The Impact of Small Cluster Size on Multilevel Models: A Monte Carlo Examination of Two-Level Models with Binary and Continuous Predictors

Bethany A. Bell[1], Grant B. Morgan[1], Jeffrey D. Kromrey[2], John M. Ferron[2]
[1]Educational Psychology, Research, and Foundations, University of South Carolina, College of Education, Wardlaw 133, Columbia, SC 29208
[2]Educational Measurement and Research, University of South Florida, College of Education, 4202 East Fowler Avenue, EDU 162, Tampa, FL 33620

Abstract
Recent methodological research has addressed the important issue of sample size at each level when estimating multilevel models. Although several design factors have been investigated in these studies, differences between continuous and binary predictor variables have not been scrutinized (previous findings are based on models with continuous predictor variables). To help address this gap in the literature, this Monte Carlo study focused on the consequences of level-2 sparseness on the estimation of fixed and random effects coefficients in terms of model convergence and both point and interval parameter estimates. The 5,760 conditions simulated in the Monte Carlo study varied in terms of level-1 sample size, number of level-2 units, proportion of singletons (level-2 units with one observation), type of predictor, collinearity, intraclass correlation, and model complexity.

Key Words: multilevel modeling, sample size, binary predictors, survey research

1. Sample Size and Multilevel Modeling

Multilevel models are being increasingly used across the social sciences 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 (also known as hierarchical linear models and mixed 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 the 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, 2002; Maas & Hox, 2004) 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.

Although many researchers attempt to adhere to these sample size guidelines, the nature of social determinants research often make these sample size recommendations difficult to achieve. More specifically, because many large-scale social science surveys utilize complex sampling procedures (i.e., stratified and clustered sampling designs), individuals are often dispersed among a large number of level-2 units with few individuals per group (e.g., thousands of census tract defined neighborhoods with few individuals in each census tract). For example, although the sampling frame for the National Longitudinal Study of Adolescent Health (Add Health) was at the school-level, neighborhood-level Add Health data are often used to answer a variety of research questions (e.g., Bruce, 2004; Cubbin, Santelli, Brindis, & Braveman, 2005; Gordon-Larsen, Nelson, Page, & Popkin, 2006; Knoester & Haynie, 2005; Regnerus, 2003; Wickrama & Bryant, 2003; Wickrama, Noh, & Bryant, 2005). However, the dispersion of adolescents across neighborhoods is less than ideal. The Wave 1 Add Health restricted use data provide observations on approximately 15,000 adolescents nested in approximately 2,600 neighborhoods, with almost 50% singleton neighborhoods (i.e., a neighborhood unit containing only one adolescent).

Given the potential problems of small sample sizes, several simulation studies have been designed to examine the effect of small sample sizes, at different levels of analysis, on various multilevel results (e.g., variance estimates, fixed effects estimates, standard errors, and convergence). In a simulation study examining the effects of data sparseness on variance estimates, Mok (1995) found that variance estimates were notably biased in balanced designs with as few as five level-2 units. Clarke and Wheaton (2007), in their Monte Carlo study focusing on a 2-level model, examined conditions in which the number of level-2 units ranged from 50 to 200 and the number of level-1 units per level-2 unit ranged from 2 to 20. They found positive bias in the intercept and slope variance estimates. They noted that "at least 10 observations per group for at least 100 groups" (p. 330) were needed for the estimated intercept variance to approach the true values; for the slope variance at least 20 observations per group for at least 200 groups were needed for the estimated slope variance to approach the true values. In this same study, Clarke and Wheaton (2007) also examined bias in the intercept and slope variance estimates as a function of group size and the proportion of singleton groups in the two level models. When singleton groups are included in the multilevel models, bias in the variance estimates was more evident than with the data without any singleton groups.

Maas and Hox (2004, 2005), who examined conditions in which the number of level-2 units ranged from 30 to 100 and the number of level-1 units per level-2 unit ranged from 5 to 50, found less bias in the variance estimates but still reported substantial difficulty in making inferences about the variance components when the number of level-2 units was only 30. Although there appear to be substantial problems in making variance inferences from small samples, results of the simulation studies regarding the fixed effects were more encouraging. Studies consistently showed little to no bias in the estimates of the fixed effects (Clarke & Wheaton, 2007; Maas & Hox, 2004, 2005; Mok, 1995; Newsom & Nishishiba, 2002). However, 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, both Clarke and Wheaton's (2007) and Mass and Hox's (2004) findings are 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 at each level, and two random effects (intercept and level-1 predictor). One exception is a previous study conducted by Bell, Ferron, and Kromrey (2008). Findings from that study, which investigated the impact of model complexity and the proportion of singletons in a model, revealed similar patterns as those conducted by others with simple models. However, in this previous study, we, like others, only included continuous predictors at each level. Thus, the impact of level-2 sparseness with more complex, realistic models with both continuous and binary predictors is currently not known.

2. Purpose

This study focuses on the consequences of level-2 sparseness on the estimation of fixed and random effects coefficients in terms of model convergence and both point estimates (statistical bias) and interval estimates (confidence interval accuracy and precision, and Type I error control of tests associated with the fixed and random effects) as a function of the level-1 sample size, number of level-2 units, proportion of singletons, type of predictor, collinearity, intraclass correlation, and model complexity. By examining more complex multilevel models (i.e., two level models with various numbers of predictors, various levels of collinearity, and both binary and continuous predictors), 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.

3. Method

For this Monte Carlo study the following design factors and conditions were examined: (a) level-1 sample sizes with conditions of small (average = 10, range 5-15) and large (average = 50, range 25-75), (b) level-2 sample sizes with conditions of 50, 100, 200, and 500, (c) proportion of singletons with conditions of 0, .10, .30, .50, and .70, (d) levels of collinearity (0 and .3), (e) level-2 error variance (.05, .10, .15, and .30), and (f) model complexity with conditions of 2, 3, and 5 level-1 predictors crossed with 1, 2, and 4 level-2 predictors for both main effect and interaction models. These factors in the Monte Carlo study were completely crossed, yielding 40 sample size conditions and 144 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 generated as a linear function of k predictors, where k = 2, 3, or 5.

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

The intercepts and slopes of the first level were simulated as a function of m predictors at the second level, where m = 1, 2, or 4. In each model, one level-1 predictor (x1) and one level-2 predictor (w1) were binary and all others were continuous. For the interaction models, models with two level-1 predictors included one cross-level interaction term, models with three level-1 predictors included two cross-level interaction terms, and models with five level-1 predictors included three cross-level interaction terms. One cross-level interaction term included both binary variables (x1*w1). In each model, the intercept and one level-1 variable were generated 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 9.1 (SAS, 2004). The level-2 errors were also generated from a normal distribution, but with variance of .05, .10, .15, or .30 to produce

different values of intraclass correlation. The data were simulated such that some predictors had no effect (for estimation of Type I error rate) and some predictors had non-null effects.

For each of the 5760 conditions (40 sample size combinations X 144 combinations of design factors), 1,000 data sets were simulated using SAS IML (SAS, 2004). 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 maximum likelihood estimation via the MIXED procedure in SAS (SAS, 2004). In all analyses the covariance matrix of the level-2 errors, $\mathbf{T}$, was modeled to be unstructured, and the covariance matrix of the level-1 errors was modeled as $\Sigma = \sigma^2\mathbf{I}$. Four primary outcomes were examined in this Monte Carlo study: rate of model convergence, bias in the estimates of the fixed and random effects, confidence interval coverage for each effect, and average confidence interval width for each effect. In addition, Type I error rates were reported.

4. Results

Model convergence was not a substantial problem with any of the conditions examined in this study. No convergence problems were evident in 98% of the conditions and the highest rate of nonconvergence in the remaining 2% of conditions was less than 2% of the simulated samples. Similarly, very low levels of statistical bias were evident for all parameter estimates. For fixed effects, min = -0.02 and max = 0.02; for random effects, min = -0.01 and max = 0.01.

Overall, binary predictors at level-1 and level-2 behaved similarly to continuous predictors, despite slightly larger CI widths. More specifically, regardless of type of predictor variable, estimated confidence interval coverage for the fixed effects (Figure 1) was relatively constant. Level-2 predictors had slightly more variability and slightly lower coverage than level-1 predictors, but, overall, predictors at both levels exhibited relatively consistent confidence interval coverage at or near the nominal .95. As shown in Figure 2, in terms of CI width, the binary level-2 predictor (w1) had the widest confidence interval of all predictors and interactions that included the binary level-2 predictor (w1) had wider confidence intervals than the main effect predictors.

Next, the proportion of singletons had no notable effect on the estimation of fixed effects for the level-1 predictors, but evidenced a clear impact on the interval estimation of the parameters for the level-2 predictors, especially when level-2 sample size was small. As shown in Figure 3, when there were only 50 level-2 units, the average confidence interval coverage quickly deteriorates as the proportion of singletons increases. However, when level-2 sample size was 500, confidence interval coverage of level-2 predictors remained relatively unaffected by changes in the proportion of singletons.

When confidence interval coverage was evaluated by computing the proportion of conditions with estimated interval coverage that fell with Bradley's (1978) "liberal" criterion for robustness (e.g., with a 95% confidence interval, the estimated coverage is greater than 92.5% and less than 97.5%), the same pattern was evident. Figure 4 presents the proportion of conditions that satisfied Bradley's criterion for small and large numbers of level-2 units. With 500 level-2 units, the proportion of singletons had no impact on the confidence interval coverage and nearly all conditions met Bradley's criterion. In

contrast, with only 50 level-2 units, the proportion of conditions meeting Bradley's criterion declined as the proportion of singletons increased. With 70% singletons, the proportion of conditions with adequate confidence interval coverage was typically less than .10.

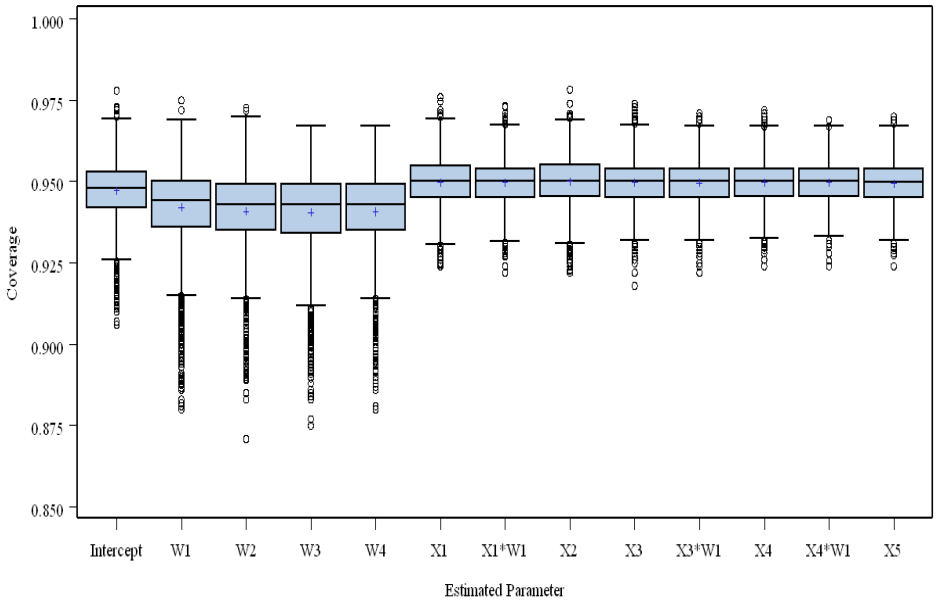

Figure 1: Estimated 95% confidence interval coverage for fixed effects

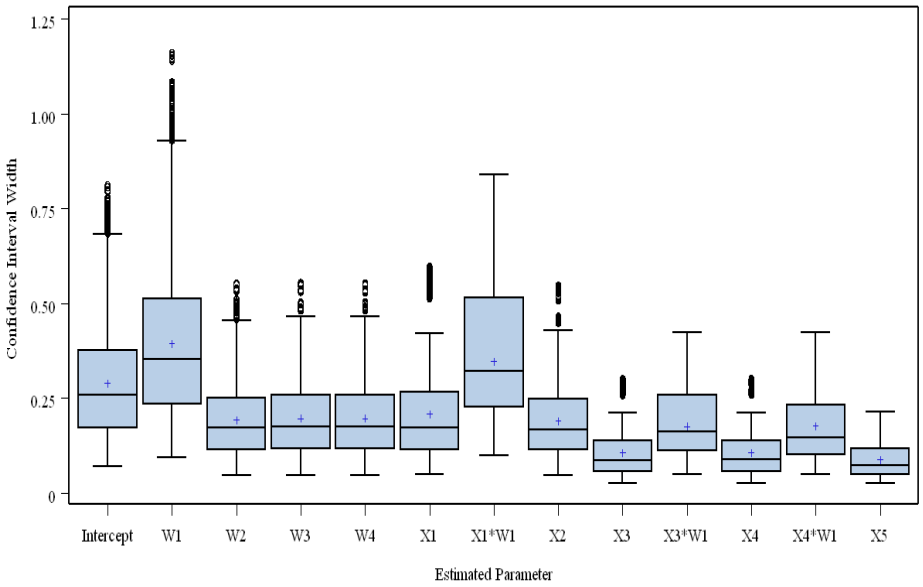

Figure 2: Estimated 95% confidence interval widths for fixed effects

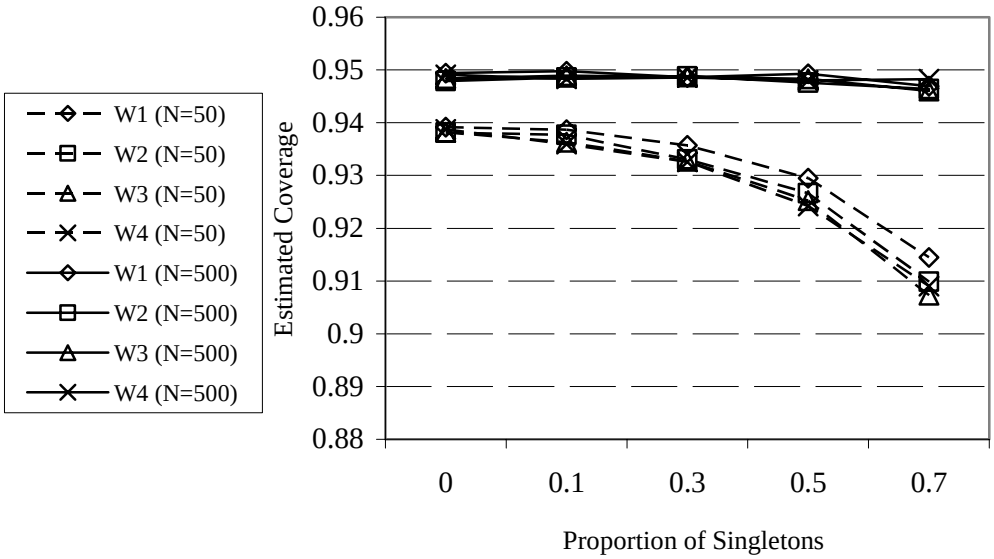

Figure 3: Average coverage of level-2 predictors by level-2 sample size and proportion of singletons

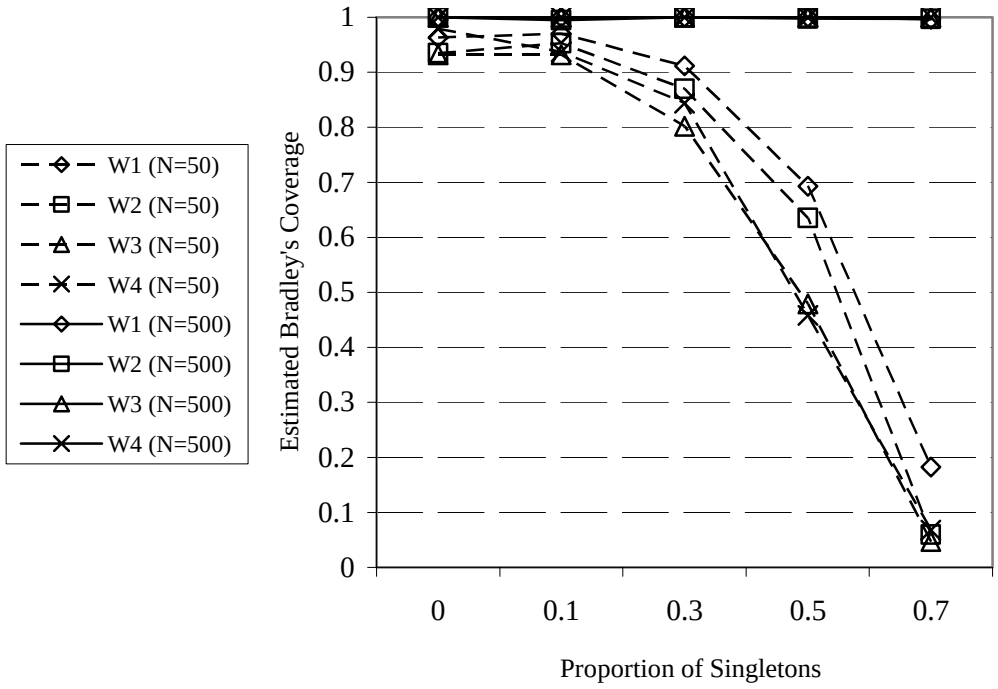

Figure 4: Bradley's coverage of level-2 predictors by level-2 sample size and proportion of singletons

Figures 5, 6, and 7 present the estimated Type I error rates for the tests of fixed effects for the binary level-2 predictor, as a function of level-2 sample size, proportion of singletons, and level-2 sample size and proportion of singletons, respectively. With smaller level-2 sample sizes, the estimated Type I error rate for the binary level-2 predictor was slightly liberal (Figure 5). Likewise, as the proportion of singletons increased, the estimated Type I error rate for the binary level-2 predictor was also liberal (Figure 6). Figure 7 displays the estimated Type I error rate of the binary level-2 predictor as a function of both level-2 sample size and proportion of singletons. When there were only 50 level-2 units, the estimated Type I error rate was slightly above the nominal .05, even when there were no singletons in the model and continued to become more liberal as the proportion of singletons increased.

An opposite pattern was observed for the estimated Type I error rates for the tests of random effects (Figure 8). With large numbers of level-2 units (N_2 = 500), the proportion of singletons had relatively no effect on Type I error control. With few level-2 units (N_2 = 50), the test became slightly conservative as the proportion of singletons increased. However, across all conditions the estimated Type I error rates were close to the nominal alpha level (.043).

5. Conclusions

The results of this study are encouraging for researchers who analyze multilevel data with sparse structures. As with previous investigations of small cluster sizes, regardless of model complexity, the proportion of singletons in the simulated samples had little impact on either the point or interval estimates of model parameters when large numbers of level-2 units were included. With smaller numbers of level-2 units, increasing the proportion of singletons led to a reduction in the accuracy of the confidence intervals for level-2 predictors and bias in the Type I error control of the binary level-2 predictor, but did not impact the accuracy of the estimates for level-1 predictors. With that said, it is imperative for researchers to remember that as with all simulation research, these findings are only generalizable to data and model conditions included in our study. Additional research is required to examine a broader array of data structures and to evaluate the impact of sparse structures for analyzing non-normal outcomes (e.g., dichotomous or count variables) using generalized multilevel models.

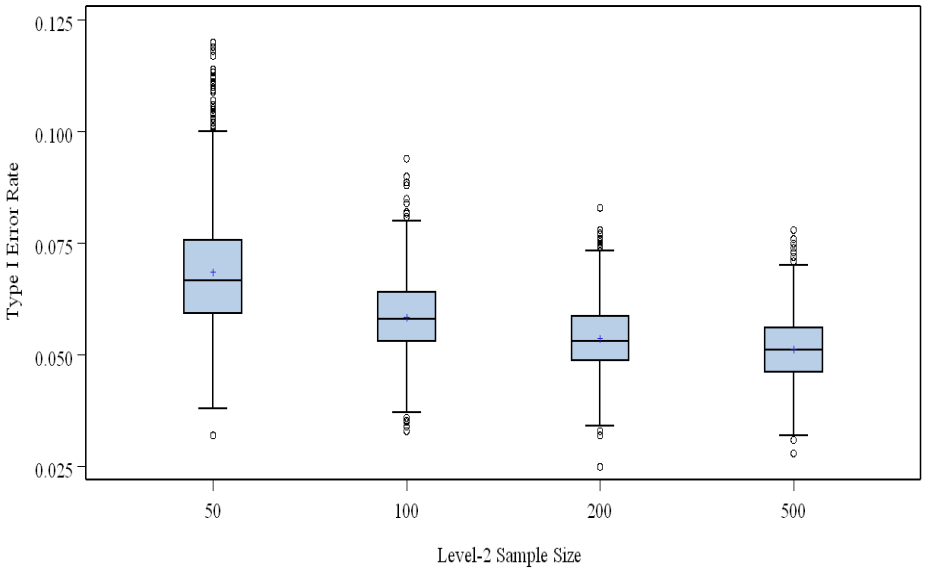

Figure 5: Distributions of Type I error rates of binary level-2 predictor (W1) by level-2 sample size

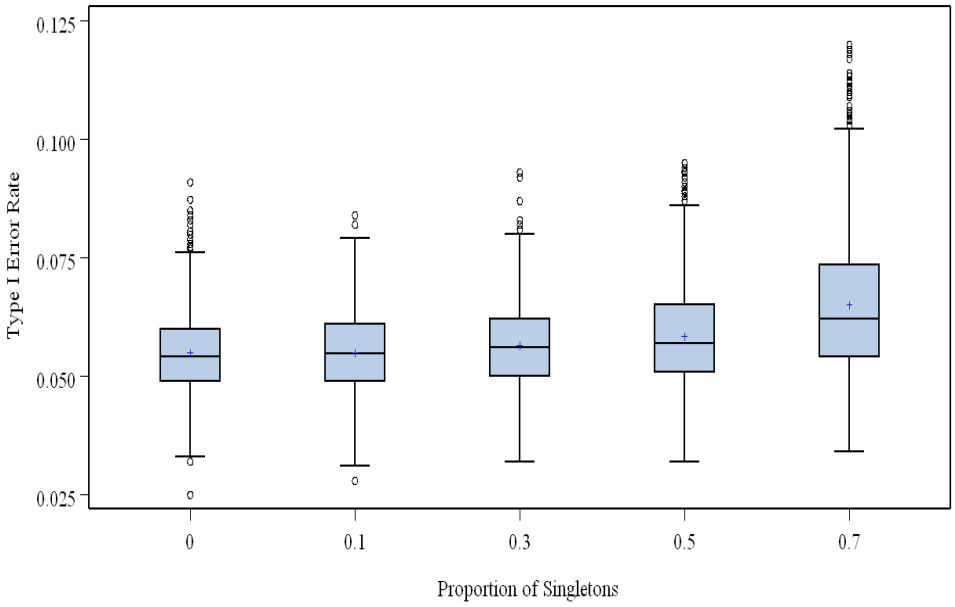

Figure 6: Distributions of Type I error rates of binary level-2 predictor (W1) by proportion of singletons

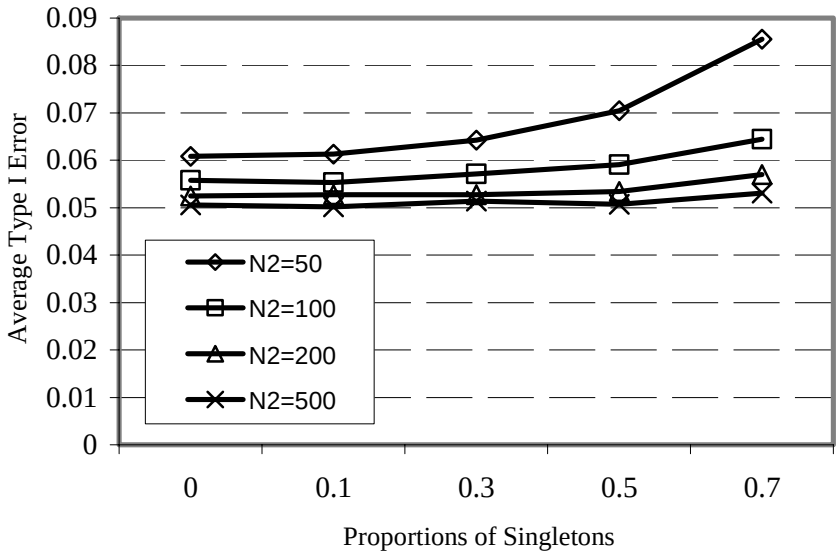

Figure 7: Average Type I error rate of binary level-2 predictor (W1) by level-2 sample size and proportion of singletons

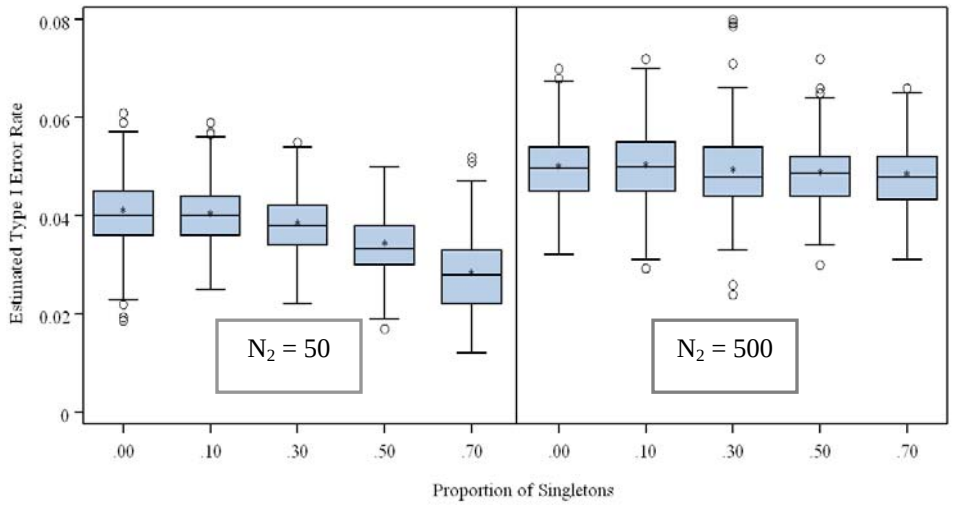

Figure 8: Distributions of Type I error rates for random effects by level-2 sample size and proportion of singletons

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. *JSM Proceedings,* Section on Survey Research Methods. Alexandria, VA: American Statistical Association. 1122 – 1129. http://www.amstat.org/Sections/Srms/Proceedings/y2008/Files/300933.pdf

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

Bruce, M. A. (2004). Inequality and adolescent violence: An exploration of community, family, and individual factors. *Journal of the National Medical Association, 96*(4), 486-495.

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.

Cubbin, C., Santelli, J., Brindis, C. D., & Braveman, P. (2005). Neighborhood context and sexual behaviors among adolescents: Findings from the national longitudinal study of adolescent health. *Perspectives on Sexual and Reproductive Health, 37*(3), 125-134.

Gordon-Larsen, P., Nelson, M. C., Page, P., & Popkin, B. M. (2006). Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics, 117,* 417–424.

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.

Knoester, C., & Haynie, D. L. (2005). Community context, social integration into family, and youth violence. *Journal of Marriage and the Family, 67*(3), 767-780.

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.

National Longitudinal Study of Adolescent Health. (2005). *National Longitudinal Study of Adolescent Health.* Retrieved October 30, 2006, from http://www.cpc.unc.edu/addhealth

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

Regnerus, M. D. (2003). Moral communities and adolescent delinquency: Religious contexts and community social control. *Sociological Quarterly, 44*(4), 523-554.

SAS Institute Inc. (2004). *SAS, release 9.12* [computer program]. Carey, NC: SAS Institute Inc.

Wickrama, K. A. S., & Bryant, C. M. (2003). Community context of social resources and adolescent mental health. *Journal of Marriage and the Family, 65*(4), 850-866.

Wickrama, K. A. S., Noh, S., & Bryant, C. M. (2005). Racial differences in adolescent distress: Differential effects of the family and community for blacks and whites. *Journal of Community Psychology, 33*(3), 261-282.