Original Article

How Low Can You Go?

An Investigation of the Influence of Sample Size and Model Complexity on Point and Interval Estimates in Two-Level Linear Models

Bethany A. Bell,[1] Grant B. Morgan,[1] Jason A. Schoeneberger,[1]
Jeffrey D. Kromrey,[2] and John M. Ferron[2]

[1]University of South Carolina, SC, USA, [2]University of South Florida, FL, USA

Abstract. Whereas general sample size guidelines have been suggested when estimating multilevel models, they are only generalizable to a relatively limited number of data conditions and model structures, both of which are not very feasible for the applied researcher. In an effort to expand our understanding of two-level multilevel models under less than ideal conditions, Monte Carlo methods, through SAS/IML, were used to examine model convergence rates, parameter point estimates (statistical bias), parameter interval estimates (confidence interval accuracy and precision), and both Type I error control and statistical power of tests associated with the fixed effects from linear two-level models estimated with PROC MIXED. These outcomes were analyzed as a function of: (a) level-1 sample size, (b) level-2 sample size, (c) intercept variance, (d) slope variance, (e) collinearity, and (f) model complexity. Bias was minimal across nearly all conditions simulated. The 95% confidence interval coverage and Type I error rate tended to be slightly conservative. The degree of statistical power was related to sample sizes and level of fixed effects; higher power was observed with larger sample sizes and level-1 fixed effects.

Keywords: Monte Carlo, multilevel models, sample size

Hierarchically organized data are commonplace in educational, clinical, and other settings in which research often occurs. Students are nested within classrooms or teachers, and teachers are nested within schools. Alternatively, service recipients are nested within social workers providing services, who may in turn be nested within local civil service entities. Conducting research at any of these levels while ignoring the more detailed levels (students) or contextual levels (schools) can lead to erroneous conclusions. As such, multilevel models have been developed to properly account for the hierarchical (correlated) nesting of data (Heck & Thomas, 2000; Hox, 2002; Klein & Kozlowski, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

Multilevel models can be conceptualized as regression models at two different levels: level-1 (e.g., students) and level-2 (e.g., schools). For example, if a researcher was interested in examining how student and school variables influenced students' reading achievement, he/she could use a two-level model with student variables at level-1 and school variables at level-2. The equations necessary for estimating the multilevel model are presented below.

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}. \quad (1)$$

Equation 1 represents a simple level-1 model with one student-level predictor where Y_{ij} is the reading achievement for student i in school j, β_{0j} is the average reading achievement for school j, X_{ij} is the centered student-level predictor for student i in school j, and β_{1j} is the slope or regression coefficient associated with X_{ij}, thus this value depicts the relationship between the student-level variable and reading achievement. The last term in the equation, r_{ij}, is the student-level error term, which is assumed to be normally distributed with covariance R.

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + \mu_{0j} \\ \beta_{1j} &= \gamma_{10}\end{aligned}. \quad (2)$$

Equation 2 is for the simple level-2 model with one school-level predictor where γ_{00}, is the intercept, which represents the grand mean of reading achievement across students and across schools, W_j is the centered school-level predictor for school j, and γ_{01} is the regression coefficient associated with W_j, μ_{0j} is an error term representing a unique effect associated with school j, and γ_{10} estimates the average effect of the student-level predictor. The absence of an error term in the equation for β_{1j} indicates that the effect of the student-level predictor is fixed, or held constant across schools. The level-2 errors are assumed to be normally distributed with covariance G.

$$\mathrm{Y}_{ij} = \gamma_{00} + \gamma_{01}\,\mathrm{W}_j + \gamma_{10}\,\mathrm{X}_{ij} + \mu_{0j} + r_{ij}. \quad (3)$$

Then, by substituting the values of β_{0j} and β_{1j} from the level-2 equation into the level-1 equation, the combined level-1 and level-2 model is created (Equation 3). From this

DOI: 10.1027/1614-2241/a000062

combined model, the regression element behind multilevel models becomes more apparent. Specifically, as shown in Equation 3, in this two-level model, there is the continuous reading outcome (Y_{ij}), an intercept (γ_{00}), level-1 and level-2 regression coefficients (γ_{10} and γ_{01}, respectively), and level-1 and level-2 error terms (r_{ij} and μ_{0j}, respectively).

Research has shown that ignoring a level of nesting in data can impact estimated variances and the available power to detect treatment or covariate effects (Donner & Klar, 2000; Julian, 2001; Moerbeek, 2004; Murray, 1998; Shadish, Cook, & Campbell, 2002), can seriously inflate the Type I error rate (Wampold & Serlin, 2000), and can lead to substantive errors in interpreting the results of statistical significance tests (Goldstein, 2003; Nich & Carroll, 1997). There are many types of multilevel models, which differ in terms of the number of levels (e.g., 2, 3), type of design (e.g., cross-sectional, longitudinal with repeated measures, cross-classified), scale of the outcome variable (e.g., continuous, categorical), and number of outcomes (e.g., univariate, multivariate). These models have been used to address a variety of research questions involving model parameters that include fixed effects (e.g., average student socioeconomic status-mathematics achievement slope across schools), random level-1 coefficients (e.g., student socioeconomic status-mathematics achievement slope at a particular school), and variance-covariance components (e.g., amount of variation in the student socioeconomic status-mathematics achievement slope across schools).

As the use of multilevel models has expanded into new areas, questions have emerged concerning how well these models work under various design conditions. One of these design conditions is sample size at each level of the analysis. This issue is central in most quantitative studies but is more complex in multilevel models because of the multiple levels of analysis. Currently there are few sample size guidelines referenced in the literature. One rule of thumb, proposed for designs in which individuals are nested within groups, calls for a minimum of 30 units at each level of the analysis. This rule of thumb is commonly cited (see, e.g., Hox, 1998; Maas & Hox, 2004, 2005) and was further developed by Hox (1998) who recommended a minimum of 20 observations (level-1) for 50 groups (level-2) when examining interactions across levels. For multilevel structural equation modeling, Hox and Maas (2001) recommend at least 100 groups (level-2). Although many researchers attempt to adhere to these sample size guidelines, practical constraints in applied research (e.g., financial costs, time) often make these sample size recommendations at one or more of the levels of analysis difficult to achieve. For example, in school effects research, recruiting and obtaining the cooperation of individual schools can be labor intensive and expensive. Once a school agrees to participate, however, it is often easy to obtain many level-1 units (e.g., students). In other cases, obtaining a large number of level-2 units (e.g., families) is straightforward, but obtaining sufficient numbers of level-1 units (e.g., family members) may be difficult or impossible. Still in other cases it may be difficult to obtain large numbers of observations at both level-1 and level-2. A review of multilevel studies in education and the social sciences (Dedrick et al., 2009) bears out the difficulties involved in achieving sample size guidelines in applied settings. This review of 99 multilevel studies from 13 peer-reviewed journals (1999–2003) identified three studies with $\leq$ 30 level-2 units and $\leq$ 30 level-1 units, three studies with $\leq$ 30 level-2 units and > 30 level-1 units, and 15 studies with > 30 level-2 units and $\leq$ 30 level-1 units.

Given the reality of small sample sizes, several simulation studies have been designed to examine the effect of small sample sizes on various multilevel results (e.g., variance estimates, fixed effects estimates, standard errors, model convergence). Results from these studies vary based on the nature of the effect being examined (i.e., fixed vs. random). For instance, with random effects, findings from simulation studies that have focused on small sample sizes, at one or both levels of two-level models, suggest that the overall functioning of random effects with small sample sizes is less than ideal (Bell, Ferron, & Kromrey, 2009; Clarke & Wheaton, 2007; Maas & Hox, 2004, 2005; Mok, 1995; Newsom & Nishishiba, 2002). Similar findings have been noted in studies examining sample size using multilevel analyses for longitudinal research (De Jong, Moerbeek, & Van Der Leeden, 2010). Specifically, researchers have noted positive parameter bias in the intercept, slope, and residual variances (Bell et al., 2009; Clarke & Wheaton, 2007; Mok, 1995; Snijders, 2005) as well as convergence difficulties and relatively poor confidence interval coverage (Maas & Hox, 2004; Newsom & Nishishiba, 2002) of random effects with small level-1 sample sizes. Browne and Draper (2000) report biased variance component standard error estimates, even with upward of 50 level-2 groups. In an investigation of sample sizes in the context of multilevel logistic regression, Moineddin, Matheson, and Glazier (2007) recommended minimum sample sizes of 50 at levels-1 and -2. They report convergence problems with small sample sizes and low prevalence of the outcome variable in logistic regression but unbiased fixed effect parameter estimates in moderate or larger sample sizes.

Although there appear to be substantial problems in making variance inferences from small samples, results of simulation studies regarding fixed effects have been more encouraging. For instance, studies have consistently shown little to no bias in the estimates of the fixed effects, regardless of level-1 or level-2 sample size (Bell, Ferron, & Kromrey, 2008; Bell et al., 2009; Clarke, 2008; Clarke & Wheaton, 2007; Hess, Ferron, Bell Ellison, Dedrick, & Lewis, 2006; Maas & Hox, 2004; Mok, 1995; Newsom & Nishishiba, 2002). The same general pattern has also been observed for the standard errors of the fixed effects, with a few exceptions. For example, some studies have shown bias in the standard errors of the fixed effects, thus decreasing average coverage rates of the 95% confidence intervals at some extreme sample size conditions (e.g., 50 level-2 units with large proportions of singletons; Bell et al., 2008, 2009; Maas & Hox, 2004).

Although the findings related to fixed effects and small sample sizes are generally encouraging, the majority of studies have only examined relatively simple models. For example, Clarke and Wheaton (2007), Maas and Hox (2004, 2005), and Hess et al.'s (2006) findings were based on simple two-level hierarchical models with one continuous criterion variable, one predictor variable at each level, one cross-level interaction between the predictors, and two random effects (intercept and single level-1 predictor). Determining the precise impact of a covariate on power a priori depends on a number of factors, including the amount of within and between variance the predictor accounts for and how multiple covariates interact (Moerbeek, 2006; Reise & Duan, 2003). Because findings from simulation studies are not generalizable beyond the models and conditions examined in the studies, simulation studies need to investigate how such models function under a wide variety of realistic model and data conditions. One such condition is the inclusion of binary predictor variables (e.g., sex, treatment vs. control group assignment, minority status), which are now commonly used in the regression framework for comparing groups, following Cohen (1968). Given the prevalence with which binary independent variables are used in regression analysis, researchers should develop a deeper understanding of how these variables function in multilevel analysis.

This study focused on the consequences of small level-1 and level-2 sample sizes on the estimation of fixed effect inferences in two-level linear multilevel models in which individuals were nested within groups. Monte Carlo methods were used to examine convergence rates, non-positive definite G-matrix rates, point estimates (statistical bias) and interval estimates (confidence interval accuracy and precision), and Type I error control and statistical power of tests associated with the fixed effects as a function of level-1 sample size, level-2 sample size, intercept variance, slope variance, collinearity, and model complexity. By examining more complex multilevel models (i.e., two-level models with various numbers of predictors, various levels of collinearity, and binary and continuous predictors at each level), this study adds information about the accuracy and precision of estimates and contributes to our understanding of the behavior of multilevel models under less than ideal conditions.

Method

For this Monte Carlo study the following design factors and conditions were examined:

(a) level-1 sample sizes (with n_j randomly selected from the intervals 5–10, 10–20, and 20–40),
(b) level-2 sample sizes with conditions of 10, 20, and 30,
(c) intercept variance (.10 or .30),
(d) slope variance (.0 or .30),
(e) levels of collinearity (level-1 and level-2 population correlation between regressors of 0, .10, .30 and cross-level population correlation between regressors of .0 and .1), and
(f) model complexity with two and three level-1 predictors crossed with two and three level-2 predictors for both main effect and three different interaction models (level-1 interaction, level-2 interaction, and cross-level interaction).

These factors in the Monte Carlo study were completely crossed, yielding nine sample size conditions and 1,152 design factor conditions.

Data were generated based on a two-level model in which observations were nested within groups. The main effects model is shown in Equation 4. At the first level, a continuous outcome was modeled as a linear function of k predictors, where $k = 2$ or 3. At the second level, the intercepts and slopes of the first level, which were allowed to vary randomly, were modeled as a function of m predictors where $m = 2$ or 3.

$$Y_{ij} = \gamma_{00} + \sum_{m=1}^{M} \gamma_{0m} W_{mj} + \sum_{k=1}^{K} \gamma_{k0} X_{kij} + \mu_{0j} + \sum_{k=1}^{K} \mu_{kj} X_{kij} + r_{ij}. \quad (4)$$

In each model, one level-1 predictor and one level-2 predictor were binary and all others were continuous. For each k, m predictor combination, four models were examined: a main effect model (shown in Equation 4), a level-1 interaction model, a level-2 interaction model, and a cross-level interaction model, yielding a total of 16 different models. The interaction models were generated by adding a term to Equation 4 based on the product of the first level-1 and first level-2 predictors (W_1X_1), the product of the first two level-1 predictors (X_1X_2), or the product of the first two level-2 predictors (W_1W_2). The level-1 errors were generated from a normal distribution with a variance of 1.0 using the RANNOR random number generator in SAS version 9.2 (SAS Institute Inc., Cary, NC, 2009). The level-2 errors were also generated from a normal distribution, but with a variance of .10 or .30 for μ_0 (intercept), a variance of .0 or .10 for μ_{kj} (slopes), and no covariance among the level-2 errors (i.e., a diagonal G-matrix).

The data were simulated such that one predictor at each level had no effect ($\gamma = 0$; for estimation of Type I error rate) and all the remaining predictors had non-null effects ($\gamma \neq 0$; for estimating statistical power). To yield statistical power of approximately .80 when level-1 sample size ranged from 20 to 40, level-2 sample size equaled 30, slope and intercept variance equaled .1, and all correlations between predictors equaled .1, fixed effects for each of the k, m predictor combinations of 2, 2; 2, 3; 3, 2; and 3, 3 were assigned γ of 0.45, 0.42, 0.39, or 0.38, respectively. The first predictor at both level-1 and level-2 was transformed into a binary variable by generating a standard normal distribution using the RANNOR random number generator in SAS version 9.2 (SAS Institute Inc., Cary, NC, 2009) such that values below the mean were recoded as "0" and values above the mean were recoded as "1."

For each of the 10,368 conditions (nine sample size combinations × 1,152 combinations of design factors), 1,000 data sets were simulated using SAS IML (SAS Institute Inc., Cary, NC, 2008). The data simulation program was checked by examining the matrices produced at each stage of data generation. After each data set was generated, the simulated sample was analyzed using a two-level multilevel model with restricted maximum likelihood estimation and *Kenward-Roger* degrees of freedom estimation via the

MIXED procedure in SAS (SAS Institute Inc., Cary, NC, 2003). Because the focus of our study was on the functioning of the estimates and tests of parameters for fixed effects, we chose to use the covariance structures that are commonly used by applied researchers who focus on fixed effects. Specifically, in all analyses the covariance matrix of the level-2 errors, G-matrix, was modeled to be diagonal (i.e., to have separate variance estimates but no covariances) and the covariance matrix of the level-1 errors, R-matrix, was modeled to have a common variance with no covariances. These choices make the analyses consistent with the commonly used default settings within the MIXED PROCEDURE of SAS. Although these restrictions help keep the model complexity focused on the fixed effects, they also limit the generalizability of the results to instances where level-2 error terms do not covary and level-1 errors are homogeneous.

Seven outcomes were examined in this Monte Carlo study:
(a) bias in the estimates of the fixed effects,
(b) rate of model convergence,
(c) rate of non-positive definite G-matrices,
(d) Type I error rates for null fixed effects,
(e) average confidence interval width for each fixed effect,
(f) average confidence interval coverage for each fixed effect, and
(g) statistical power for non-null fixed effects.

In our study, *coverage* refers to the proportion of replications in which the produced confidence interval contains the true value of the parameter in the population from which the replications were drawn. With a defined Type I error rate of 5%, the expected confidence interval coverage rate is .95 (i.e., one minus Type I error rate) in this study. Confidence interval *width* is directly related to standard error; smaller standard errors result in narrower confidence intervals (i.e., more estimation precision), and larger standard errors result in wider confidence intervals (i.e., less estimation precision). Given that smaller standard errors are more desirable, narrower confidence interval widths (i.e., widths closer to zero) are favorable.

The results were analyzed by describing the distributions of these outcomes rather than testing null hypotheses about their values. Conducting hypothesis tests on simulation study results typically does not contribute to interpretability because nearly all mean differences among the design factors will be statistically significant (Burton, Altman, Royston, & Holder, 2006). For example, with our 1,000 replications per condition, more than 10 million data points comprise the results of the current study.

Results

Initial inspection of the statistical bias of the fixed effect estimates suggested that main effect, level-1 interaction, and cross-level interaction models did not evidence substantial bias; however, bias from the level-2 interaction model fixed effects was heavily influenced by a few extreme values (binary predictors: $M = -0.0074$, $SD = 0.30$, min = -2.23, max = 14.29; continuous predictors: $M = -0.0027$, $SD = 0.08$, min = -3.04, max = 0.87). Among the bias estimates of the binary predictors, 9 of 2,592 level-2 interaction conditions were greater than ±3 standard deviations from the mean, and 16 of 2,592 level-2 interaction conditions had bias estimates greater than ±3 standard deviations from the mean for continuous predictors. Each of the extreme conditions was in the smallest level-2 sample size condition (i.e., $n_2 = 10$). The conditions under which each outlier was observed are presented in Table 1. To address these outliers, and to better understand the nature of our data, 90% winsorization was carried out such that the lower and upper 5% of bias values were set equal to the 5th and 95th percentiles, respectively. Subsequent inspection of the bias of the fixed effects from level-2 interaction models was not suggestive of substantial bias. Thus, overall statistical bias estimates of binary predictors across the four models were not problematic as each of the four models' mean bias was less than 0.0003 in absolute value, with minimum of -0.013 and maximum of 0.012. Similarly, observed bias of continuous variables was not problematic across models as each model's mean bias was less than 0.0002 in absolute value, with minimum of -0.006 and maximum of 0.006.

Convergence was also not viewed as problematic as 100% convergence was obtained under 99.19% of the conditions studied. Of the conditions that did not have perfect convergence rates, 85.71% had a convergence rate of 99.9%.

A non-positive definite G-matrix indicates that one or more of the variance components of random effects is/are estimated to be zero. The frequency with which such G-matrices were observed among replications within each condition was recorded. Rates of non-positive definite G-matrices are presented by level-1 sample size by slope variance by level-2 sample size in Figure 1. In the condition where level-1 and -2 sample sizes were maximized and slope variance was set at 0.3, the rate of non-positive definite G-matrices approached zero ($M = 0.015$) with very little variability. Under the same slope variance ($\tau_{11} = 0.3$) and level-2 sample size ($n_2 = 30$) condition, the mean rate of non-positive definite G-matrices was .252 in the smallest level-1 sample size condition. Next, holding slope variance at 0.3, the mean rate of non-positive definite G-matrices for the smallest level-1 ($n_1 = 5$–10) and smallest level-2 sample size category ($n_2 = 10$) was .677. These observations suggest that, even when the data were generated and modeled to vary randomly, models with the smallest sample sizes at each level only correctly estimated a positive definite G-matrix roughly one-third of the time. Conversely, when the variance components were generated to be null in the population ($\tau_{11} = 0$) but were modeled to vary randomly, the average rate of non-positive definite G-matrices was .850 in the largest sample size conditions ($n_1 = 20$–40, $n_2 = 30$). Thus, even with 20–40 level-1 units and 30 level-2 units, on average 15% of the time, nonzero variance components were estimated for null effects.

Table 1. Replications with outlying bias estimates for level-2 interaction model estimates by condition

Bias estimate										
Binary	Continuous	n_1	n_2	τ_{00}	τ_{11}	Corr. *X*s	Corr. *Z*s	Corr. $X \times Z$	No. of *X*s	No. of *Z*s
−0.91	0.48	5–10	10	0.1	0.3	0.3	0.1	0.0	2	2
1.06	−0.43	5–10	10	0.3	0.3	0.0	0.0	0.1	3	2
2.58	−1.30	10–20	10	0.1	0.0	0.0	0.1	0.0	2	2
14.29	−3.04	10–20	10	0.1	0.0	0.1	0.3	0.0	3	3
1.92	−0.75	10–20	10	0.1	0.3	0.0	0.1	0.1	3	3
−2.23	0.87	10–20	10	0.1	0.3	0.3	0.0	0.1	2	3
2.08	−0.56	10–20	10	0.3	0.0	0.0	0.3	0.1	3	3
−0.93	0.30	20–40	10	0.1	0.3	0.1	0.1	0.0	3	3
1.08	−0.54	20–40	10	0.3	0.0	0.3	0.1	0.1	2	2
−0.05*	−0.75	5–10	10	0.1	0.0	0.3	0.0	0.1	3	3
−0.24*	−0.48	5–10	10	0.1	0.3	0.0	0.1	0.1	2	3
0.67*	−0.86	5–10	10	0.3	0.3	0.0	0.1	0.0	2	2
−0.83*	0.29	5–10	10	0.3	0.3	0.3	0.1	0.0	2	3
0.79*	−0.27	10–20	10	0.1	0.3	0.1	0.3	0.0	3	3
0.44*	−0.25	10–20	10	0.3	0.3	0.1	0.0	0.1	2	2
−0.03*	0.43	20–40	10	0.3	0.3	0.1	0.3	0.0	3	2

Notes. Corr. *X*s refers to the correlation between level-1 predictors; Corr. *Z*s refers to the correlation between level-2 predictors; Corr. *X* × *Z* refers to the correlation between level-1 and level-2 predictors. *Estimate did not exceed ± 3 standard deviations.

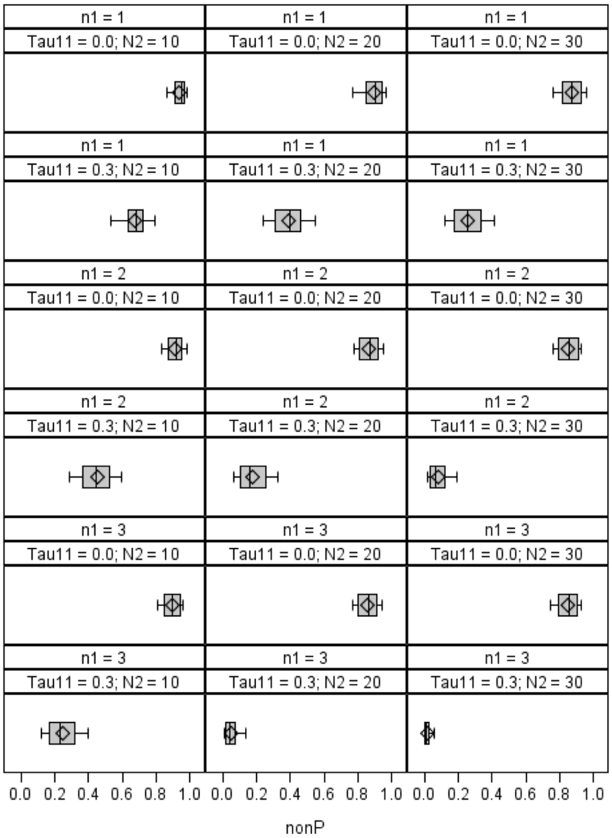

Figure 1. Rates of nonpositive G-matrices by level-1 (n_1) sample size, slope variance (τ_{11}), and level-2 (n_2) sample sizes. "n_1 = 1," "n_1 = 2," and "n_1 = 3" reflect level-1 sample size ranges of 5–10, 10–20, and 20–40, respectively.

The average rate of positive definite G-matrices for models with null variance components did not vary much within sample size combinations; the standard deviation within each of the 18 level-1 sample sizes by level-2 sample size by slope variance conditions was .089 or less. Moreover, non-positive definite G-matrices for null variance components occurred most frequently under the smallest sample size condition of 5–10 level-1 units and 10 level-2 units (M = 0.94).

Overall, the estimated Type I error rates for tests of the fixed effect regression parameters of two continuous predictors were close to the nominal alpha level ($M = 0.046$, min = 0.021, max = 0.068). Thus, although there was some variability in Type I error rates, overall, they did not appear to be overly influenced by any of the design factors included in the study.

Estimates of the 95% confidence interval widths across models were also not considered problematic. The minimum and maximum widths for binary variables across models were 0.597 and 2.685, respectively. The cross-level interaction model had the highest mean width ($M = 1.49$) followed by the level-2 interaction ($M = 1.37$), the level-1 interaction ($M = 1.18$), and the main effect ($M = 1.17$) models. Among continuous predictors, the largest width was observed in the level-2 interaction model ($M = 1.20$) followed by cross-level interaction ($M = 0.76$), main effects ($M = 0.76$), and level-1 interaction ($M = 0.72$) models. The minimum and maximum widths for continuous variables across models were 0.354 and 2.237, respectively.

The distributions of estimated 95% confidence interval coverage for level-1 and level-2 fixed effect parameters are presented in Figures 2 and 3, respectively. For the level-1 fixed effects, all models provided near nominal level coverage for the majority of conditions examined.

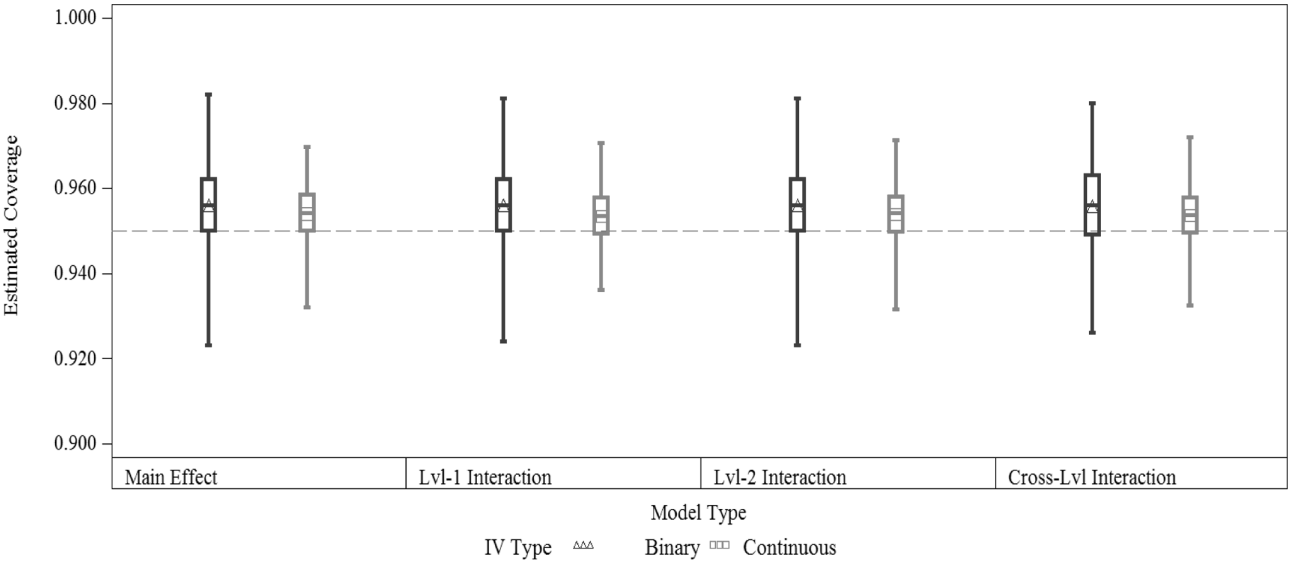

Figure 2. 95% confidence interval coverage for level-1 binary and continuous fixed effects by model type.

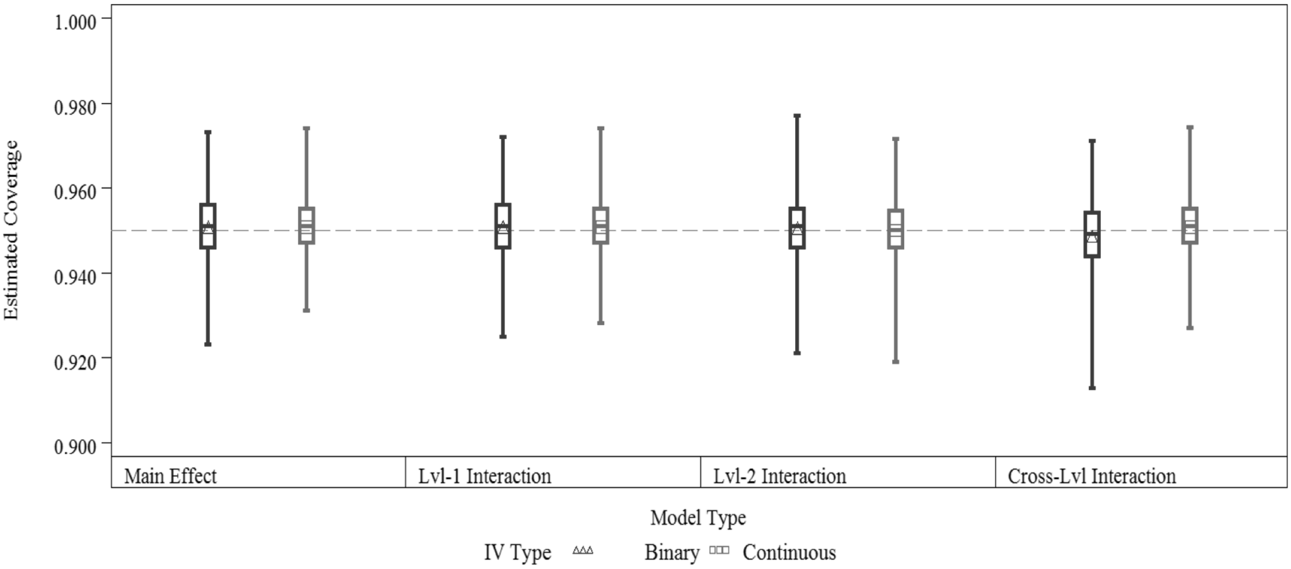

Figure 3. 95% confidence interval coverage for level-2 binary and continuous fixed effects by model type.

Specifically, the mean coverage estimates of binary and continuous level-1 fixed effects across models were .956 and .954, respectively, whereas the 95% confidence interval coverage estimates for continuous predictors were slightly less variable than for those of binary predictors. The mean coverage estimates for binary level-1 predictors in main effect, level-1 interaction, level-2 interaction, and cross-level interaction models were each .956. The mean coverage estimates for continuous level-1 predictors in main effect, level-1 interaction, level-2 interaction, and cross-level interaction models were each .954. Coverage estimate distributions for level-1 predictors are presented in Figure 2.

Among level-2 fixed effects, the mean 95% confidence interval coverage estimates in all models were also very near nominal levels whether the predictors were binary or continuous; the mean coverage estimates of binary and continuous level-2 fixed effects across models were .950 and .951, respectively. For binary level-2 predictors, the mean coverage estimates in main effect, level-1 interaction, level-2 interaction, and cross-level interaction models were .951, .951, .950, and .949, respectively. The mean coverage estimates for continuous level-2 predictors in main effect, level-1 interaction, level-2 interaction, and cross-level interaction models were .951, .951, .950, and .951, respectively. Thus, coverage of level-1 and level-2 fixed effects was not viewed as a problem. Coverage estimate distributions for level-2 predictors are presented in Figure 3.

As with the level-1 and level-2 predictors, the 95% confidence interval coverage for the cross-level interaction fixed effect was also near nominal levels (data not shown). The effect represented the interaction between the binary level-1 and level-2 predictors and had an observed mean of 0.956, minimum of 0.928, and maximum of 0.984.

On the whole, the mean estimates of statistical power for binary ($M = 0.390$, min = 0.065, max = 0.941) and continuous predictors ($M = 0.553$, min = 0.113, max = 0.996) fell below the typically desired power of .80. However, varying levels of estimated power were observed across conditions in the study. Of particular interest was the estimated power across level-1 and level-2 sample sizes by the level and type of predictor. That is, power estimates were examined separately for binary and continuous predictors at levels-1 and -2 (see Figures 4 and 5 below). As one might expect, power estimates for level-1 predictors (both binary and continuous) increased as level-1 and level-2 sample sizes increased. Yet, for the smallest level-1 sample size of 5–10 units, the estimated power rarely reached the .80 level, even as level-2 sample size increased. Not until level-1 and level-2 sample sizes reached 20–40 and 30, respectively, did estimated power reach a mean at or above .80.

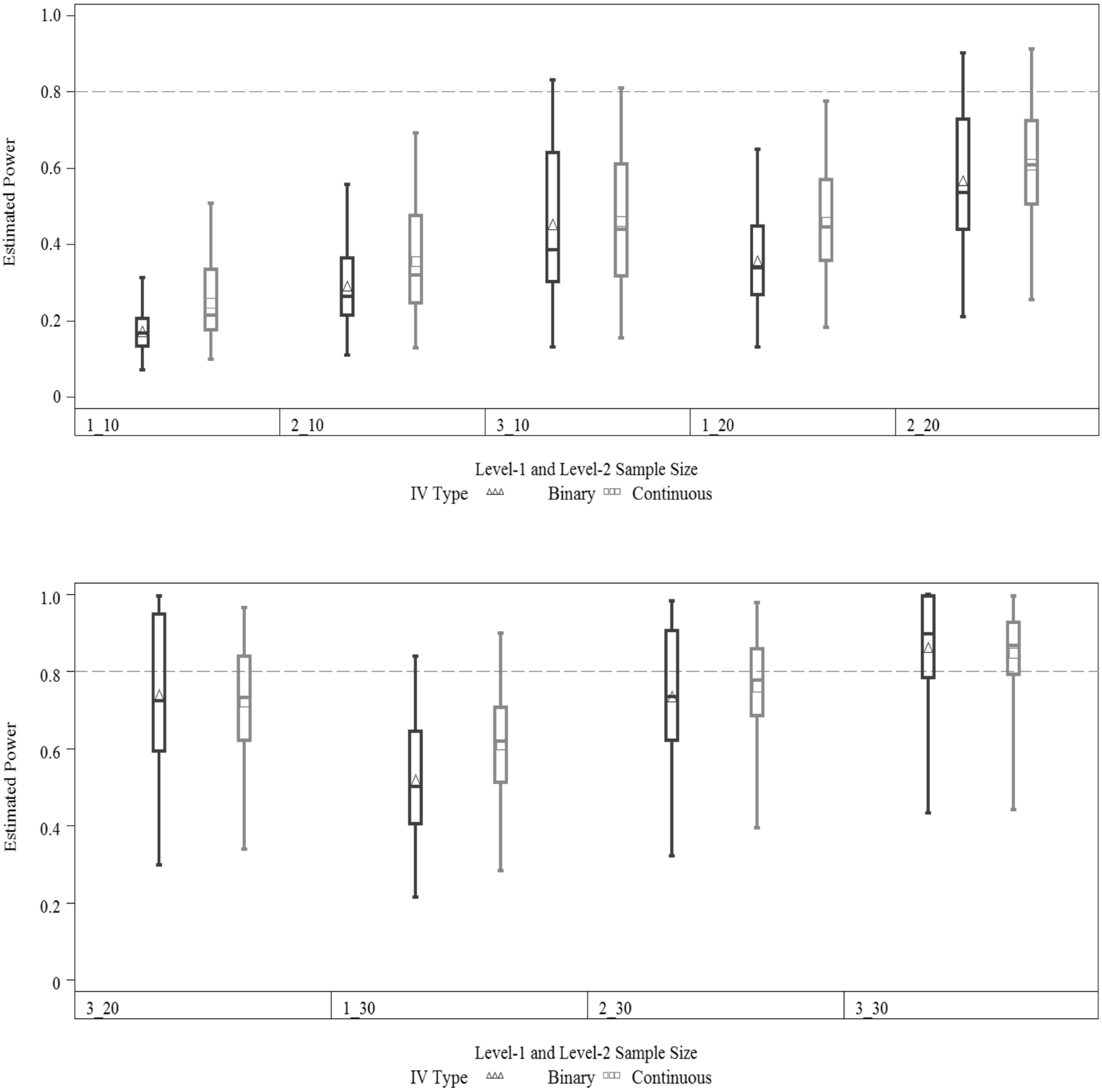

Figure 4. Power distributions of binary and continuous level-1 fixed effects by level-1 and level-2 sample size. Sample size indicators of 1, 2, and 3 reflect the level-1 sample size ranges of 5–10, 10–20, and 20–40, respectively. Sample size indicators of 10, 20, and 30 reflect level-2 sample sizes.

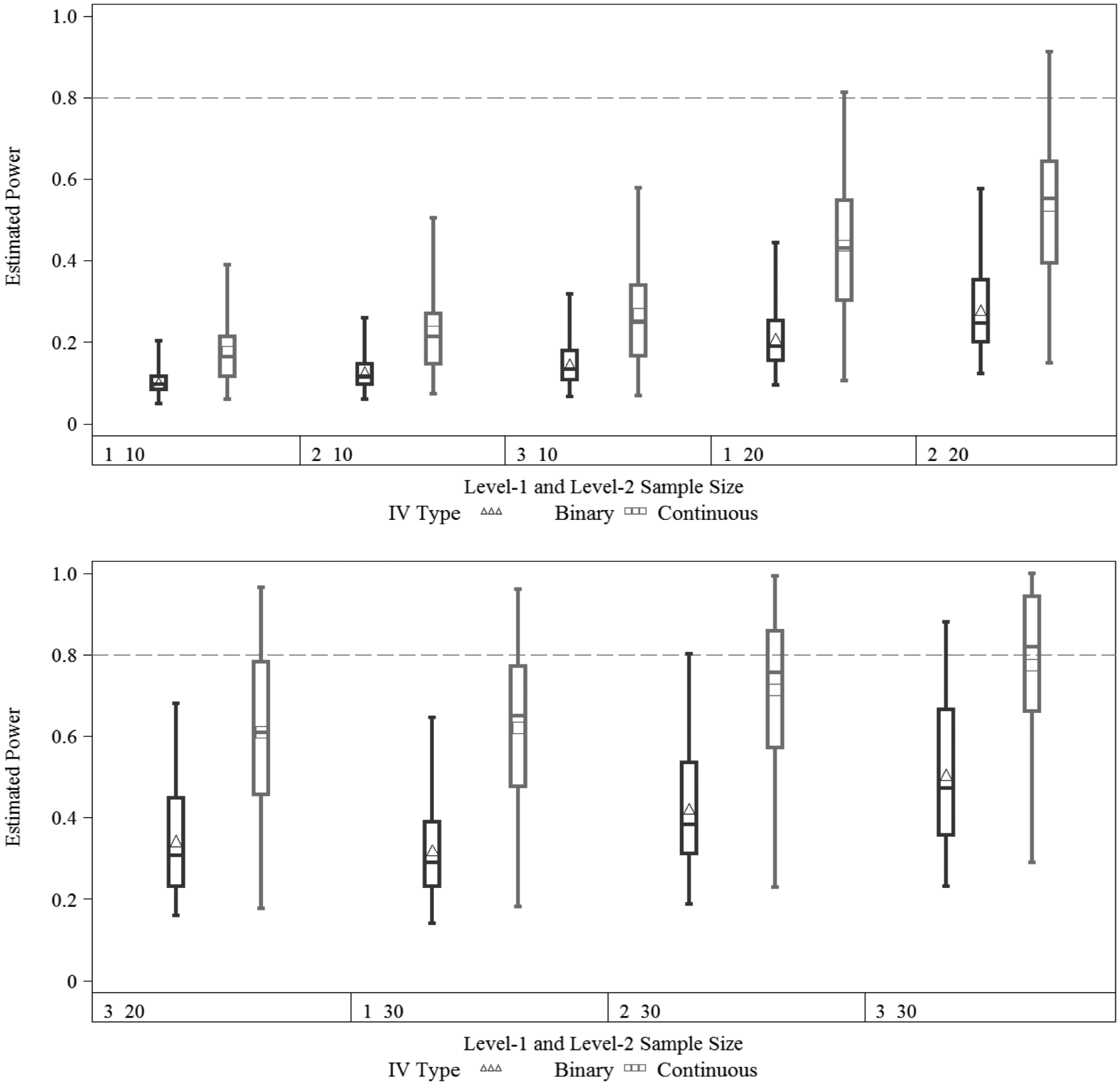

Figure 5. Power distributions of binary and continuous level-2 fixed effects by level-1 and level-2 sample sizes. Sample size indicators of 1, 2, and 3 reflect the level-1 sample size ranges of 5–10, 10–20, and 20–40, respectively. Sample size indicators of 10, 20, and 30 reflect level-2 sample sizes.

As shown in Figure 5, the mean statistical power estimates for level-2 predictors did not reach the .80 level in any of the sample size categories although, as expected, it did improve with larger sample sizes at each level. When level-2 sample size was set at 10, mean power estimates did not exceed .270, regardless of level-1 sample size or predictor type (i.e., binary or continuous). When level-1 sample size ranged from 5 to 10, the mean power estimates for binary and continuous level-2 predictors were highest when level-2 sample size was set at 30 (M = 0.321 and 0.621, respectively). Across each sample size classification, the distribution of power estimates for continuous level-2 predictors had larger interquartile ranges than for binary level-2 predictors. Generally, the statistical power estimates of level-2 predictors appear to be more heavily influenced by sample sizes at each level than for level-1 predictors.

Statistical power estimates of cross-level interaction effects were less than favorable. As sample sizes at each level increased, estimated power likewise increased. How-

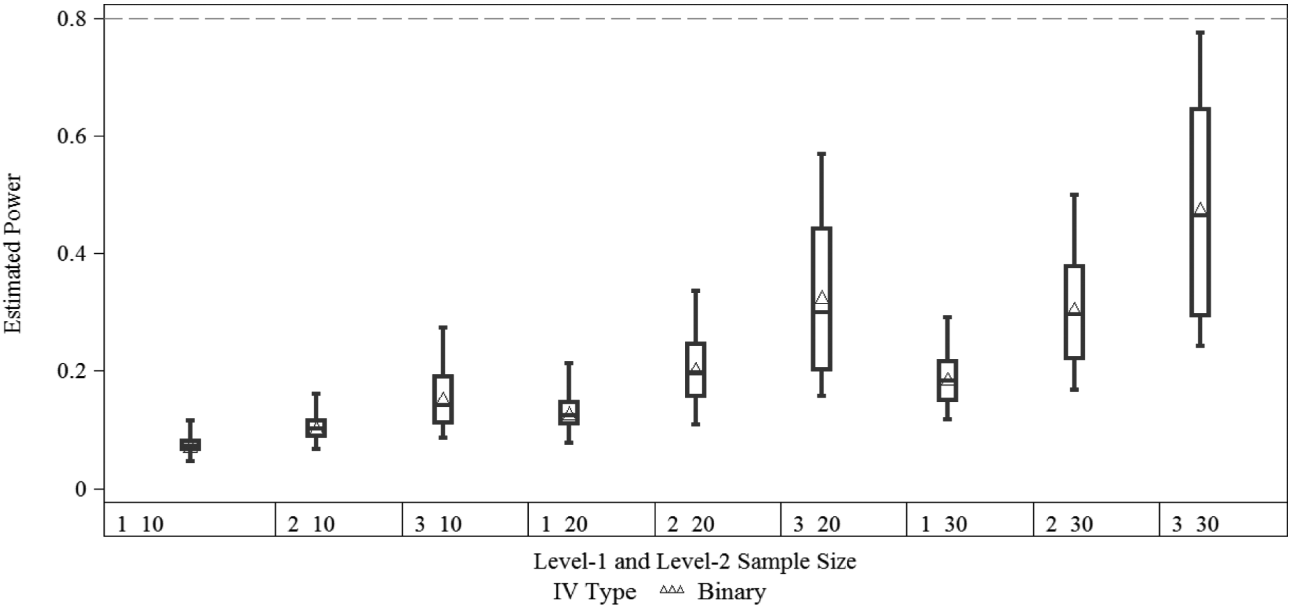

Figure 6. Power distributions of cross-level interaction fixed effects by level-1 and level-2 sample sizes. Sample size indicators of 1, 2, and 3 reflect the level-1 sample size ranges of 5–10, 10–20, and 20–40, respectively. Sample size indicators of 10, 20, and 30 reflect level-2 sample sizes.

ever, regardless of the level-1 and level-2 sample size combination, statistical power estimates of cross-level interactions failed to reach the .80 threshold. Figure 6 contains the distribution of statistical power estimates by sample size combinations for binary cross-level interaction effects.

Discussion

As stated previously, within the multilevel model framework, general recommendations for the minimum number of units at each level have been offered, despite their lack of feasibility in a variety of contexts. Moreover, to date, sample size recommendations that have been put forth are based on relatively simple models containing minimal numbers of continuous predictors at each level. In an effort to help applied researchers make sound decisions regarding the use of more complex two-level linear models with small sample sizes, the current study sought to build on the existing body of literature with specific regard to binary and continuous independent variables, level-1 and level-2 sample sizes, levels of collinearity among predictors, the number of predictors, and the type of model estimated (i.e., main effect, level-1 interaction, level-2 interaction, and cross-level interaction).

Three aspects of model complexity, the number of predictors at each level, the type of model estimated, and the correlation among predictor variables, did not pose substantial problems in any of the statistical outcomes examined in our study. However, slope variances impacted rates of nonpositive definite G-matrices, and samples sizes impacted power, as expected. In addition, estimates of power varied by predictor type.

After running all 10,368 conditions, what do we know? First, as found in previous studies, except for a handful of conditions with the smallest level-2 sample size, estimates of bias were not viewed as problematic regardless of sample size at each level. Similarly, by and large model convergence was not an area of concern, and surprisingly, Type I error rates were not substantially inflated across models and conditions, which may in part be attributable to using the Kenward-Roger adjusted degrees of freedom in the analyses. Thus, across the many design factors included in the current study, these findings suggest that bias was minimal, and that 95% confidence interval coverage and Type I error rates tend to be slightly conservative but are fairly well controlled even when modeling hierarchically structured data with smaller sample sizes.

Second, in terms of statistical power, the results are not quite as encouraging. The parameter value for the fixed effect (gamma) was set to obtain a power around .80. Often, the observed power never reached the typically desired level of .80 in conditions where sample sizes at levels-1 and -2 were limited. Setting gamma in this way ensured we could make power comparisons across types of predictors and models. The process of selecting gamma led to the use of values around 0.40 (for each of the k, m predictor combinations of 2, 2; 2, 3; 3, 2; and 3, 3 the assigned values of γ were 0.45, 0.42, 0.39, and 0.38, respectively). Coupling these values with the level-1 variance of 1.0 and the small level-2 variances would lead to standardized effect estimates in the small to medium range for the social sciences. Although the distributions of power estimates on occasion reached a power of .80, the more likely scenario suggests that adequate power will not be realized, given smaller level-1 and level-2 sample size combinations unless larger effect sizes are present. Moreover, estimated statistical power of level-2 predictors never reached a mean of .80 across any sample size combination. Thus, contradictory to what we saw regarding Type I error control, it appears that the commonly cited rule of 30 level-1 units and 30 level-2 units would likely not yield high levels of statistical power for the fixed effects at both levels of the model.

Third, the examination of the frequency with which non-positive definite G-matrices were produced provides insights that are consistent with some previous simulation studies. On the one hand, when slope variances were generated to be null ($\tau_{11} = 0$), non-positive definite G-matrices were expected, but not always obtained because sampling error would in some cases lead to non-zero variance estimates. Sampling error is greatest when the sample sizes are the smallest, and consequently positive definite G-matrices were obtained more frequently under the small sample size conditions. On the other hand, as we would expect, when slope variances were generated to vary randomly ($\tau_{11} = 0.3$), non-positive definite G-matrices were more frequently produced, with smaller level-1 sample sizes. Thus, even though we did not investigate the specific statistical properties of the random effects per se, these findings support previous research that suggests substantial bias in the estimates of random effects with small sample sizes (Bell et al., 2009; Clarke & Wheaton, 2007; Maas & Hox, 2004, 2005; Mok, 1995; Newsom & Nishishiba, 2002).

In conclusion, the current study adds to the understanding of the statistical considerations of multilevel modeling given a variety of conditions. The results provide applied researchers with valuable information regarding the impact of certain design factors on her or his results. However, as with all simulation research, it is important to remember that our findings are only generalizable to data conditions similar to those examined in the study. Nonetheless, in conjunction with findings from previous studies, it appears that researchers can more confidently apply multilevel modeling techniques with relatively small samples sizes, across a variety of model types, and make appropriate inferences regarding the point and interval estimates for fixed effects.

Acknowledgments

The authors would like to thank Brandon Loudermilk for his assistance with this study. In addition, the helpful comments and constructive critiques provided by the anonymous reviewers led to a much improved manuscript.

References

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008, August). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. In *Proceedings of the joint statistical meetings, survey research methods section* (pp. 1122–1129). Alexandria, VA: American Statistical Association.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2009, April). *The effect of sparse data structures and model misspecification on point and interval estimates in multilevel models.* Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*, 391–420. doi: 10.1007/s001800000041

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*, 4279–4292. doi: 10.1002/sim.2673

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health, 62*, 752–758. doi: 10.1136/jech.2007.060798

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Methods & Research, 35*, 311–351. doi: 10.1177/0049124106292362

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426–443. doi: 10.1037/h0026714

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

De Jong, K., Moerbeek, M., & Van Der Leeden, R. (2010). A prior power analysis in longitudinal three-level multilevel models: An example with therapist effects. *Psychotherapy Research, 20*, 273–284. doi: 10.1080/10503300903376320

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed). London: Edward Arnold.

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Erlbaum.

Hess, M. R., Ferron, J. M., Bell Ellison, B., Dedrick, R., & Lewis, S. E. (2006, April). *Interval estimates of fixed effects in multi-level models: Effects of small sample size.* Presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA.

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

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J. J., & Maas, C. J. M (2001). The accuracy of multilevel structural equation modeling with psuedobalanced groups and small samples. *Structural Equation Modeling, 8*, 157–174. doi: 10.1207/S15328007SEM0802_1

Julian, M. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8*, 325–352. doi: 10.1207/S15328007SEM0803_1

Klein, K. & Kozlowski, S. W. J. (Eds.). (2000). *Multilevel theory, research, and methods in organizations*. San Francisco, CA: Jossey-Bass.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*, 127–137. doi: 10.1046/j.0039-0402.2003.00252.x

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

Moerbeek, M. (2004). The consequences of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 129–149. doi: 10.1207/s15327906mbr3901_5

Moerbeek, M. (2006). Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine, 25*, 2607–2617. doi: 10.1002/sim.2297

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMS Medical Research Methodology, 7*, 1–10. doi: 10.1186/1471-2288-7-34

Mok, M. (1995). *Sample size requirements for 2-level designs in educational research.* Unpublished manuscript, Macquarie University, Sydney, Australia.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Newsom, J. T., & Nishishiba, M. (2002). *Nonconvergence and sample bias in hierarchical linear modeling of dyadic data.* Unpublished manuscript, Portland State University.

Nich, C., & Carroll, K. (1997). Now you see it, now you don't: A comparison of traditional versus random-effects regression models in the analysis of longitudinal follow-up data from a clinical trial. *Journal of Consulting and Clinical Psychology, 65*, 252–261. doi: 10.1037//0022-006X.65.2.252

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Newbury Park, CA: Sage.

Reise, S. P., & Duan, N. (2003). Design issues in multilevel studies. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: methodological advances, issues and applications* (pp. 285–298). Mahwah, NJ: Erlbaum.

SAS Institute Inc. (2003). *SAS, release 9.1* [computer program]. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2008). *SAS/IML® 9.2 User's guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., (2009). *SAS® 9.2 Language Reference: Dictionary* (2nd ed.). Cary, NC: SAS Institute Inc.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MD: Houghton Mifflin.

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 1570–1573). Chicester, UK: Wiley.

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

Wampold, B. E., & Serlin, R. C. (2000). The consequences of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5*, 425–433. doi: 10.1037//1082-989X.5.4.425

Received March 4, 2011
Accepted May 15, 2012
Published online July 20, 2012

Bethany A. Bell

University of South Carolina
College of Education
820 Main Street, Wardlaw #133
Columbia, SC 29208
USA
Tel. +803-777-2387
E-mail babell@sc.edu