variation, when in reality, this variation is smaller than expected. A quantitative synthesis can be informative for applied researchers to show which assumption violations are crucial to achieving valid inferences. The MC literature can also be better informed through the ability of this study to include moderator variables that were not directly manipulated within a study, but were between studies, such as number of fixed effects.

Research Questions

Based on the prior MC literature, the following research questions were explored in the current meta-analysis:

Research Question 1: Is there evidence that the type I error rate is different from the nominal rate of 0.05 for fixed and random effects?
Research Question 2: To what extent does the independence assumption of the within-cluster residuals affect the empirical type I error rate?
Research Question 3: To what extent does the normality assumption of the random effects affect the empirical type I error rate?
Research Question 4: To what extent do the MC study characteristics moderate the relationships found in the above questions?

Method

Data Collection

Articles, dissertations, conference papers, or unpublished documents were gathered to attempt to answer the research questions above. Documents were selected if they are simulation studies that reported empirical type I error rates for the fixed effects. Only studies with continuous outcome variables were included to keep the comparison consistent. The simulation studies must include longitudinal data conditions, specifically multiple measurement occasions for individuals that are often smaller than cross-sectional models (Singer & Willett, 2003).

Based on the above criteria, the population of studies is defined as all possible MC LMM studies exploring data conditions similar to longitudinal studies using a continuous dependent variable. An initial search was performed in March 2012 and follow-up searches were performed in April 2013 and June 2014. Articles were selected for relevance based on their title. Abstracts from the articles selected by their titles were read to determine their relevance. If the study met the criteria established above, it was set aside to be read and code information from the study.

A Boolean search was used with the Eric, PsycInfo, and Dissertation Abstract databases to search for documents to be coded. The Boolean search string took the following form: (“Monte Carlo” or simulation) and (“linear mixed model” or “hierarchical linear model” or “mixed effects” or “mixed-effects” or longitudinal or LMM or HLM or LMEM) and generalized and (nonlinear or Bayesian or SEM or “structural equation model”).

The search identified 223 articles for review. Of those 223 articles, a total of 25 were selected for inclusion and were read further to include in the meta-analysis. There were three primary reasons why studies were excluded: (a) The article did not report empirical type I error rates (or it was not an outcome), (b) the study was cross sectional, or (c) the study was done in the structural equation modeling framework.

After studies were found using the above manner, Google Scholar was used to find articles that cited the articles found above. Footnote chasing was also used by exploring the titles in the reference list of the articles selected. The titles and abstracts of studies identified through Google Scholar were screened for inclusion.

From each of the sampled studies, effect sizes and a number of study characteristics were coded to help with the data analysis step of the meta-analysis. Three independent readers, who had completed PhD training or were in a PhD program related to quantitative methods, completed the coding of the studies included in this meta-analysis. One individual coded all the studies and the other two individuals independently coded approximately half of the studies to evaluate coding consistency. Internal consistency was high on all coded variables, including the primary dependent variable,

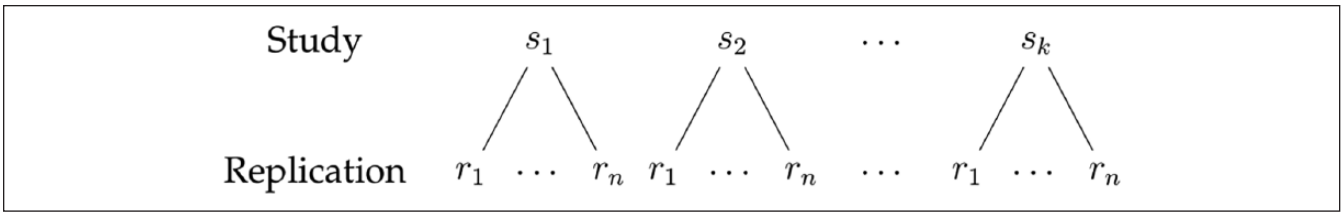

Figure 1. Data structure for model.

the empirical type I error rate. For the primary dependent variable, 98% of the values were coded the same across coders. Those that did not match were reviewed by looking at the original study for verification of the correct values.

Data Evaluation

Studies were first checked to ensure that they contained the dependent variable of interest, the empirical type I error rate. If the studies do not contain the empirical type I error rate or if the study used an empirical type I error rate other than 0.05, the study was not coded. The studies were then checked for methodological/coding flaws. Evidence that the data were generated accurately was explored first. If studies appear to show inaccuracies when the model assumptions have been met compared with the body of MC literature, they will be excluded from the sample of studies due to severe methodological or coding flaws. No studies were removed due to methodological or coding flaws.

All studies were read one at a time to code the variables of interest. Each MC study contributes many effect sizes to the meta-analysis; however, due to the quasi-random number generation within each study, the effect sizes are assumed to be independent within a study. However, there may be coder or author effects that need to be considered, and this potential dependency was adjusted by using an LMM and is discussed in more detail below.

Accuracy in coding was checked to ensure that all the effect sizes were coded properly. Summary statistics and plots were used to examine the distribution of effect sizes looking for possible extreme values because of erroneous coding. Very large or small empirical type I error rates were checked against the values published in the manuscripts. After errors were corrected, exploratory and inferential data analyses were used to attempt to explain variation in effect sizes. The coded variables and analyses are described in more detail below.

Dependent Variable

The primary dependent variable in the current meta-analysis was the empirical type I error rate for each condition in the MC studies. The type I error rate is commonly reported as the proportion of tests that reject a true null hypothesis. Statistical theory informs us that the proportion of rejected tests when the null hypothesis is true should be very close to the α value set by the researcher. Deviations in the proportion of tests that reject a true null hypothesis from the α value reflect problems in estimation; as a result, hypothesis tests are too conservative (proportion less than the α) or liberal (proportion greater than the α).

Independent Variables

The primary independent variables were the conditions that the MC studies directly manipulated. Variables commonly manipulated are the cluster sample size (i.e., how many individuals), presence of serial correlation, what kind of serial correlation structure is assumed, and the number of measurement occasions. Other conditions that are commonly not manipulated within an MC study that were coded included how many fixed effects are in the model, the number of random effects in the model, the number of replications within a cell of the study design, estimation method (e.g., full information maximum likelihood, restricted maximum likelihood [REML]), and whether the design was balanced or unbalanced.

The random effects are commonly assumed to follow a normal distribution, and violating this assumption has been studied thoroughly with numerous MC studies. The simulated random effect distribution was coded. In addition to the name of the coded distribution, the theoretical and empirical skewness and kurtosis values were coded for the random effects distribution. These independent variables were used to help determine whether the skewness or kurtosis of the distribution has a larger impact on the type I error rate.

Data Analysis

Exploratory data analyses were used to explore variation in the effect sizes. If significant variation in the empirical type I error rates was found, an LMM was used to see whether any moderator variables explain variation in the type I error rates. The LMM was chosen due to the hierarchical structure of the empirical type I error rates, which is illustrated in Figure 1.

The empirical type I error rates are proportions. As a result, the variance is a function of the specific value of the empirical type I error rates and the sampling distribution is unlikely to be normally distributed. Therefore, the empirical type I error rates were transformed using the Freeman–Tukey

transformation (Freeman & Tukey, 1950). This transformation takes the following form:

$$t_k = \frac{1}{2}\arcsin\left(\sqrt{\frac{x_k}{n_k+1}}\right) + \arcsin\left(\sqrt{\frac{x_k+1}{n_k+1}}\right), \quad (2)$$

where t_k is the transformed proportion, x_k is the number of type I errors, and n_k is the total number of replications. Many articles reported the empirical type I error rate as a proportion, not the number of type I errors for each cell of the MC design. To calculate the transformation, the number of empirical type I error rates made in each cell of the design was found by taking $x_k = \pi_k \times n_k$, where π_κ is the empirical type I error rate reported by the study.

The variance of the transformed proportions (t_k) is

$$v_k = \frac{1}{(4 \times n_k + 2)}, \quad (3)$$

where v_k is the variance and n_k is the number of replications. Miller (1978) defined a back-transformation to convert back into the raw proportion metric defined as follows:

$$\pi_k = \frac{1}{2}\left[1 - \sin\left(\cos t_k\right)\sqrt{\left(1 - \left(\sin t_k + \frac{\left(\sin t_k - \frac{1}{\sin t_k}\right)^2}{n_k}\right)\right)}\right], \quad (4)$$

where π_k is the back-transformed empirical type I error rate, t_k is the transformed value from Equation 2, and n_k is the number of replications. Results will be back-transformed to the empirical type I error rate metric for use in figures and tables.

Inferential model. The LMM was fitted with REML as this has been shown to produce less biased estimates of the random components (Fitzmaurice, Laird, & Ware, 2004; Raudenbush, 2009). The LMM took the following general form:

$$t_k = \beta_0 + \beta_1 X_{k1} + \ldots + \beta_t X_{kt} + b_k + e_k. \quad (5)$$

In Equation 5, t_k represents the transformed empirical type I error rate coded from the articles. β_0 is an intercept and $\beta_1, \ldots, \beta_t$ represent the relationship between the predictor variables, $X_{k1}, \ldots, X_{kt}$, and the dependent variable, t_k. Finally, this model contains b_k, which represent study-specific residual terms that are assumed to follow a normal distribution with mean zero and variance g^2. The e_k represent known sampling variances that are assumed to follow a normal distribution with mean zero and known v_k calculated from Equation 3 above. Within a study, the empirical type I error rates were treated as independent due to the quasi-random number generation used by MC studies (Rubinstein & Kroese, 2016). The moderators chosen to be included in the LMM were informed by the exploratory data analysis.

First, an omnibus model with no covariates was used to explore the heterogeneity in the effect sizes. The *Q* test was used to assess the amount of heterogeneity (Cooper, Hedges, & Valentine, 2009). If this test was significant, covariates were added to attempt to explain variation in the effect sizes with a meta-regression (Cooper et al., 2009). The covariates will take the form of simulation conditions that were coded and discussed above. Descriptive analyses will help to inform which covariates are included in the model. Significant predictors will be identified when the *z* value is greater than 2.33 in absolute value, representing a *p* value less than .01. The level of significance was selected to help control for compounding type I error rates from many tests of predictors in the analysis and better reflect covariates of practical significance.

To assess the amount of explanatory power of the predictors, an R^2_{Meta} statistic defined by Aloe, Becker, and Pigott (2010) will be used. The statistic takes the following form:

$$R^2_{\text{Meta}} = 1 - \frac{\hat{g}^2_{\text{cond}}}{\hat{g}^2_{\text{uncond}}}, \quad (6)$$

where $\hat{g}^2_{\text{cond}}$ and $\hat{g}^2_{\text{uncond}}$ represent the conditional and unconditional estimates of the between study variation.

A similar analysis to that described above was also done using the empirical type I error rate for the random effects. In this analysis, the dependent variable was t_k for the empirical type I error of the random effects and independent variables were the simulation conditions described in more detail above. There were fewer studies that studied the empirical type I error rate for the random effects; therefore, many study conditions coded did not have variation and were omitted from the model.

Proportional Odds Models (POMs). POMs (Yee, 2015) were also explored to further attempt to understand variation in the empirical type I error rates for the fixed effects. A new dependent variable was defined that represented three ordinal categories. These ordinal categories represented conservative tests (with a level of significance less than .05), accurate tests, and liberal tests (with a level of significance greater than .05).

To determine which group each observation belonged in, confidence intervals (CIs) were created for the empirical

type I error rate reported for each condition. These took the following form:

$$\mathrm{CI} = \pi_k \pm 2 \times \sqrt{\frac{\pi_k\left(1-\pi_k\right)}{n_k}}, \tag{7}$$

where π_k represents the empirical type I error rate for the fixed effects and n_k represents the number of replications for each simulation condition coded. If 0.05 is contained by the CI, then the dependent variable was coded as 1. If 0.05 was greater than the upper bound of the CI, the dependent variable took a value of 0 (an example of a conservative test); if it was less than the lower bound of the CI, the dependent variable took a value of 2 (an example of a liberal test). There was a total of 162 π_k (4%) that were below the lower bound of the CI, 2,897 π_k (72%) that were within the confidence band, and 943 π_k (24%) that were above the confidence band.

This variable was then modeled using a POM that had three categories. This model took the following general form:

$$\log\left(\frac{P\left(Y \le c\right)}{1-P\left(Y \le c\right)}\right) = \alpha_c + \beta X_k, \tag{8}$$

where the log odds of being less than or equal to a given category c is modeled as a function of the covariates (X_k). In the current example, where $c = 3$, there were two cumulative logits that were modeled simultaneously, $L_1 = \log(\pi_1/\pi_2 + \pi_3)$ and $L_2 = \log(\pi_1 + \pi_2/\pi_3)$.

The POM assumes that the regression weights are consistent between the two cumulative logits defined above, referred to as the parallel regression assumption (Yee, 2015). This assumption will be explored empirically using nested model chi-square comparisons (Yee, 2015). If there is evidence that the slopes vary, these will be allowed to vary between the two cumulative logits modeled. This strategy revealed that the covariate, model term (i.e., intercept, linear slope for time, other within slopes), did not satisfy the parallel regression assumption; as a result, these were allowed to vary between the cumulative logits described above.

A POM was also fitted for the empirical type I error rate of the random effects. There was a total of 160 π_k (26%) that were below the lower bound of the CI, 320 π_k (52%) that were within the confidence band, and 132 π_κ (22%) that were above the confidence band. As in fixed-effects analysis, the parallel regression assumption was not tenable for the covariate reflecting the variance component (i.e., within-cluster residual, variance of intercept, etc.), and these were allowed to vary between the cumulative logits.

Software. Data analysis was performed with R (R Core Team, 2015) using the metafor package (Viechtbauer, 2010). This included the fitting of the model shown in Equation 5 and the calculation of the Freeman–Tukey transformation, back-transformation, and the variance shown in Equations 2 through 4. POMs were fitted with the VGAM package (Yee, 2010, 2015). Figures were generated with the ggplot2 package within R (Wickham, 2009).

Limitations. There are at least three limitations of this study. The first stems from the nature of the meta-analysis, in that, studies that did not report the empirical type I error rate cannot be included in this meta-analysis. Second, the studies included in this meta-analysis are not a random sample of the population of MC studies; therefore, the results are fixed to the MC conditions coded from the included studies. The extent to which the studies not included (due to omission or did not report the empirical type I error rate) are significantly different than the included studies could bias the results. For this reason, some care needs to be taken when interpreting the results and the external validity may be affected. Both limitations overlap as the MC studies that did not report the type I error rate were also more likely to be journal articles. The degree to which these studies are different than the ones included in this meta-analysis may bias the results. Finally, many of the coded studies included in the meta-analysis were published in social science journals or were other document types completed within a social science context. The degree to which the data conditions included in the primary studies is different in disciplines outside of social science domains may affect the external validity.

Results

Fixed Effects

Summary information about the 13 articles coded in the meta-analysis is shown in Table 1 (two articles, Kwon [2011] and LeBeau [2012], are represented in four rows of Table 1 as they had two studies within a single article). As can be seen from Table 1, there are a total of 4,002 empirical type I error rates for the fixed effects with an average unweighted type I error rate of 0.063% and 95% CI = [0.055, 0.070]. The distribution of weighted effect sizes was highly concentrated between 0 and 0.1, but there were effect sizes greater than 0.15 and even one effect size larger than 0.4.

Additional summary statistic information for weighted back-transformed empirical type I error rates separated by various potential moderators can be seen in Table 2. From the table, there appears to be differences based on many of these moderators. For example, missing a random effect appears to have a strong impact on type I error rate with a mean of 0.078 compared with 0.055 when all random effects are modeled. Figure 2 shows the interaction between missing a random effect and the fixed effect term (e.g., intercept, time, or other within). As can be seen from Figure 2, the

Table 1. Article Summary Information of Unweighted Empirical Type I Error Rates for the Fixed-Effects and Other Monte Carlo Conditions.

Author	Source	*K*	Repl	Avg T1	Med T1	Min T1	Max T1	FE	Range CS	Range wCS
Black (2011)	Jour	40	1,000	0.088	0.075	0.000	0.310	4	(50, 50)	(20, 20)
Browne (2000)	Jour	20	930	0.058	0.053	0.003	0.099	2	(12, 48)	(18, 18)
Browne (2002)	Jour	14	10,000	0.053	0.054	0.047	0.057	2	(65, 65)	(62, 62)
Delpish (2006)	Diss	64	500	0.054	0.052	0.044	0.091	4	(30, 100)	(30, 30)
Ferron (2002)	Jour	192	10,000	0.054	0.052	0.045	0.079	4.75	(30, 500)	(3, 12)
Kwon (2011)	Diss	324	250	0.051	0.048	0.000	0.448	6	(40, 160)	(45, 45)
Kwon (2011)	Diss	540	250	0.049	0.044	0.008	0.192	10	(40, 160)	(45, 45)
LeBeau (2013)	Conf	244	300	0.059	0.057	0.033	0.100	4	(30, 50)	(6, 12)
LeBeau (2012)	Diss	1,500	500	0.063	0.061	0.015	0.119	5	(25, 50)	(6, 8)
LeBeau (2012)	Diss	750	500	0.082	0.070	0.024	0.274	5	(25, 25)	(6, 8)
Maas (2004)	Jour	144	1,000	0.059	0.058	0.038	0.088	4	(30, 100)	(5, 50)
Maas (2005)	Jour	108	1,000	0.054	0.054	0.037	0.075	4	(30, 100)	(5, 50)
Mallinckrokt (2001)	Jour	32	3,000	0.059	0.058	0.050	0.072	4	(100, 100)	(7, 7)
Murphy (2009)	Jour	64	10,000	0.059	0.052	0.047	0.125	4	(30, 200)	(5, 8)
Overall (2010)	Jour	66	1,500	0.062	0.063	0.039	0.087	4.33	(100, 100)	(9, 9)
Total	—	4,002	1,123	0.063	0.058	0.000	0.448	—	—	—

Note. *K* = number of effect sizes; Repl = replication; T1 = type I error rate; Med = median; Min = minimum; Max = maximum; FE = fixed effects; CS = cluster size; wCS = within-cluster size; Jour = journal article; Diss = dissertation.

empirical type I error is only inflated for the fixed effect terms associated with time, the others are similar to one another.

Table 2 also shows that there are some differences in the empirical type I error rate for differing simulated random effect distributions. The empirical type I error rate for the Laplace and chi-square(1) distributions were inflated at 0.067 and 0.066, respectively, compared with the other two distributions at 0.054 and 0.056 for normal and uniform, respectively. The distributions of the empirical type I error rate for the simulated random effect distributions also showed evidence of being positively skewed. This can be seen from the median being less than the mean and with large maximum values most notably for the normal, Laplace, and chi-square(1) distributions.

The effect of sample size is difficult to see from Table 2. For many sample sizes, the empirical type I error rate is close to the theoretical value of 0.05. Some deviations from this occurs when the cluster sample size is 25 and the within-cluster sample size is 6, 8, and 20. There may be more complicated effects that underlie the data here, and these differences will be explored in more detail with the inferential model.

Finally, considering the variance and the number of replications, the weighted average empirical type I error rate was 0.058 with a 95% CI = [0.054, 0.062]. In addition, the omnibus Q test was significant, $Q(4,001) = 22,798, p < .0001,$ suggesting that there is significant variation in the empirical type I error rates. The estimate for the between study variance (i.e., g^2) was 0.003 for the omnibus model using 3,842 empirical type I error rates. The sample size differs compared with Table 1 due to missing data on the cluster and within-cluster sample sizes arising from reporting practices. For example, some articles did not provide tables for the entire factorial research design. Instead, in the reported tables from the article, there was some aggregation over simulation conditions.

Inferential statistics. Expanding on the significant Q test, simulation conditions were added to the model to attempt to explain variation in the empirical type I error rate. The predictors explained significant variation as shown by the significant chi-square test for moderators, $Q_M(43) = 7,108,$ $p < .0001, R^2_{\text{Meta}} = .049$ Although the predictors are shown to be highly significant based on the moderator chi-square test, the explanatory power of the model is small. This could be attributable to the small amount of variation between studies. The significant predictors can be seen in Table 3. Note, a handful of predictors were significant, but had back-transformed estimates of zero to three decimal places (i.e., 0.000).

The average empirical type I error rate for the intercept term (i.e., initial status) is very close to 0.05 at 0.048. This suggests that, on average, the empirical type I error rate control is very good for the reference group. More specifically, this is for dissertations, a normal random effect distribution, and independent fitted and generated serial correlation structures. More simply, this would represent the situations where model assumptions have been adequately met. Assumption violations, such as nonnormal random effects, does not affect the empirical type I error rate for the intercept. This can be

Table 2. Summary Statistics of Weighted Back-Transformed Empirical Type I Error Rate for the Fixed Effects by Parameter, Level of Parameter, Article Source, Cluster Size, Within-Cluster Size, Serial Correlation, Random Effect Distribution, and Missing Random Effect.

Moderator	Avg T1	Med T1	Min T1	Max T1	*K*
Term					
Intercept	.056	.055	.003	.448	1,835
Time	.060	.057	.000	.274	1,693
Within	.065	.066	.023	.110	474
Level parameter					
Level 1	.061	.059	.003	.274	2,044
Level 2	.057	.055	.000	.448	1,958
Cluster sample size					
12	.059	.079	.003	.099	10
25	.072	.067	.021	.274	1,500
30	.061	.060	.036	.124	264
40	.045	.045	.009	.177	288
48	.048	.048	.036	.058	10
50	.058	.058	.000	.310	922
65	.053	.053	.046	.056	14
80	.048	.049	.003	.224	288
100	.054	.052	.036	.086	258
160	.052	.049	.013	.448	288
200	.050	.049	.046	.057	32
500	.050	.050	.046	.056	40
Random effect distribution					
Chi-square(1)	.066	.062	.021	.268	846
Laplace	.067	.064	.015	.272	846
Normal	.054	.053	.000	.448	2,262
Uniform	.056	.056	.037	.068	48
Article source					
Dissertations	.059	.057	.003	.448	3,178
Journal	.058	.056	.000	.310	680
Conference paper	.059	.057	.033	.100	144
Missing random effect					
No	.055	.054	.000	.448	3,252
Yes	.078	.070	.024	.274	750
Within-cluster sample size					
3	.051	.050	.047	.057	24
4	.055	.054	.044	.078	72
5	.057	.056	.040	.082	92
6	.066	.063	.023	.274	1,221
7	.058	.056	.049	.071	32
8	.067	.062	.015	.272	1,205
9	.060	.062	.038	.086	66
12	.060	.060	.034	.100	96
18	.054	.053	.003	.099	20
20	.078	.079	.000	.310	40
30	.054	.052	.037	.091	124
45	.049	.049	.003	.448	864
50	.054	.055	.036	.086	60
62	.053	.053	.046	.056	14

Note. T1 = type I error rate; Med = median; Min = minimum; Max = maximum; *K* = number of effect sizes; Intercept = regression terms modeling the intercept; Time = regression terms modeling time; Within = regression terms of another within-cluster variable.

seen from Table 3, where the back-transformed values associated with the intercept are zero to two or three decimal places.

On average, the time or linear trend fixed effect did not show evidence of being significantly different from the initial status. However, there was evidence of inflation, on

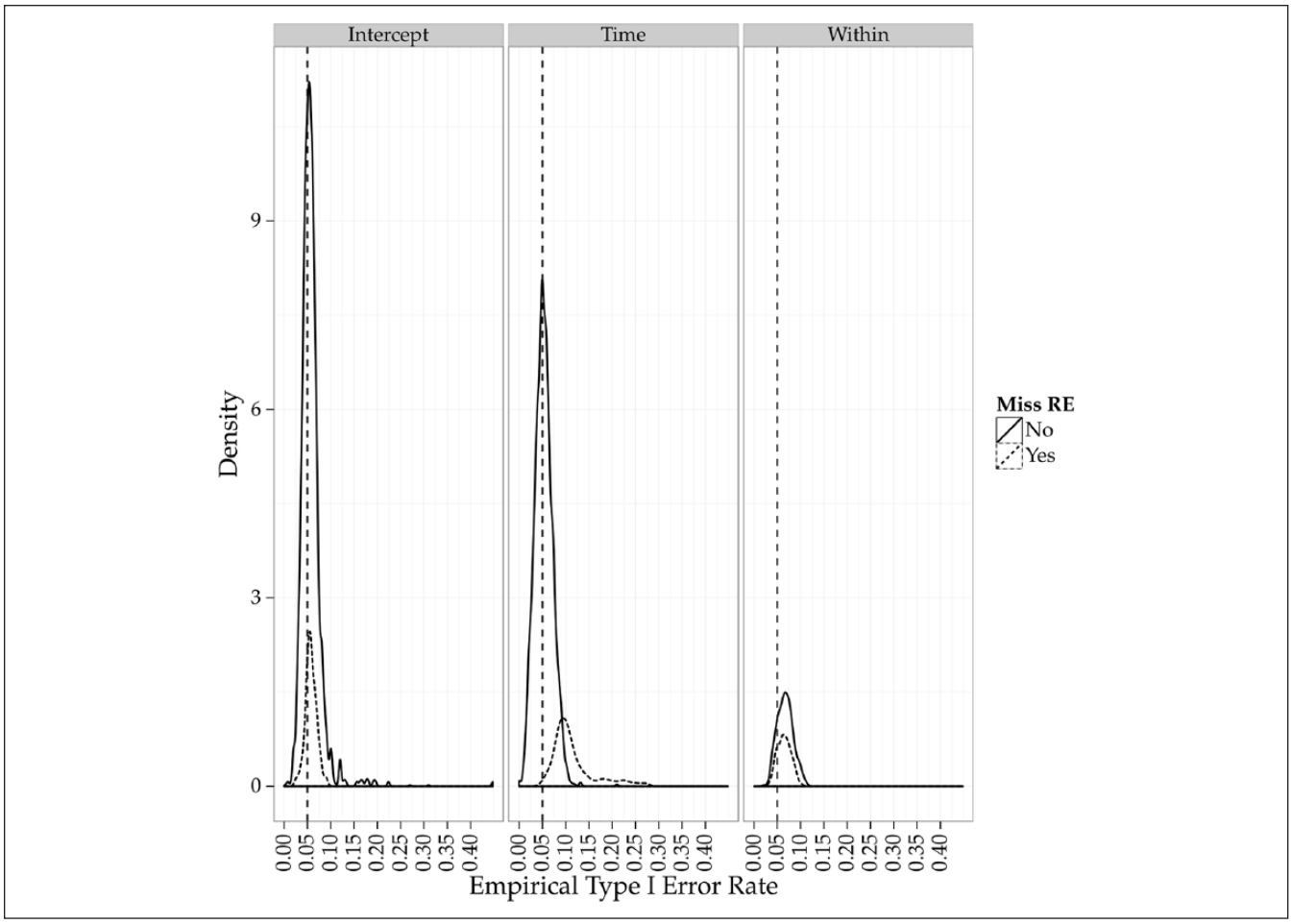

Figure 2. Density plot of weighted empirical type I error rates by missing a RE and fixed effect term.
Note. RE = random effect.

Table 3. Fixed Effect Meta-Regression Results for Significant Predictors in the Transformed $(\hat{\beta}_{t_k})$ and Type I Error Rate Metrics $\hat{\beta}_{\pi_k}$.

Variable	$(\hat{\beta}_{t_k})$	$\hat{\beta}_{\pi_k}$	99% CI $\hat{\beta}_{\pi_k}$	*Z*
Intercept	0.224	0.048	[0.019, 0.090]	6.296
w/ sample size	−0.025	−0.000	[−0.000, −0.000]	−3.629
Fit SC UN	0.051	0.002	[0.001, 0.002]	24.716
Time: Miss RE	0.171	0.028	[0.025, 0.031]	41.378
Fit SC ARMA: Miss RE	−0.024	−0.000	[−0.000, −0.000]	−5.439
Within: Fit SC ARMA	0.028	0.000	[0.000, 0.000]	8.097
Within: Fit SC MA2	0.035	0.000	[0.000, 0.001]	9.473
Time: Fit SC AR1:Miss RE	−0.119	−0.000	[−0.000, −0.000]	−19.527
Within: Fit SC AR1:Miss RE	0.027	0.001	[0.003, 0.001]	3.438
Time: Fit SC ARMA:Miss RE	−0.115	−0.000	[−0.000, −0.000]	−18.905
Within: Fit SC ARMA:Miss RE	0.028	0.001	[0.002, 0.001]	3.596
Time: Fit SC MA1:Miss RE	−0.084	−0.000	[−0.000, −0.000]	−12.953
Time: Fit SC MA2:Miss RE	−0.105	−0.000	[−0.000, −0.000]	−16.131
Within: Fit SC MA2:Miss RE	0.027	0.002	[0.004, 0.001]	3.352

Note. Reference groups were dissertations, normal random effect distributions, intercept fixed effect terms, and independent fitted and generated serial correlation structures. CI = confidence interval; RE = random effect; Fit = fitted; SC = serial correlation; ARMA = autoregressive moving average; MA = moving average; Miss = missing; w/ = within; colon (:) = an interaction.

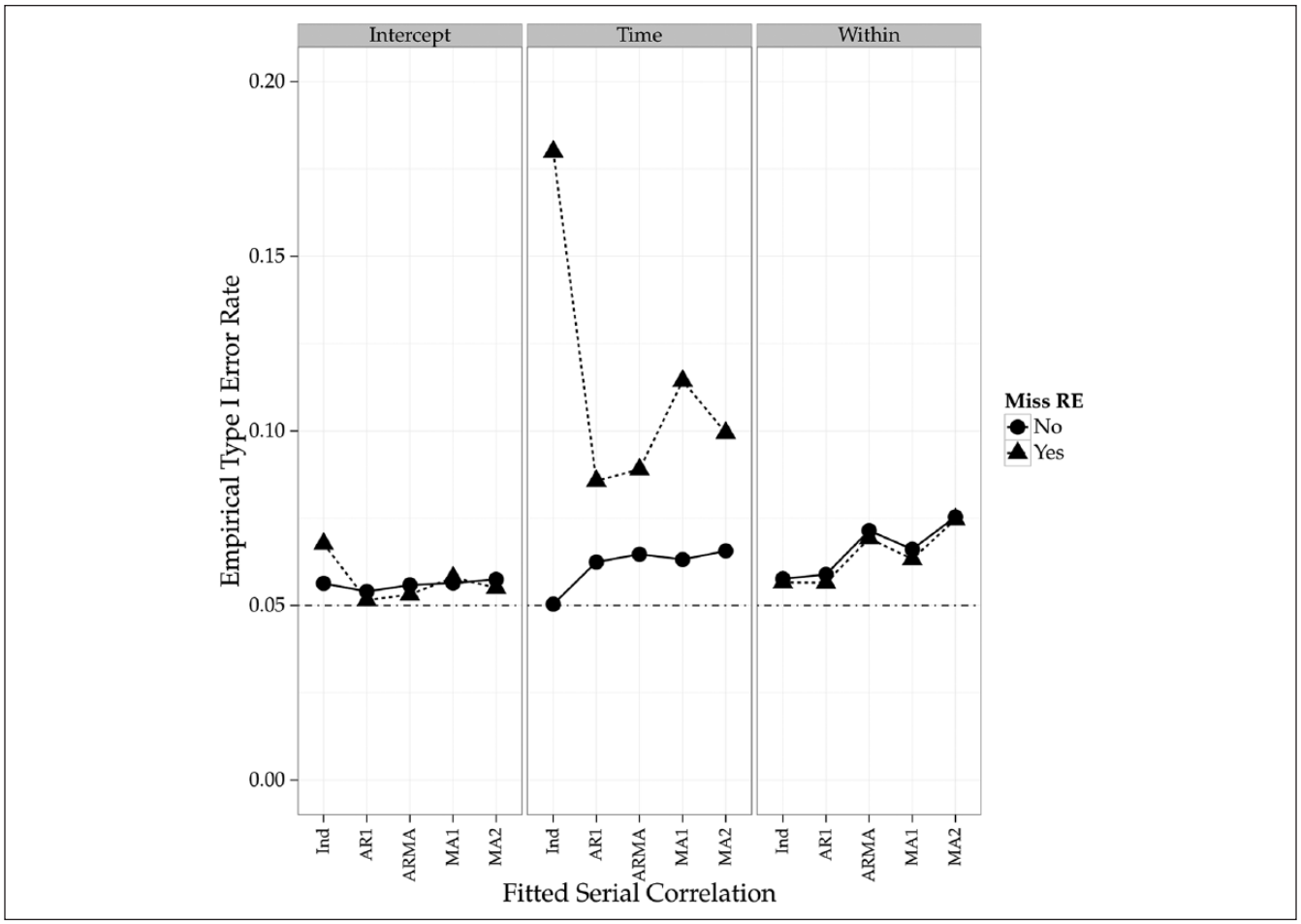

Figure 3. Interaction plot showing three variable interaction between fixed effect term, fitted serial correlation structure (Toeplitz and unstructured omitted from figure), and misspecification of the random effect structure.
Note. AR = autoregressive; ARMA = autoregressive moving average; MA = moving average; RE = random effect.

average, when the random effect distribution was misspecified, an increase of 0.028 in the empirical type I error rate metric. This results in an estimated empirical type I error rate of 0.076, an approximately 50% increase in the empirical type I error rate metric.

There is evidence that including some form of serial correlation does improve the empirical type I error rate, but does not fully overcome the inflation due to misspecification of the random effect structure. This can be seen by the negative coefficients for the time by fitted serial correlation by misspecification of the random effect three-way interaction. Estimates for these terms are not as large as the time by misspecification of the random effect two-way interaction, suggesting an adjustment but not full correction of the inflation. The first-order autoregressive (AR(1)) and first-order autoregressive moving average (ARMA(1, 1)) seem to do the best job of correcting the inflation, shown in Figure 3. The figure shows the fixed effects associated with time and a misspecification of the random effect structure, the empirical type I error rate is inflated. The effect is particularly large when the fitted serial correlation structure is independent (i.e., assuming no serial correlation underlies the data). Including some form of serial correlation does help to limit the strong inflation however.

Other within individual variables behave very similar to the intercept term. There are significant terms labeled as "Within" in Table 3, but many of these have very small estimates suggesting terms of little practical significance.

POM. Results for the final POM can be seen in Figure 4 for covariates of primary interest. This plot depicts an interaction between the generated and fitted serial correlation structure, the model term, whether the random structure was misspecified, and the two cumulative logits shown with solid lines (L_1) and dashed lines (L_2). As can be seen for the intercept term (left most panel), the second cumulative logit is very close to 1, indicating that very few of the observations were in the third group. Furthermore, the probability of being in Group 1 is also very small, indicating that fixed effect terms associated with the intercept tend to have the nominal

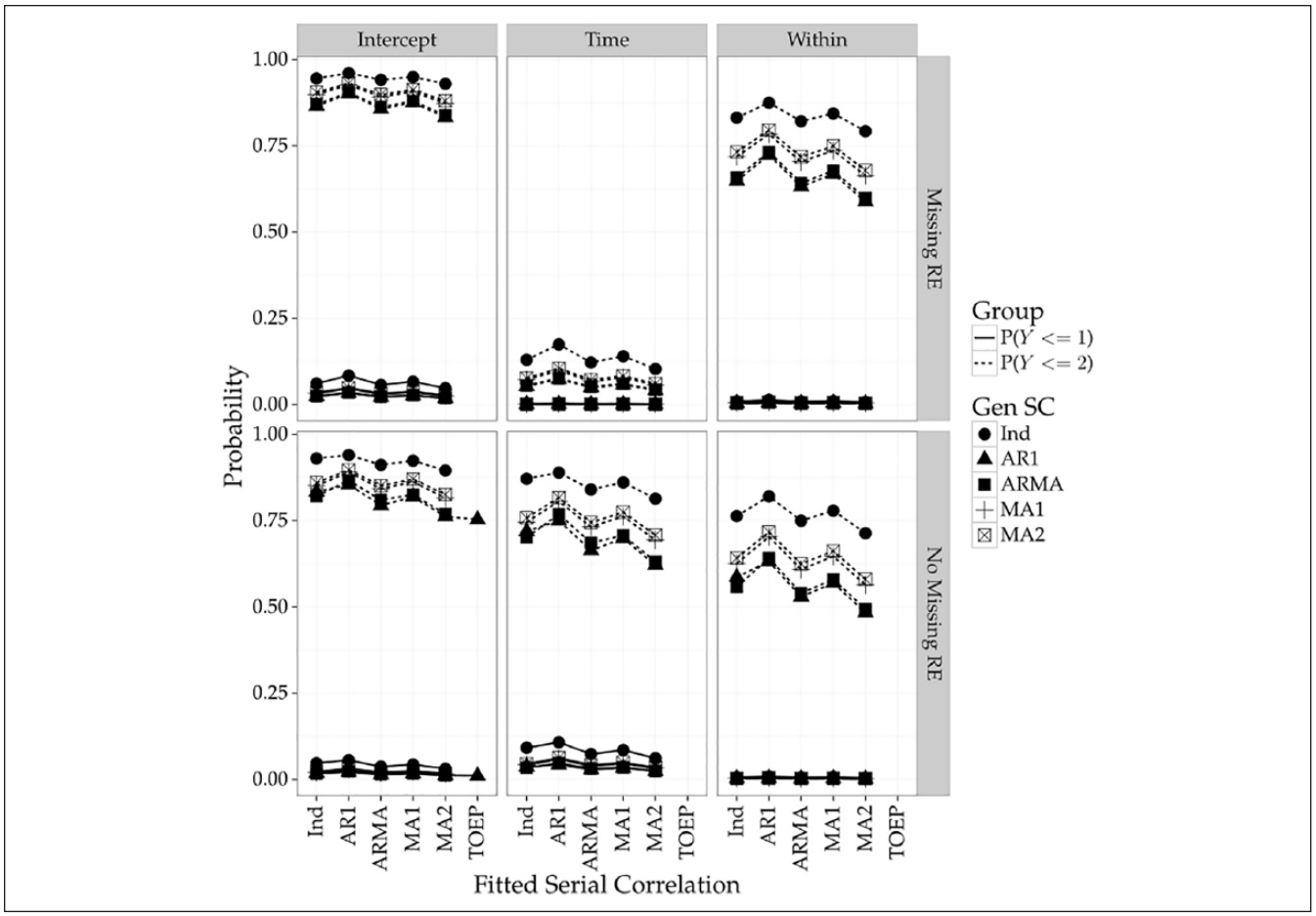

Figure 4. Interaction plot showing results of the partial proportional odds model by the fitted serial correlation structure, generated serial correlation structure, fixed effect term, and misspecification of the random effect structure.
Note. $P(Y \leq 1)$ represents the probability of being a conservative test (a level of significance less than .05) compared with accurate or liberal tests (a level of significance greater than .05). $P(Y \leq 2)$ represents the probability of being a conservative or accurate test compared with liberal tests. AR = autoregressive; ARMA = autoregressive moving average; MA = moving average; RE = random effect.

Table 4. Article Summary Information of Unweighted Empirical Type I Error Rates for the Random Effects and Other Monte Carlo Conditions.

Author	Source	*K*	Repl	Avg T1	Med T1	Min T1	Max T1	RE	Range CS	Range wCS
Black (2011)	Jour	20	1,000	0.153	0.080	0.000	0.860	2	(50, 50)	(20, 20)
Browne (2000)	Jour	20	930	0.103	0.091	0.032	0.186	2	(12, 48)	(18, 18)
Browne (2002)	Jour	11	1,000	0.057	0.059	0.040	0.075	3	(65, 65)	(62, 62)
Delpish (2006)	Diss	48	500	0.100	0.052	0.045	0.400	2	(30, 100)	(30, 30)
Kwon (2011)	Diss	162	250	0.043	0.048	0.000	0.132	2	(40, 160)	(45, 45)
Kwon (2011)	Diss	162	250	0.043	0.048	0.000	0.124	2	(40, 160)	(45, 45)
Maas (2004)	Jour	108	1,000	0.097	0.051	0.005	0.429	2	(30, 100)	(5, 50)
Maas (2006)	Jour	81	1,000	0.067	0.064	0.035	0.116	2	(30, 100)	(5, 50)
Total	—	621	561	0.066	0.052	0.000	0.860	—	—	—

Note. *K* = number of effect sizes; Repl = replication; T1 = type I error rate; Med = median; Min = minimum; Max = maximum; RE = random effects; CS = cluster size; wCS = within-cluster size; Jour = journal article; Diss = dissertation.

α rate enclosed in the 95% CI, indicating adequate type I error rate coverage.

The middle panel in Figure 4 shows the model results for fixed effects associated with the linear slope. As can be seen

Table 5. Random Effect Meta-Regression Results for Significant Predictors in the Transformed $(\hat{\beta}_{t_k})$ and Type I Error Rate Metrics $(\hat{\beta}_{\pi_k})$.

Variable	$\hat{\beta}_{t_k}$	$(\hat{\beta}_{\pi_k})$	99% CI $(\hat{\beta}_{\pi_k})$	Z
Intercept	0.334	0.106	[0.055, 0.172]	8.246
Var b_0	0.067	0.003	[0.002, 0.004]	24.168
Var b_1	−0.055	−0.000	[−0.000, −0.000]	−19.909
Number FE 6	−0.134	−0.000	[−0.000, −0.000]	−2.806
Number FE 10	−0.134	−0.000	[−0.000, −0.000]	−2.796

Note. Reference groups were dissertations, two fixed and random effects, within-cluster residuals. CI = confidence interval; FE = fixed effect.

from the figure, there is an interaction effect where misspecification of the random effect structure significantly decreases the probability of being in Group 1 or 2. This suggests that the probability of having an inflated type I error rate is larger than .75 in many cases. When all the random effect terms are correctly modeled, the probability of having an inflated type I error rate decreases significantly as shown in the bottom plot of the middle panel.

Finally, the rightmost panel in Figure 4 shows the effect for terms associated with other Level 1 slopes not including the intercept or linear slope for time. Compared with the intercept, the probability of being in both Groups 1 and 2 is slightly smaller, suggesting that the probability of being in Group 3 is higher than the intercept and the linear slope when the random effect structure is correctly modeled.

Throughout all the conditions, the AR(1) fitted serial correlation structure did increase the probability of being in Group 1 or 2, especially for the linear slope and other within slopes (middle and right panels of Figure 4). The independent generated serial correlation structures also provided the highest probability of being in Group 1 or 2, a situation that is not surprising, given that these are cases where serial correlation is not present in the data, and the random effects should adequately represent the dependency due to repeated measures. Finally, Figure 4 also depicts across all conditions that the probability of being in Group 1 (type I error rate smaller than the nominal rate) is not common. Liberal tests are a much larger concern than conservative significant tests, particularly when serial correlation is present in the underlying data.

Random Effects

The unweighted empirical type I error rate for the random effects can be seen in Table 4. Of the 13 articles coded, only seven reported empirical type I error rates for the random effects, resulting in a total of 621 effect sizes. On average, the unweighted empirical type I error rate was at 0.066 with evidence of positive skew and large variation as shown by the minimum and maximum values.

Considering the number of replications and variance, the weighted average empirical type I error rate was 0.066 with a 95% CI = [0.048, 0.087]. The omnibus Q test was also significant $(Q(491) = 14{,}891, p < 0.001)$, suggesting that there is significant variation that may be able to be explained by study conditions that were coded. The estimate of the between study variation was 0.003.

Inferential models. With the significant Q test, a meta-regression was performed to attempt to explain the significant variation in the empirical random effect type I error rates. Significant predictors at the .01 level are shown in Table 5. The conditional model explained significant variation as shown by the significant chi-square moderator test, $Q_M(9) = 2{,}633, p < .001; R^2_{\text{Meta}} = .759$. Table 5 shows the weighted empirical type I error rate was 0.106 for the reference condition being those with average sample size, dissertations, two random and fixed effects, and for the variance of the within-cluster residuals. The other two significant predictors shown in the table are adjustments for the variance of the random effect terms. Specifically, the empirical type I error rate of the variance of the random intercepts tends to be slightly inflated compared with the within-cluster residuals. In contrast, the empirical type I error rate of the variance of the random slopes tends to be smaller (i.e., less biased) compared with the within-cluster residuals. Of note, the cluster sample size was also significant, however, this term was not reported in the table as the parameter estimate was −0.000 (in the transformed scale) and not deemed significant in practice.

POM. Results for the POM can be seen in Figure 5 for significant covariates. This plot shows the effect of the number of fixed effects and the random effect term on the probability of the empirical type I error rate to be conservative, accurate, or liberal. In this figure, the solid lines represent the probability of a conservative test, the dashed lines represent the probability of an accurate or conservative test, and one minus the probability of the dashed lines represent liberal tests. Increasing the number of fixed effects significantly increases the probability of being in Group 1 or 2 compared with Group 3, suggesting that these tests are more likely to be accurate or conservative rather than liberal, with the probability being very close to 1. However, with only a couple

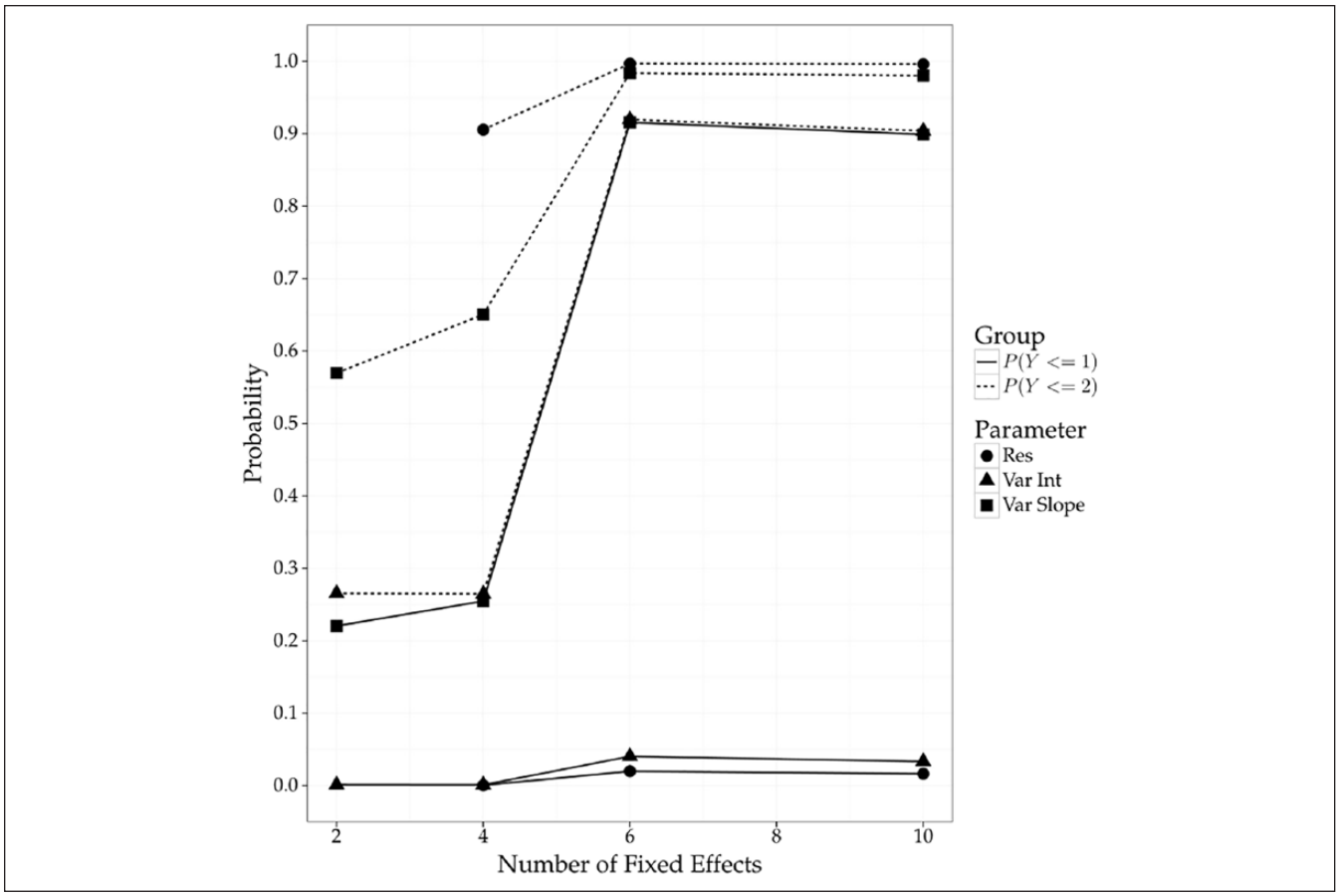

Figure 5. Interaction plot showing results of the partial proportional odds model by the random effect term and number of fixed effects.

Note. $P(Y \leq 1)$ represents the probability of being a conservative test (a level of significance less than .05) compared with accurate or liberal tests (a level of significance greater than .05). $P(Y \leq 2)$ represents the probability of being a conservative or accurate test compared with liberal tests.

fixed effect terms in the model, the probability of a liberal test is much larger, as likely as .7 for the significant test associated with the variance of the intercept.

The within-cluster residuals (labeled as Res) and the variance of the intercept tend to be accurate tests (most likely to be in Group 2), particularly when more fixed effects are included in the model. In contrast, there is a rather large probability (about .90) of the empirical type I error rate to be conservative for the variance of the slope when more than six fixed effects are included in the model as shown in Figure 5.

Discussion

The purpose of this meta-analysis was to help improve the external validity of MC LMM studies, better understand gaps in simulation conditions, and inform applied researchers of assumption violations that can affect the study results. A meta-analysis of the empirical type I error rates from MC studies was performed and the manipulated and nonmanipulated study conditions were combined to empirically understand which are most important.

There are a handful of takeaway messages from this meta-analysis. First, not modeling a random effect when that term is underlying the data leads to inflated type I error rates. For example, in the study by LeBeau (2012), the data were generated with a random effect for time, but this random effect was omitted during model fitting. This leads to inflated type I error rates for the fixed effect parameters associated with that random effect. This inflation of the type I error rate could lead researchers to reject true null hypotheses more often than one would expect given their specified α level. Given the level of significance is one aspect that researchers have direct control over, the disconnect between the empirical value and the one specified is problematic for applied researchers. The problem of misspecification of the random effect structure in this way may occur most often when models fail to converge. In this situation, fixing a random effect to zero, can drastically improve the convergence rate and may be a step taken by applied researchers. Understanding ways to overcome this inflation or better detecting serial correlation would be meaningful for the literature.

Fitting a serial correlation structure has the effect of helping to correct the inflation in type I error rates when a random effect is missing. However, fitting the serial correlation structure does not fully recover the type I error rate to nominal levels. In addition, by looking at the parameter estimates and Figure 3, the AR(1) and ARMA(1, 1) structures seem to do a better job of reducing the type I error rate than the MA(1) or MA(2) serial correlation structures (−.013 and −.012 vs. −.0006 and −.010, respectively). This may be due to AR(1) and ARMA(1, 1) structures better representing longitudinal data structures (i.e., correlation decreasing between measurement occasions as the time lag increases), whereas the MA structures may better align with other nested data structures.

No relationship was found between the random effect distribution and the type I error rate for the fixed effects, which has been found previously in the literature (LeBeau, 2013; Maas & Hox, 2004, 2005). Similarly, the generated serial correlation structure does not have an impact on the type I error rates of the fixed effects. This suggests that the random effects do an adequate job of accounting for the dependency due to repeated measurements, regardless of the type of serial correlation structure underlying the data.

Finally, the small effects related to the sample sizes is an interesting finding. With maximum likelihood being an asymptotic estimation method, the larger sample sizes may have provided better estimates. The small effect helps to inform researchers with relatively small sample sizes that the type I error rate can be held in check with few observations and relatively few clusters (12 clusters was the smallest condition in the current meta-analysis). However, sample size may play an important role in the estimates of the variances of the random components and would be worthy of further study.

Informing Applied Researchers

There are three recommendations for applied researchers with respect to empirical type I error rates for the LMM. First, results suggest that conservative tests are less common with the LMM for the fixed effect terms (see Figure 4). Instead, significant tests are much more likely to be liberal. One aspect to be particularly careful with is when the random effect for time is not included in the final model. This could be a case where the random effect structure is misspecified, and severe inflation of the empirical type I error rate may be present for fixed effects associated with the linear slope. In these situations, exploring variation in the trajectories of individuals would be helpful to see whether there is variation in the data not being captured by the LMM. If there is evidence of misspecification of the random effect structure and the inclusion of the random effect for the linear slope leads to nonconvergence, using robust standard errors may be a way to alleviate elevated empirical type I error rates.

A second recommendation is to check for the presence of serial correlation in the data, especially when the repeated measurements are measured close in time. Adding serial correlation to the LMM can help to statistically adjust for a potential source of random effect misspecification and help alleviate severe empirical type I error rate inflation. Finally, this meta-analysis also found that simpler serial correlation structures (AR(1)) perform just as well as more complicated structures (ARMA(1, 1)) as shown in Figure 3 and Figure 4.

Finally, statistical tests for the random effects also show evidence of being too liberal on average (see Table 5), particularly with few fixed effects included in the model (see Figure 5). This can be problematic if applied researchers are counting on these tests for inclusion of random effects in the LMM. However, as shown by the fixed effect results, it is extremely problematic to underspecify the random effect structure; therefore, including more random effects than needed is likely less problematic as long as the model still converges. Research into this would be worthy of further study.

Informing Future MC Studies

This meta-analysis can help inform additional areas of study. None of the MC studies allowed for variables to vary besides the initial status and linear trend. The implications for the empirical type I error rates when the trajectory of additional variables is allowed to vary between clusters would be interesting to explore, especially in relation to cases when the random structure is misspecified.

Second, the number of fixed effects in the MC studies coded were homogeneous and were not significant in the final inferential model. As can be seen in Table 1, most studies had between four and six fixed effects and only a single study included more with 10 fixed effects. Many applied studies using the LMM tend to have more than four to six fixed effects, for example, Harwell, Post, Medhanie, Dupuis, and LeBeau (2013) has 33 covariates included in a three-level longitudinal LMM. Better understanding the implications when there are many fixed effects related to assumption violations and small sample size conditions would be a welcome addition to the MC literature. Relatedly, little attention in the MC LMM literature has been given to three-level models and would be worthy of more attention.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

References marked with an asterisk indicate studies included in the meta-analysis

Aloe, A. M., Becker, B. J., & Pigott, T. D. (2010). An alternative to R2 for assessing linear models of effect size. *Research Synthesis Methods*, *1*, 272-283.

*Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics*, *38*, 1845-1865.

*Browne, W., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391-420.

*Browne, W., Draper, D., Goldstein, H., & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics & Data Analysis*, *39*, 203-225.

Browne, W., & Goldstein, H. (2010). MCMC sampling for a multilevel model with nonindependent residuals within and between cluster units. *Journal of Educational and Behavioral Statistics*, *35*, 453-473.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

*Delpish, A. (2006). *Comparison of estimators in hierarchical linear modeling: Restricted maximum likelihood versus bootstrap via minimum norm quadratic unbiased estimators* (Doctoral dissertation). Florida State University, Tallahassee.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford, UK: Oxford University Press.

*Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, *37*, 379-403.

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley-IEEE.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, *21*, 607-611.

Goldstein, H. (2010). *Multilevel statistical models*. West Sussex, UK: Wiley.

Harwell, M. R., Post, T. R., Medhanie, A., Dupuis, D. N., & LeBeau, B. (2013). A multi-institutional study of high school mathematics curricula and college mathematics achievement and course taking. *Journal for Research in Mathematics Education*, *44*, 742-774.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*, 315-339.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, *29*, 122-126.

Kwok, O., West, S., & Green, S. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, *42*, 557-592.

*Kwon, H. (2011). *A Monte Carlo study of missing data treatments for an incomplete level-2 variable in hierarchical linear models* (Doctoral dissertation). The Ohio State University, Columbus.

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-974.

*LeBeau, B. (2012). *The impact of ignoring serial correlation in longitudinal data analysis with the linear mixed model: A Monte Carlo study* (Doctoral dissertation). University of Minnesota, Twin Cities.

*LeBeau, B. (2013, April). *Impact of non–normal level one and two residuals on the linear mixed model*. Paper presented at the American Educational Research Association Conference, San Francisco, CA.

*Maas, C., & Hox, J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*, 427-440.

*Maas, C., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86-92.

*Mallinckrodt, C. H., Clark, W. S., & David, S. R. (2001). Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Information Journal*, *35*, 1215-1225.

Miller, J. J. (1978). The inverse of the Freeman–Tukey double arcsine transformation. *The American Statistician*, *32*, 138-138.

*Murphy, D., & Pituch, K. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education*, *77*, 255-284.

*Overall, J. E., & Tonidandel, S. (2010). The case for use of simple difference scores to test the significance of differences in mean rates of change in controlled repeated measurements designs. *Multivariate Behavioral Research*, *45*, 806-827.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*, 287-312.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. *The Handbook of Research Synthesis and Meta-Analysis*, 2, 295-316.

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

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from https://www.R-project.org/

Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo method* (Vol. *10*). Hoboken, NJ: John Wiley.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*, 137-167.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1-34.
Yee, T. W. (2015). *Vector generalized linear and additive models: With an implementation in R*. New York, NY: Springer.

Author Biographies

Brandon LeBeau is an assistant professor of Educational Statistics and Measurement at the University of Iowa. His research interests include longitudinal data, reproducible workflows, and research software development.

Yoon Ah Song is a graduate student at the University of Iowa in the Educational Measurement and Statistics program. She is interested in item response theory and linear mixed models.

Wei Cheng Liu is a graduate of the Educational Measurement and Statistics program at the University of Iowa. He is interested in program evaluation.