Paper 197-2010

Dancing the Sample Size Limbo with Mixed Models: How Low Can You Go?

Bethany A. Bell, Grant B. Morgan, Jason A. Schoeneberger, Brandon L. Loudermilk
University of South Carolina
Jeffrey D. Kromrey, John M. Ferron
University of South Florida

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.

Keywords: MONTE CARLO, MULTILEVEL MODELS, SAMPLE SIZE, SAS/IML, SAS/STAT, PROC MIXED

INTRODUCTION

Multilevel models are increasingly employed across a variety of disciplines to analyze nested or hierarchically-structured data. 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, for example, Hox, 1998; Maas & Hox, 2004; Maas & Hox, 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. However, Hox (1998) conducted the study using continuous predictors and did not examine the effect of including cross-level interactions with binary variables with small sample sizes.

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, Ferron, Hess, Hogarty, Kromrey, Lang, Niles, & Lee, 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 example, for 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). Specifically, studies have noted positive parameter bias in the intercept, slope, and residual variances (Bell et al., 2009; Clarke & Wheaton, 2007; Mok, 1995) 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.

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 (i.e., 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 more encouraging, the majority of studies have only examined relatively simple models. For example, Clarke and Wheaton (2007), Mass 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 continuous predictor variable at each level, one cross-level interaction between the predictors, and two random effects (intercept and level-1 predictor). Because findings from simulation studies are not generalizable beyond the models and conditions examined in the studies, findings based on relatively simple models and design factors are often not helpful to the applied researcher. Instead, simulation studies need to investigate how models with varying numbers and type of predictors function under a variety of data conditions.

This study focused on the consequences of small level-1 and level-2 sample sizes on the estimation of fixed effect inferences in linear 2-level 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 2 and 3 level-1 predictors crossed with 2 and 3 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 9 sample size conditions and 1152 design factor conditions.

Data were generated based on a two-level model in which observations were nested within groups. At the first level, a continuous outcome was modeled as a linear function of k predictors, where k = 2 or 3.

$$y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} X_{kij} + r_{ij} .$$

At the second level, the intercepts and slopes of the first level were modeled as a function of m predictors where m = 2 or 3. 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, a level-1 interaction model, a level-2 interaction model, and a cross-level interaction model, yielding a total of 16 different models. See Figure 1 for example PROC MIXED code used for each of the four models. When estimating the models, the intercept and level-1 variables were allowed to vary randomly. 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., 2009). The level-2 errors were also generated from a normal distribution, but with variance of .10 or .30 for u_0 (intercept) and a variance of .0 or .10 for u_1 (slopes). The data were simulated such that one predictor at each level had no effect (γ = 0; for estimation of Type I error rate) and all remaining predictors had non-null effects (γ ≠ 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 equal to 30, slope and intercept variance equal to .1, and all correlations between predictors equal to .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.45,.42,.39, or.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., 2009) such that values below the mean were recoded as “0” and values above the mean were recoded as “1”.

```
title 'Main Effect Model K1=2, K2=2';
proc mixed data=b1;
class IDlevel2;
model y = x1 x2 w1 w2 /s cl alpha=.05 ddfm=kenwardroger;
random int x1 x2 / sub=IDlevel2 type=vc;
run;

title 'Cross-level Interaction Model K1=2, K2=2';
proc mixed data=b1;
class IDlevel2;
model y1 = x1 x2 w1 w2 x1*w1/s cl alpha=.05 ddfm=kenwardroger;
random int x1 x2 / sub=IDlevel2 type=vc;
run;

title 'Level-1 Interaction Model K1=2, K2=2';
proc mixed data=b1;
class IDlevel2;
model y2 = x1 x2 w1 w2 x1*x2/s cl alpha=.05 ddfm=kenwardroger;
random int x1 x2 / sub=IDlevel2 type=vc;
run;

title 'Level-2 Interaction Model K1=2, K2=2';
proc mixed data=b1;
class IDlevel2;
model y2 = x1 x2 w1 w2 w1*w2/s cl alpha=.05 ddfm=kenwardroger;
random int x1 x2 / sub=IDlevel2 type=vc;
run;
```

Figure 1. Example PROC MIXED Code for Main Effect and Interaction Models with Two Level-1 Predictors and Two Level-2 Predictors

For each of the 10,368 conditions (9 sample size combinations * 1152 combinations of design factors), 1,000 data sets were simulated using SAS IML (SAS Institute Inc., 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 2-level multilevel model with restricted maximum likelihood estimation and *Kenward-Roger* degrees of freedom estimation via the MIXED procedure in SAS (SAS, 2003b). 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 was modeled as $\Sigma = \sigma^2 \mathbf{I}$. Five primary outcomes were examined in this Monte Carlo study: (a) rate of model convergence, (b) rate of non-positive definite G-matrices, (c) bias in the estimates of the fixed effects, (d) average confidence interval coverage for each fixed effect, and (e) average confidence interval width for each fixed effect. In addition, the Type I error rates were estimated for null effects and statistical power was estimated for non-null effects.

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 being heavily influenced by a few extreme values (binary predictors: *M* = -.0064, min = -2.23, max = 14.29; continuous predictors: *M* = .0022, min = -3.04, max = 0.87). 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 |0.0006| with minimum of -0.013 and maximum of 0.014. Similarly, observed bias of continuous variables was not problematic across models as each model’s mean bias was less than |0.0003| with minimum of -0.0007 and maximum of 0.006.

Convergence was also not viewed as problematic as 100% convergence was obtained under 96.75% of the conditions studied. Of the conditions that did not have perfect convergence rates, 68.97% had a convergence rate of

3.8%. Further investigation revealed that these problematic conditions occurred in models with three level-1 and three level-2 predictors. Within these models, lack of convergence appeared most often among conditions with cross-level collinearity of .1, level-2 collinearity of .0 or .1, and all levels of level-1 collinearity (.0, .1. or .3). When there was no cross-level correlation among predictor variables, only models with level-1 collinearity of .3 and level-2 collinearity of .1 showed convergence problems.

Rates of non-positive definite G-matrices are presented by level-1 sample size by slope variance by level-2 sample size in Figure 2. Surprisingly, 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 with very little variability. *Ceteris paribus*, when level-1 sample size was reduced to the 5-10 sample size classification, the mean rate of non-positive definite G-matrices increased to .242. Next, holding slope variance at 0.3, as one would expect, the mean rate of non-positive definite G-matrices for the level-1 and 2 sample size category of 5-10 and 10, respectively, increased substantially to .652.Thus, even when the data were generated and modeled to vary randomly, models with the smallest sample sizes at each level estimated a positive definite G-matrix less than 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, in largest sample size conditions, the rate of non-positive G-matrices still only occasionally reached 100% (*M* = .86). Thus, even with 20-40 level-1 units and 30 level-2 units, on average 14.02% of the time, positive G-matrices were estimated for null effects. Interestingly, the average proportion of positive definite G-matrices for models with null variance components did not vary much across sample size combinations. Moreover, positive definite G-matrices occurred most frequently under the smallest sample size condition of 5-10 level-1 units and 10 level-2 units (*M* = .93).

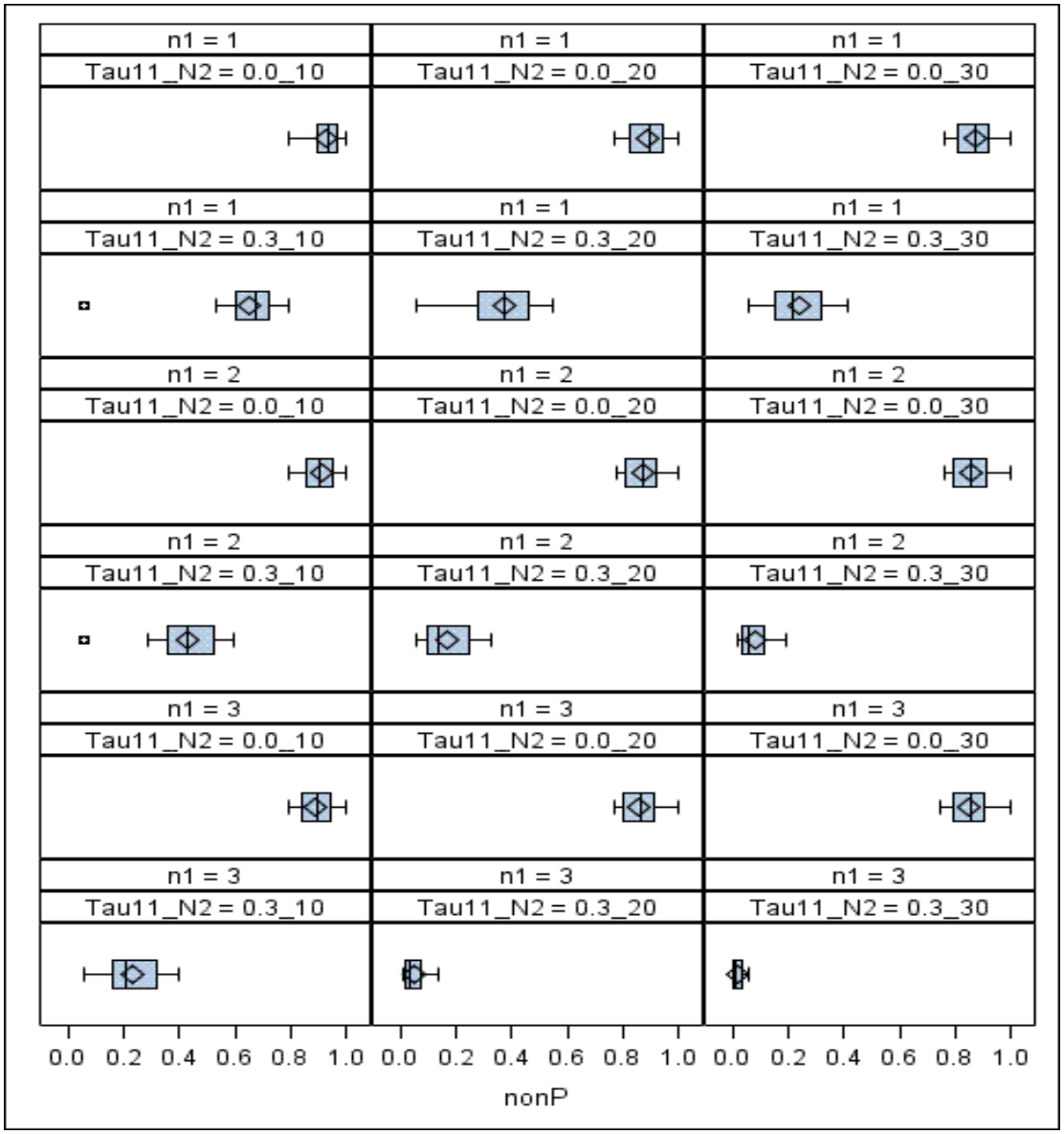

Figure 2. Rates of Non-Positive G-Matrices by Slope Variance (Tau11) by Level-1 (N1) and Level-2 (N2) sample sizes. [NOTE: “n1 = 1”, “n1 = 2”, and “n1 = 3” reflect level-1 sample size ranges of 5-10, 10-20, and 20-40, respectively.]

Overall, the estimated Type I error rates for tests of the fixed effect regression parameters of two continuous predictors maintained rates close to the nominal alpha level (*M* = .046, min = .021, max = .066). Thus, although there was some variability in the Type I error rates, overall, they did not appear to be overly influenced by any of the design factors included in the study.

After the data were winsorized as specified above, estimates of the 95% confidence interval widths across models were also not considered problematic. The minimum and maximum width for binary variables across models was 0.597 and 2.175, 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* = 1.17), and the main effect (*M* = 1.16) models. Among continuous predictors, the largest width was observed in the level-2 interaction model (*M* = 1.20) followed by cross-level interaction (*M* = 0.75), main effects (*M* = 0.74), and level-1 interaction (*M* = 0.71) models. The minimum and maximum widths for continuous variables across models was 0.359 and 2.285, respectively.

The distributions of estimated 95% confidence interval coverage for level-1 and level-2 fixed effect parameters are presented in Figures 3 and 4, respectively. The mean coverage estimates of binary and continuous level-1 fixed effects across models were .955 and .952, respectively. The mean coverage estimates of binary and continuous level-2 fixed effects across models were .950 and .951, respectively.

For the level-1 fixed effects, all models provided near nominal level coverage for the majority of conditions examined; the 95% confidence interval coverage estimates for continuous predictors was slightly less variable than for that 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 was .956, .956, .956, and.954, respectively. The mean coverage estimates for continuous level-1 predictors in main effect, level-1 interaction, level-2 interaction, and cross-level interaction models was .954, .948, .954, and .954, respectively. Coverage estimate distributions for level-1 predictors are presented in Figure 3 below.

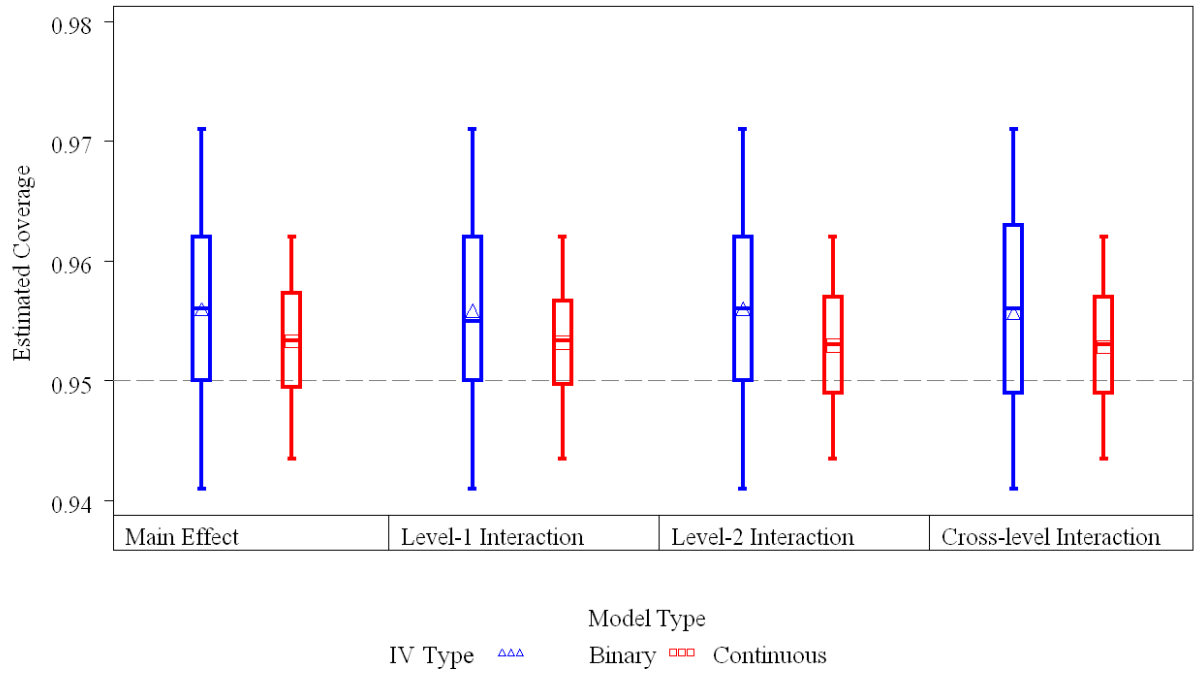

Figure 3. 95% Confidence Interval Coverage for Level-1 Binary and Continuous Fixed Effects by Model Type

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 for binary level-2 predictors in main effect, level-1 interaction, level-2 interaction, and cross-level interaction models was .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 was .951, .951, .951, and .950, 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 4 below.

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 .955, minimum of .921, and maximum .984.

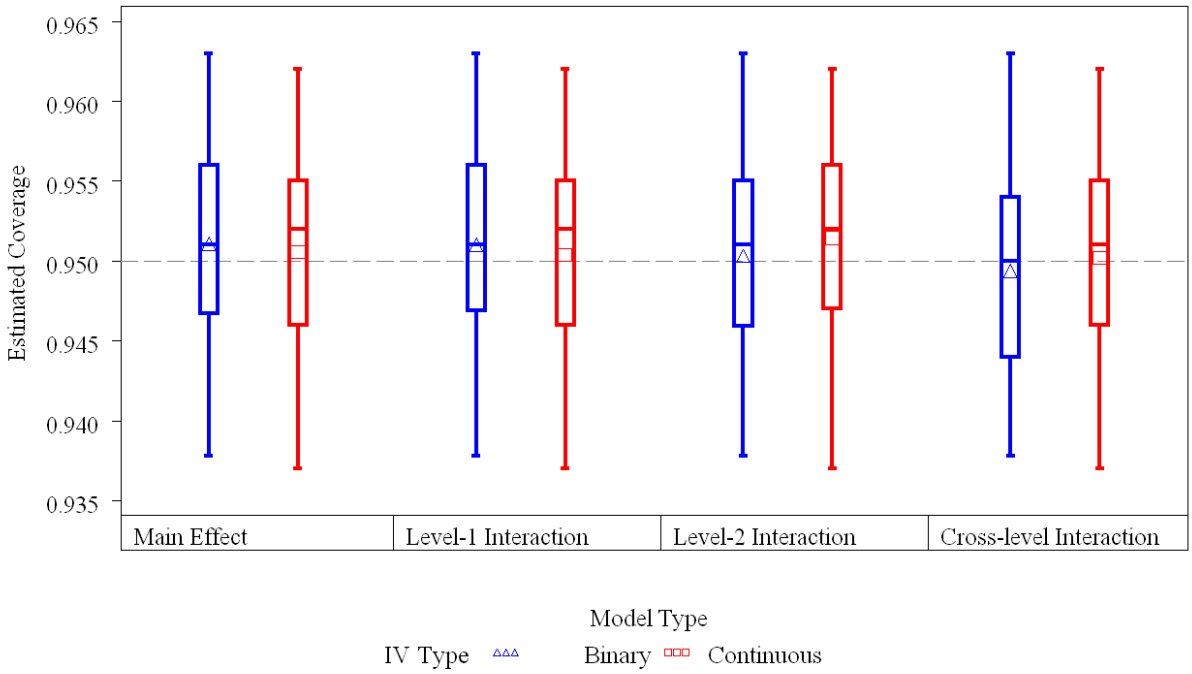

Figure 4. 95% Confidence Interval Coverage for Level-2 Binary and Continuous Fixed Effects by Model Type

On the whole, the mean estimates of statistical power for binary (*M* = .389, min = .065, max = .939) and continuous predictors (*M* = .556, min = .113, max = .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 one and two (see Figures 5 and 6 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 to 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 10-20 and 30, respectively, did estimated power reach a mean at or above .80.

As shown in Figure 6, 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 .31 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* = .318 and *M* = .626, 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 predictors were less than favorable. As sample sizes at each level increased, observed power likewise increased. However, regardless of the level-1, level-2 sample size combination, statistical power estimates of cross-level predictors failed to reach the .80 threshold. In all sample size classifications save the condition with 20-40 level-1 and 30 level-2 units, power estimates had an upper bound of .569. The remaining condition had a lower bound of .105 and upper bound at .775. Figure 7 contains the distribution of statistical power estimates by sample size combinations for binary cross-level interaction fixed effects.

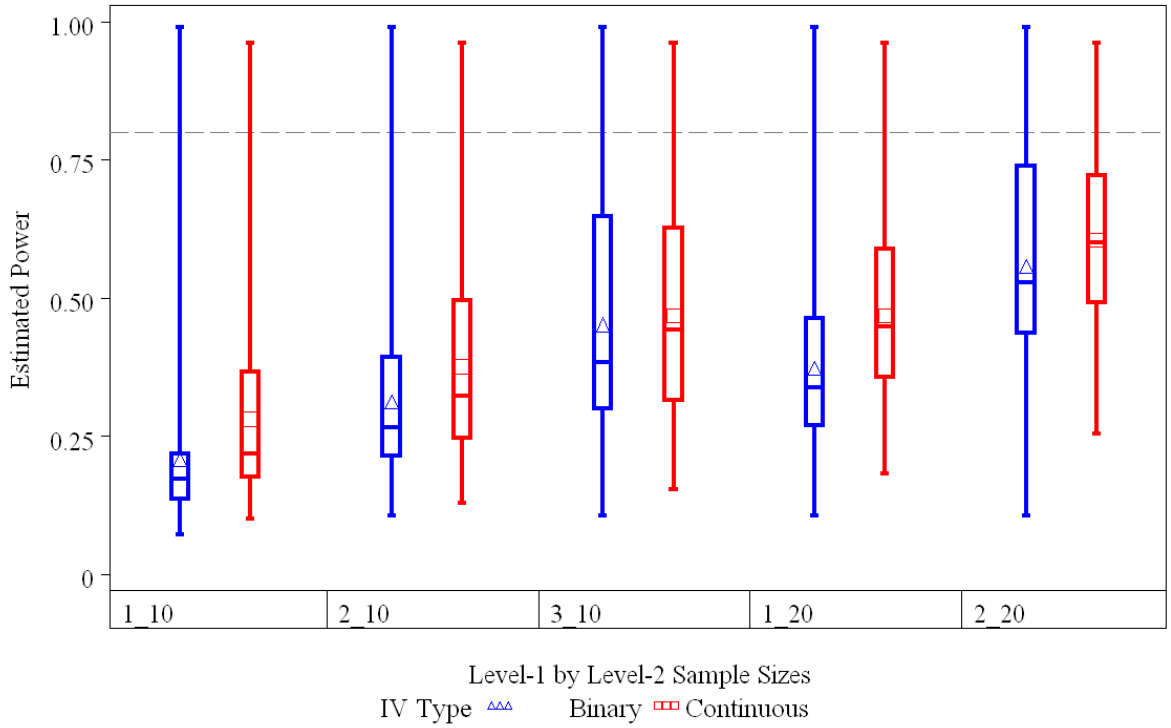

Figure 5. Power Distributions of Binary and Continuous Level-1 Fixed Effects by Sample Size. **[**NOTE: "n1 = 1", "n1 = 2", and "n1 = 3" reflect level-1 sample size ranges of 5-10, 10-20, and 20-40, respectively.]

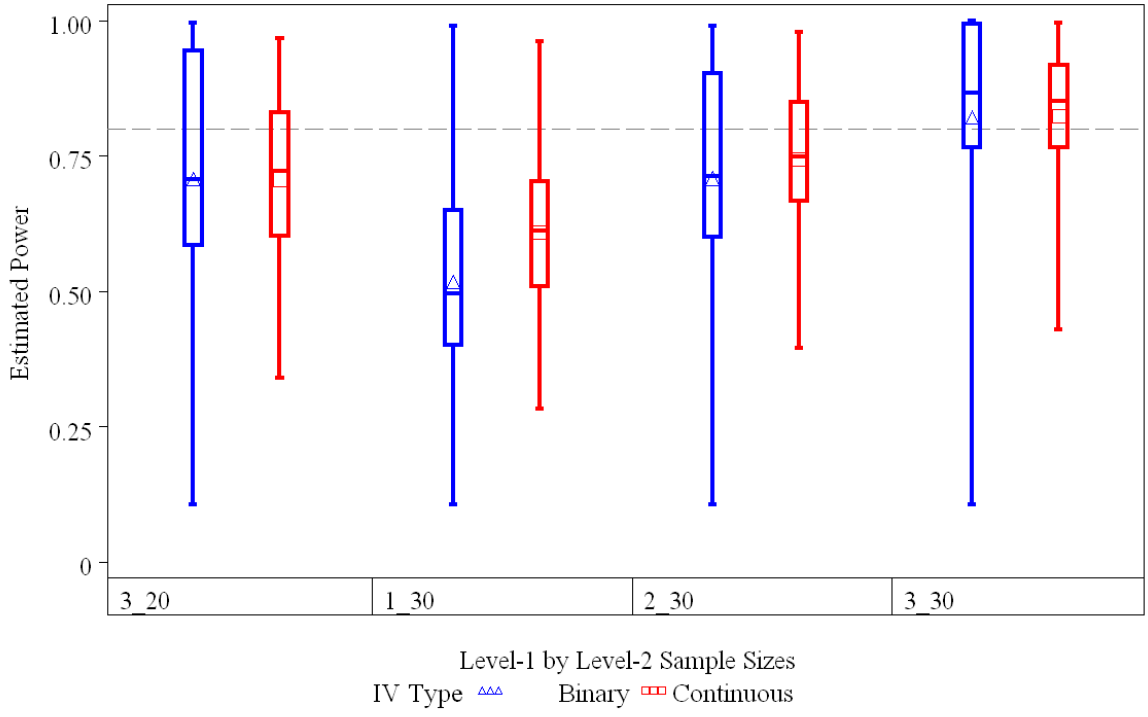

Figure 5 (cont). . Power Distributions of Binary and Continuous Level-1 Fixed Effects by Sample Size. **[**NOTE: "n1 = 1", "n1 = 2", and "n1 = 3" reflect level-1 sample size ranges of 5-10, 10-20, and 20-40, respectively.]

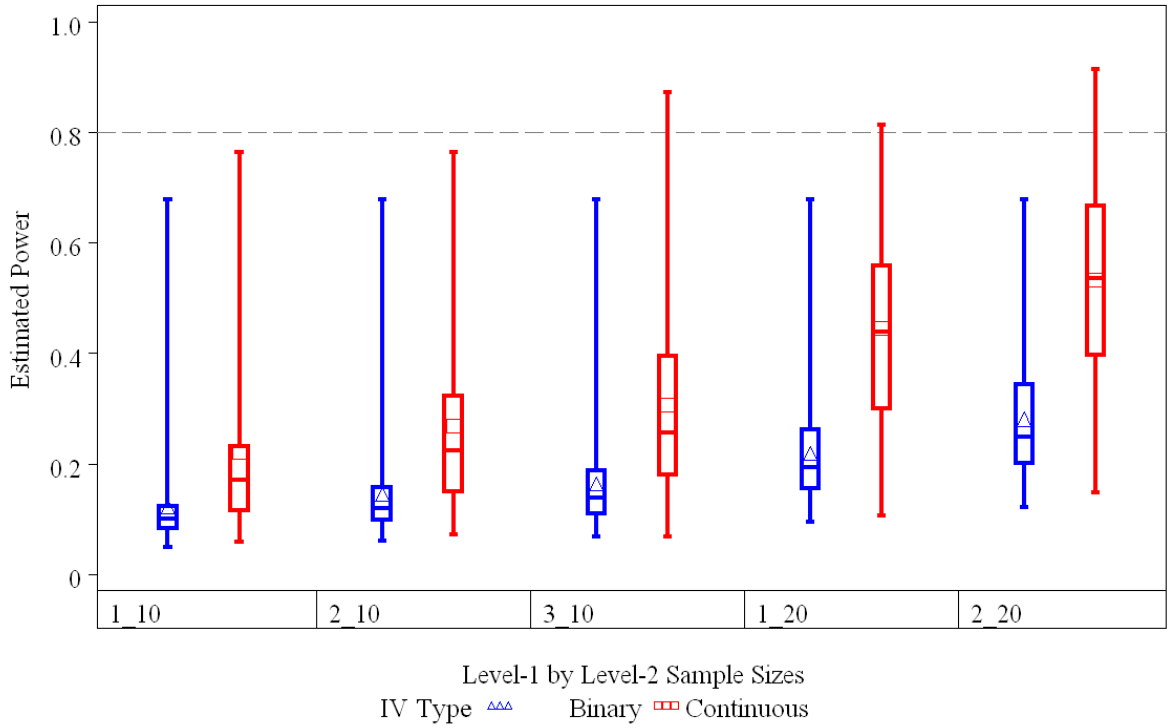

Figure 6. . Power Distributions of Binary and Continuous Level-2 Fixed Effects by Sample Size. **[**NOTE: "n1 = 1", "n1 = 2", and "n1 = 3" reflect level-1 sample size ranges of 5-10, 10-20, and 20-40, respectively.]

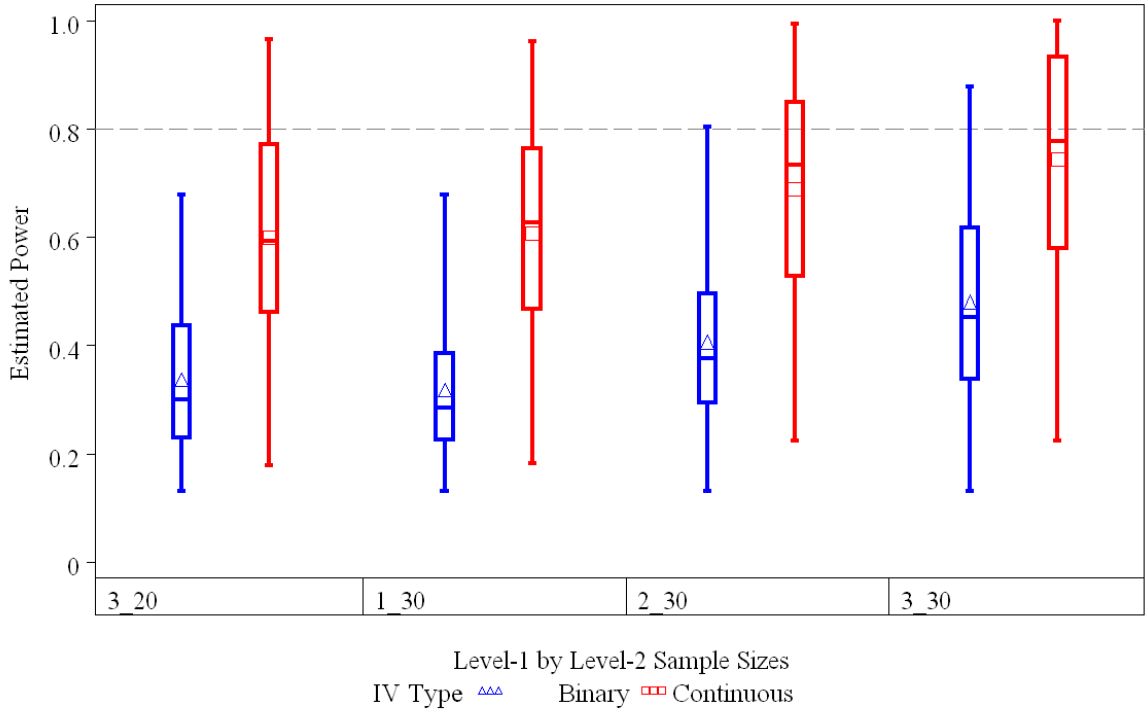

Figure 6 (cont). Power Distributions of Binary and Continuous Level-2 Fixed Effects by Sample Size. **[**NOTE: "n1 = 1", "n1 = 2", and "n1 = 3" reflect level-1 sample size ranges of 5-10, 10-20, and 20-40, respectively.]

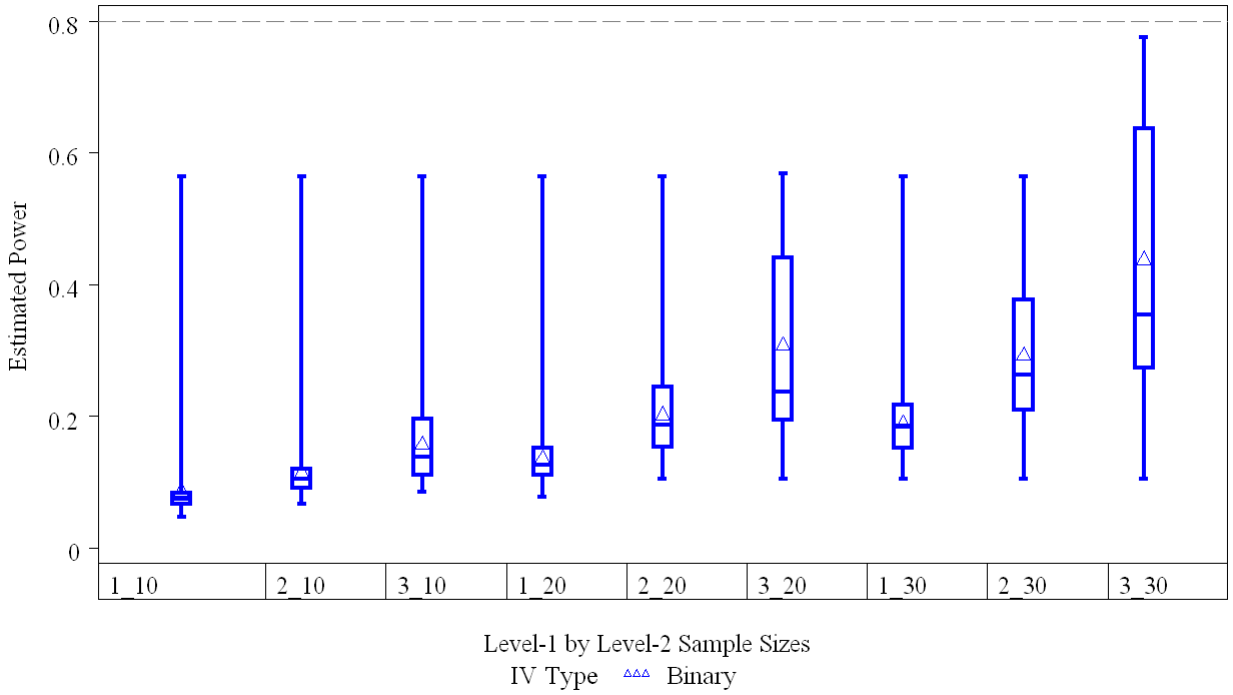

Figure 7). Power Distributions of Cross-Level Interaction Fixed Effects by Sample Size. **[**NOTE: "n1 = 1", "n1 = 2", and "n1 = 3" reflect level-1 sample size ranges of 5-10, 10-20, and 20-40, respectively.]

CONCLUSIONS

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. Such models are often not helpful to the applied researcher. Thus, in an effort to help applied researchers make sound decisions regarding the use of two-level linear models with small sample sizes, the current study sought to build on the existing body of literature with specific regard to level-1 and level-2 sample sizes, levels of collinearity among predictors, and model complexity, including the number and type of predictors, as well as the type of model estimated (i.e., main effect, level-1 interaction, level-2 interaction, and cross-level interaction).

Two aspects of model complexity, the number of predictors at each level and the type of model, and 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 non-positive 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, 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 increased across models and conditions. Thus, across the many design factors included in the current study, these findings suggest that 95% confidence interval coverage, Type I error, and statistical bias 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. Often, the observed power never reached the typically desired level of .80 in conditions where sample sizes at levels one and two were limited. Only when level-1 and level-2 sample sizes were 20-40 and 30, respectively, did estimated power of level-1 predictors reach a mean at or above .80. In the smaller sample size combinations, although the distributions of power estimates on occasion reached 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 very large effect sizes are present. Moreover, estimated statistical power of level-2 predictors never reached a mean of .80 across any sample size combinations. 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 ($\tau_{11} = .3$), non-positive 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 suggest substantial bias in the estimates of random effects with small sample sizes (Bell, Ferron, & Kromrey, 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.

REFERENCES

Bell, B.A., Ferron, J.M., & Kromrey, J.D. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*.

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.

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.

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.

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

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: Springer Verlag.

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

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86-92.

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

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

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, Second Edition*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bethany A. Bell
University of South Carolina
College of Education
Wardlaw 133
Columbia, SC 29208
Work Phone: 803-777-2387
E-mail:babell@sc.edu