A Comparison of the Performance of Univariate and Multivariate Multilevel Models for a Cluster Randomized Two-Group Design

Wanchen Chang Boise State University
Keenan A. Pituch University of Texas at Austin
S. Natasha Beretvas University of Texas at Austin

The multivariate multilevel model (MVMM) is an extension of the univariate multilevel model (MLM) that may be used in the presence of multiple outcomes. In a two-group cluster randomized design, one approach is to apply the MLM separately to each outcome variable whereas an alternate approach is to use the MVMM, which incorporates all outcomes simultaneously in a single analysis model. This Monte Carlo study investigated the degree to which results from the two models differ across a set of conditions that can be considered to favor the use of univariate analysis. Our results showed there were no differences in the performance of the MLM and MVMM with respect to estimation bias, power, and Type I error rate. We discuss the implications of these findings for applied researchers.

In educational research, clustered data are often collected from individuals who are nested within clusters and then analyzed with hierarchical linear or multilevel models (MLMs; Raudenbush & Bryk, 2002). As is well known, with such clustered data, MLMs help maintain accurate Type I error rates for tests of fixed effects while use of traditional multiple linear regression (MR) models may result in underestimated standard errors. In addition, in applied studies, data collected from participants often include scores on multiple outcome variables (Pituch, Whittaker, & Chang, 2016). Since conventional MLMs include just one participant-level outcome, and are thus univariate models, one approach, when multiple outcomes are collected, is to conduct as many MLM analyses as there are outcomes.

However, in recent years, there has been a growing interest in using a multivariate analysis approach when multiple outcomes are collected in a multilevel design. This approach, known as the multivariate multilevel model (MVMM), is an extension of the standard MLM that allows for multiple outcomes to be analyzed jointly (Goldstein, 2011; Hox, 2010; Snijders & Bosker, 2012). It is also a multilevel extension of single-level multivariate multiple linear regression (MMR) models if multiple outcomes are jointly regressed on a set of independent variables that are common to all the outcomes, and Zellner's (1962) seemingly unrelated regression (SUR) models if the independent variables differ across the outcomes (Timm, 2002).

In applied studies with continuous clustered data, MVMMs have been used to investigate the effects of (a) emotion regulation on multiple indicators of concordance (Butler, Gross, & Barnard, 2014); (b) cognitive ability (Freund, Hooling, & Preckel, 2007) and reading ability (Korpershoek, Kuyper, & Van Der Werf, 2015) on measures of academic achievement; (c) intergroup contact on subtle and blatant prejudice in adolescents (Olaizola, Diaz, & Ochoa, 2014); and (d) individual and contextual variables on health and happiness (Pierewan & Tampubolon, 2015). In addition, several pedagogical articles have analyzed real and simulated multilevel data for the purpose of illustrating the application of MVMMs in cross-sectional (Hauck & Street, 2006; Hoffman & Rovine, 2007; Paterson, 1998; Tate & Pituch, 2007) and longitudinal (Baldwin et al., 2014; Plewis, 2005) contexts, as well as in situations with missing data (Yang, Goldstein, Browne, & Woodhouse, 2002). Other studies have proposed methods of specifying Bayesian priors (Turner, Omar, & Thompson, 2006) and incorporating sample weights (Veiga, Smith, & Brown, 2014) when estimating MVMMs.

There are several advantages to using the multivariate approach. According to Snijders and Bosker (2012) and Hox (2010), the MVMM tends to be more powerful than the MLM for tests of specific dependent variables, especially if the outcomes are highly correlated and there is a large amount of incomplete outcome data. In addition, with the MVMM, one can test whether the impact of a predictor variable differs across multiple outcomes, provided these outcomes are on the same scale (Baldwin, Imel, Braithwaite, & Atkins, 2014; Pituch & Stevens, 2016; Snijders & Bosker, 2012). In addition, with multilevel data, the MVMM allows one to obtain estimates of correlations among the outcomes at each of the various levels (e.g., within cluster, between cluster). These correlations may be of substantive interest and would not be estimable with a strictly univariate approach. Finally, the MVMM procedure can be implemented using an overall multivariate test which is useful in controlling the Type I error rate. Note

though that Frane (2015) showed, for a standard (non-multilevel) design, use of a Bonferroni-adjusted alpha approach for the tests of specific dependent variables also provides for accurate family-wise and per-family Type I error rates and can often provide for as much or greater power compared to using a multivariate omnibus test, particularly when the number of outcomes are greater than two.

While applied researchers have begun to use the MVMM, there has been limited research regarding how well the MVMM performs compared to separate univariate only analyses. Park, Pituch, Kim, Chung, and Dodd (2015) compared MVMM to traditional multivariate and univariate analyses with regard to power and Type I error rate in a non-multilevel setting across several conditions. Their simulation study found that the MVMM was often more powerful, particularly as outcome missingness and the correlation among outcomes increased, even when data were missing completely at random. Further, Hauck and Street (2006) and Baldwin et al. (2014) compared results from MVMM analysis with multilevel data to those from MLM analyses. Hauck and Street analyzed existing cross-sectional data on the performance of health organizations, and Baldwin et al. simulated longitudinal data in the context of a clinical study on depression with a single fixed condition. Both studies found little to no difference in MVMM and MLM estimates of the common parameters (i.e., fixed effects and variances). According to Baldwin et al. (2014), MLM and MVMM provided essentially the same estimates in their study because the fixed effects and variances ("univariate" parameters) are estimated using the data for their respective outcome, independent of the data for all other outcomes. Note, though, that no missing data were present in their study. However, because the MVMM allows outcomes to be correlated, the MVMM models were a better fit to the data than the corresponding MLMs. According to Hauck and Street, the correlations obtained from MVMM analysis provides more information, which "improves the quality of the statistical analysis" and "provides insight into the potential trade-offs or synergies between" outcomes (p. 1047). Furthermore, allowing outcomes to be correlated leads to more powerful tests of whether the effect of a treatment varies across outcomes (Baldwin et al., 2014).

Given the similar performance of the MVMM and MLM in Hauck and Street (2006) and Baldwin et al. (2014), this study explores the degree to which these statistical models perform similarly under conditions generally favorable to the univariate model. That is, we assume multiple correlated outcomes are present, but that there is no missing data. Further, we suppose that researchers have no interest in testing whether the effect of a given predictor varies across outcomes and are not particularly interested in estimating the within- and between-cluster correlations among the outcomes.

We are not aware of any study that has compared the performance of MVMM and MLM in this context. As such, we conduct a simulation study to compare the performance of these models with regard to estimation bias, power, and Type I error rate accuracy. Similar performance between these approaches would suggest, of course, that researchers could potentially forgo using the more complex multivariate approach.

We also note that we do not examine the possible use of descriptive discriminant analysis in this context. As is well known, this procedure may be used to find linear composites of the outcome variable that best differentiate between groups. However, our study context assumes that researchers are interested in each outcome in its own right, a condition for focusing on separate outcomes as noted in Pituch and Stevens (2016, p. 408). That is, our context assumes that researchers wish to test whether there are group mean differences present for each outcome as well as estimating the association between an individual-level predictor and each outcome. A primary reason that researchers focus on outcome variables separately in such a multivariate setting is that the variables are not thought to represent indicators of a smaller number of constructs (as would be the case in discriminant analysis). Instead, the outcome variables are thought to be correlated yet distinct variables or be what Biskin (1980, p. 70) referred to as "conceptually independent." Raykov and Marcoulides (2008, p. 143) also note that a focus on separate outcomes may be considered when an "unambiguous" multivariate research question cannot be formulated due to insufficient information. Accordingly, we focus on whether the benefits found for MVMM over univariate procedures hold for the research context described above. In the next section, we present the multilevel models that are the focus of this study, after which we outline the simulation study conditions and highlight study results.

The Univariate and Multivariate Multilevel Models

In this study, we focus on random-intercept multilevel models for two reasons. First, the random-intercept MLM is commonly used in multilevel studies where the primary research objective is to examine group mean differences, that is, in cluster randomized trials, or CRTs (e.g., Carlson, Borman, & Robinson, 2011; Clements, Sarama, Spitler, Lange, & Wolfe, 2011; McCoach, Gubbins, Foreman, Rubenstein, & Rambo-Hernandez, 2014; Snyder, Vuchinich, Acock, Washburn, & Flay, 2012). In addition, estimates of design parameters for use in sample size determination and power analysis of CRTs are based on analyses of data with random-intercept MLMs (Hedges & Hedberg, 2007; Jacob, Zhu, & Bloom, 2010; Kelcey & Phelps, 2013; Xu & Nichols, 2010). A second reason we focus on random-intercept models is that they are often the model of choice in applied MVMM studies. For example, in the studies we reported above that used MVMM, all, except Freund et al. (2007), reported results *only* from random-intercept MVMMs fit to their data. We note though that even Freund et al. reported results from the random intercept model because the slopes associated with the primary predictors in their initial model did not vary significantly across clusters.

The Random-Intercept Multilevel Model

Conventional two-level MLMs partition the total variance in a single continuous outcome into within-cluster variance at the individual level (level-1) and between-cluster variance at the cluster level (level-2). In a two-level random-intercept model that includes predictor variables at each level, the individual-level equation is

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}_j) + r_{ij}, \tag{1}$$

where Y_{ij} represents the score on a normal outcome variable for individual *i* in cluster *j*, X_{ij} represents individual *i*'s score on a covariate, and $\overline{X}_j$ represents the mean covariate in cluster *j*. The term $(X_{ij} - \overline{X}_j)$ indicates that the covariate is group-mean centered. Such centering is useful because the regression coefficient associated with this predictor reflects *only* the within-cluster association (and not the between association) for *X* and *Y*, which is often preferred (Enders & Tofighi, 2007; Pituch & Stevens, 2016, Chapter 13). The intercept, β_{0j}, represents the unadjusted *Y* mean for cluster *j*. The slope, β_{1j}, represents the expected increase in *Y* for a one unit increase in *X* (or the "effect" of the covariate) in cluster *j*. The residual, r_{ij}, is assumed to be normally distributed with a mean of zero and some variance var(*r*).

The cluster-level model, assuming a two-group cluster randomized trial, is

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\ \beta_{1j} = \gamma_{10} \end{cases}, \tag{2}$$

where Z_j represents a dummy-coded treatment variable for cluster *j*, γ_{00} represents the predicted cluster mean when $Z_j = 0$, γ_{01} represents the difference between the treatment groups on the outcome mean for cluster *j*, and γ_{10} represents the overall average of the within-cluster slopes of of *Y* on *X*. The residual, u_{0j}, is assumed to be normally distributed with a mean of zero and variance $\text{var}(u_0)$. This residual allows the individual-level intercept, β_{0j}, to vary across clusters. The individual-level slope, β_{1j}, on the other hand, is treated as constant across all clusters, as indicated by the absence of a residual term. Note that, if desired, $\overline{X}_j$ may be included as a predictor in Equation 2, which may result in increased power for the test of the treatment if $\overline{X}_j$ were strongly related to β_{0j}.

Substituting Equation 2 into Equation 1 yields the combined MLM:

$$Y_{ij} = \gamma_{00} + \gamma_{01} Z_j + \gamma_{10}(X_{ij} - \overline{X}_j) + u_{0j} + r_{ij}. \tag{3}$$

Note that if multiple outcomes were present, this model could then be estimated separately for each outcome.

The degree of correlation among individual scores within a cluster, after controlling for the covariates, is measured by the residual intraclass correlation coefficient (ICC) (Snijders & Bosker, 2012). The formula for the residual ICC is as follows:

$$\text{ICC} = \frac{\text{var}(u_0)}{\text{var}(u_0) + \text{var}(r)}. \tag{4}$$

The Random-Intercept Multivariate Multilevel Model (MVMM)

The MVMM counterpart to the univariate model described above is often expressed using three model levels. The level-1 equation models the multivariate structure of the data. For data with two continuous outcomes, the equation is

$$Y_{hij} = \pi_{1ij} d_{1ij} + \pi_{2ij} d_{2ij}, \tag{5}$$

where Y_{hij} represents the score on outcome h (h = 1 or 2) for individual i in cluster j, d_{1ij} and d_{2ij} are dummy coded variables so that $d_{1ij} = 1$ for the first outcome and zero for the second, and $d_{2ij} = 1$ for the second outcome and zero for the first. As such,

$$\begin{aligned} Y_{1ij} &= \pi_{1ij}(1) + \pi_{2ij}(0) = \pi_{1ij} \\ Y_{2ij} &= \pi_{1ij}(0) + \pi_{2ij}(1) = \pi_{2ij} \end{aligned}.$$

The individual-level (level-2) equation resembles the level-1 equation of the MLM model:

$$\begin{cases} \pi_{1ij} = Y_{1ij} = \beta_{10j} + \beta_{11j}(X_{ij} - \overline{X}_j) + r_{1ij} \\ \pi_{2ij} = Y_{2ij} = \beta_{20j} + \beta_{21j}(X_{ij} - \overline{X}_j) + r_{2ij} \end{cases}, \tag{6}$$

where group-mean centering is used for each predictor. The within-cluster residuals, r_{1ij} and r_{2ij}, are assumed to be multivariate normally distributed with means of zero and constant covariance matrix $\boldsymbol{\Omega}_r$, where

$$\boldsymbol{\Omega}_r = \begin{bmatrix} \text{var}(r_1) & \\ \text{cov}(r_1, r_2) & \text{var}(r_2) \end{bmatrix}. \tag{7}$$

The cluster-level (level-3) equation is similar to the level-2 equation of the MLM model:

$$\begin{cases} \beta_{10j} = \gamma_{100} + \gamma_{101} Z_j + u_{10j} \\ \beta_{11j} = \gamma_{110} \\ \beta_{20j} = \gamma_{200} + \gamma_{201} Z_j + u_{20j} \\ \beta_{21j} = \gamma_{210} \end{cases}, \tag{8}$$

where Z_j represents a dummy-coded treatment variable for cluster j as in the MLM. The between-cluster intercept residuals, u_{10j} and u_{20j}, are assumed to be multivariate normally distributed with means of zero and constant covariance matrix $\boldsymbol{\Omega}_u$, where

$$\boldsymbol{\Omega}_u = \begin{bmatrix} \text{var}(u_{10}) & \\ \text{cov}(u_{10}, u_{20}) & \text{var}(u_{20}) \end{bmatrix}. \tag{9}$$

The combined model for the random-intercept MVMM with two outcomes is

$$\begin{cases} Y_{1ij} = \gamma_{100} + \gamma_{101} Z_j + \gamma_{110}(X_{ij} - \overline{X}_j) + u_{10j} + r_{1ij} \\ Y_{2ij} = \gamma_{200} + \gamma_{201} Z_j + \gamma_{210}(X_{ij} - \overline{X}_j) + u_{20j} + r_{2ij} \end{cases}. \tag{10}$$

The variances in Equations 7 and 9 can be used to calculate the residual ICC for each outcome *h*, as follows:

$$\text{ICC}_h = \frac{\text{var}(u_{h0})}{\text{var}(u_{h0}) + \text{var}(r_h)}. \quad (11)$$

In addition, the variances and covariances can be used to calculate the conditional correlation (*r*) between pairs of outcomes (*h* and *h′*) at the individual- and cluster-levels:

$$\rho_{hh'} = \frac{\text{cov}(r_h, r_{h'})}{\sqrt{\text{var}(r_h) * \text{var}(r_{h'})}} \quad (12)$$

$$\rho_{hh'} = \frac{\text{cov}(u_{h0}, u_{h'0})}{\sqrt{\text{var}(u_{h0}) * \text{var}(u_{h'0})}} \quad (13)$$

The conditional correlations represent the correlations among the outcomes that remain after inclusion of the covariates in the model.

Method

A Monte Carlo simulation study was conducted to compare the performance of the univariate and multivariate MLMs across 324 conditions, each replicated 1,000 times. SAS 9.4 was used to generate the data and estimate the parameters.

Generating and Estimating Models

Data were generated according to MVMMs with three and then with four outcomes. Data were analyzed using the random-intercept MLMs and MVMMs discussed previously.

Simulation Conditions

The six manipulated factors that comprised the 324 conditions were: number of outcomes (three and four), number of clusters (10, 30, and 50), cluster size (five, 10, and 30), residual ICC (.10, .20, and .30), conditional correlation between pairs of outcomes (.40, .60, and .80), and degree of imbalance of the two-group design (50/50 and 70/30). The 50/50 ratio represents a balanced design in which the clusters are divided equally into two treatment groups. The 70/30 ratio indicates that 70% of the clusters are assigned to one treatment group and 30% to the comparison group. The selection of factors and their levels began with conditions that have been shown in previous studies (described below) to be generally sufficient or necessary for acceptable univariate MLM results. Then, we included additional factors that would allow us to determine whether we could generalize what is already known from the univariate procedure to situations with multiple correlated outcomes.

The numbers of clusters, cluster size, and ICC values chosen for this study are commonly found in the simulation design of studies on adequate sample sizes for MLMs (e.g., Bell et al., 2010; Maas & Hox, 2004, 2005). Assuming a balanced design, Maas and Hox (2004, 2005) found that as few as 10 clusters along with a cluster size of five were sufficient for obtaining fairly unbiased estimates of the fixed effects, while at least 30 clusters were needed for reasonable estimates of the standard errors and variances. Maas and Hox also varied the ICC from .10 to .30, although no statistically significant effect was found on either parameter or standard error bias. For statistical power, Bell et al. (2010) recommended sample sizes larger than 30 clusters with a cluster size of 30.

In addition, methodological studies on MLM have investigated the extent to which unequal cluster sizes, unequal number of clusters, or both can impact estimates, power, and Type I error rates (e.g., Bell, Ferron, & Kromrey, 2008; Browne & Draper, 2000; Cools, van den Noortgate, & Onghena, 2009; Konstantopoulos, 2010; Steele, Mundfrom, & Perrett, 2011). The overall conclusion across these studies is that mild or moderate imbalance generally can be ignored. In the current study, the 70/30 ratio represents a moderate imbalance at the cluster level (Konstantopoulos, 2010).

The number of outcomes and values for the outcome correlations were drawn from a range of values found in methodological studies on the power of MANOVA (e.g., Cole, Maxwell, Arvey, & Salas, 1994; Frane, 2015; Stevens, 1980). These studies found that power was a function of the degree of intercorrelations among outcomes. Furthermore, the intercorrelations interacted with effect sizes, resulting in different "spots" of power advantage between the traditional MANOVA and ANOVA

procedures (Frane, 2015). Note that referencing MANOVA here is appropriate for studies on MVMMs as both procedures are multivariate and because MANOVA is analogous to (albeit less flexible and less powerful than) a *two*-level MVMM with outcomes nested within clusters (Hox, 2010; Park et al., 2015).

Generating Parameter Values

The generating parameter values for the fixed effects and variances are shown in Tables 1 and 2, respectively. To allow for an assessment of the Type I error rate, a parameter value of zero was used for the regression coefficients associated with the individual- and cluster-level predictors for the first outcome. For the other outcomes, the cluster-level coefficients (γ_{h01}) were manipulated to correspond to a range of Cohen's *d* values and obtained using the following equations for calculating effect sizes from clustered data (Hedges, 2007):

$$d_h = \frac{\gamma_{h01}}{\sqrt{\text{var}(u_{h0}) + \text{var}(r_h)}} \tag{14}$$

$$d_h = \frac{\gamma_{h10}}{\sqrt{\text{var}(u_{h0}) + \text{var}(r_h)}} \tag{15}$$

The values of the cluster-level coefficients were then repeated for the individual-level effects (γ_{h10}).

The values of the within- and between-cluster variances were a function of the ICC (see Equation 11) and of the total variance. A total (within and between) variance of 562.50 was used and then fixed across all outcomes. The values of the covariances were conditional on the outcome correlations and calculated using Equations 12 and 13.

Predictor Variables

Values for the individual-level predictor were randomly drawn from a normal distribution with a mean of zero and a standard deviation of one. For the dummy-coded cluster-level variable in the 50/50 design, the first half of the clusters within each replication was assigned a value of zero and the other half assigned a value of one. In the 70/30 design, 70% of the clusters were assigned a value of zero and the rest assigned a value of one.

Table 1. Generating Parameter Values for the Fixed Effects

Outcome	Intercept	γ_{h10}	γ_{h01} (Cohen's *d*)
1	90	0.00	0.00 (*d* = 0.00)
2	90	4.74	4.74 (*d* = 0.20)
3	90	11.86	11.86 (*d* = 0.50)
4	90	18.97	18.97 (*d* = 0.80)

Note. γ_{h10} = individual-level fixed effect for outcome *h*; γ_{h01} = cluster-level fixed effect for outcome *h*.

Table 2. Generating Parameter Values for the Variances and Covariances

ICC	Correlation	$\text{var}(r_h)$	$\text{var}(u_{h0})$	$\text{cov}(r_h, r_{h'})$	$\text{cov}(u_{h0}, u_{h'0})$
.10	.40	506.25	56.25	202.50	22.50
	.60	506.25	56.25	303.75	33.75
	.80	506.25	56.25	405.00	45.00
.20	.40	450.00	112.50	180.00	45.00
	.60	450.00	112.50	270.00	67.50
	.80	450.00	112.50	360.00	90.00
.30	.40	393.75	168.75	157.50	67.50
	.60	393.75	168.75	236.25	101.25
	.80	393.75	168.75	315.00	135.00

Note. $\text{var}(r_h)$ = individual-level variance for outcome *h*; $\text{var}(u_{h0})$ = cluster-level variance for outcome *h*; $\text{cov}(r_h, r_{h'})$ = individual-level covariance between outcomes *h* and *h′*; $\text{cov}(u_{h0}, u_{h'0})$ = cluster-level covariance between outcomes *h* and *h′*.

Estimation Method

Parameters were estimated using SAS PROC MIXED with restricted maximum likelihood (REML) estimation. In addition, given the inclusion of unbalanced data sets and small sample sizes in the simulation, the Kenward and Roger (1997) degrees of freedom was specified for the fixed effects test statistics, as recommended by Schaalje, McBride, and Fellingham (2001).

Analyses

For each condition, 1,000 sets of converged model estimates were analyzed in terms of relative parameter bias, relative standard error bias, power, and Type I error. Since our primary interest is in comparing the MLM and MVMM results, we focused our analysis on the parameters that are estimated by both models (i.e., the fixed effects and variances).

Relative parameter bias (RPB). Relative parameter bias was calculated to evaluate the accuracy of estimates of the non-null fixed effects and variances in each condition. The formula for relative parameter bias (Hoogland & Boomsma, 1998) is:

$$RPB = \frac{\bar{\theta}_i - \theta_i}{\theta_i} \tag{16}$$

where θ_i is the population value of parameter *i* and $\bar{\hat{\theta}}_i$ is the parameter estimate averaged across the 1,000 replications in each condition. Absolute values of relative parameter bias less than Hoogland and Boomsma's recommended cutoff of .05 were considered acceptable.

Relative standard error bias (RSEB). The accuracy of the standard error estimates of the non-null fixed effects and variances was assessed using relative standard error bias. The formula for relative standard error bias (Hoogland & Boomsma, 1998) is:

$$RSEB = \frac{\overline{SE}_{\hat{\theta}_i} - SE_{\theta_i}}{SE_{\theta_i}} \tag{17}$$

where SE_{θ_i} is the standard deviation of the parameter estimates (θ_is) across the 1,000 replications in each condition and $\overline{SE}_{\hat{\theta}_i}$ is the mean standard error estimate of θ_i. Relative standard error bias values with a magnitude less than Hoogland and Boomsma's cutoff of .10 were deemed acceptable.

Power. For each non-null fixed effect in each condition, the empirical power rate was calculated as the number of times each null hypothesis of no effect was correctly rejected at an alpha-level of .05, divided by 1,000 replications.

Type I error. For each null fixed effect in each condition, the empirical Type I error rate was calculated as the number of times each null hypothesis was incorrectly rejected at an alpha-level of .05, divided by 1,000 replications. A Type I error rate within the interval of .025 to .075 was considered acceptable (Bradley, 1978).

Hypotheses

Based on the Baldwin et al. study (2014) and given no missing data, we hypothesize that differences between the MLM and MVMM in the common parameters estimated (i.e., fixed effects and variances) are expected to be minimal. Similarly, there should be essentially no differences between these two statistical models with regard to power and Type I error rate. However, given the small sample size conditions included in this study, it is reasonable to expect that MVMM may experience greater convergence and/or estimation problems due to the larger number of estimated parameters. That is, in the 3 outcome condition, the MVMM will estimate 6 outcome covariances (3 within- and 3 between-cluster covariances), and in the 4 outcome conditions, will estimate 12 covariances (6 within- and 6 between-cluster covariances). No outcome covariances are estimated in the MLM approach.

Results

Results from 1,000 sets of converged estimates per condition show virtually no difference between MLM and MVMM, as values of relative bias, power, and Type I error rates for the common parameters (i.e., fixed effects and variances) were almost all identical to the third decimal place. Details of the results,

including convergence issues, are presented below, followed by a discussion of the implications of our finding of no difference between models.

Convergence. Convergence issues in the form of negative intercept variance estimates occurred in equal frequency between MLM and MVMM. Non-convergence resulted in the need for additional replications in order to obtain 1,000 sets of converged estimates, particularly in the conditions with fewer than 30 clusters and an ICC of .10.

Relative parameter bias (RPB). For the individual-level non-null fixed effects, relative bias values ranged from –0.053 to 0.046 (M = 0.001, SD = 0.013) for estimates of γ_{210}, from –0.019 to 0.023 (M = 0.000, SD = 0.005) for estimates of γ_{310}, and from –0.014 to 0.014 (M = 0.000, SD = 0.003) for estimates of γ_{410}. With the exception of one condition in which γ_{210} was substantially underestimated (RPB = –0.053), these results indicate that estimates of the individual-level fixed effects were unbiased.

All estimates of the cluster-level fixed effects with Cohen's d effect sizes of 0.50 and 0.80 were also unbiased. For estimates of γ_{301} (d = 0.50), relative bias values ranged from –0.045 to 0.047 (M = 0.000, SD = 0.014). For estimates of γ_{401} (d = 0.80), relative bias values ranged from –0.027 to 0.024 (M = -0.001, SD = 0.008). On the other hand, the relative bias of estimates of γ_{201} (d = 0.20) ranged from -0.131 to 0.133 (M = 0.000, SD = 0.038). The estimates were substantially negatively biased (RPB = -0.131 to –0.053) in 8.9% of the conditions and substantially positively biased (RPB = 0.051 to 0.133) in 6.8% of the conditions, with substantial bias more likely to occur with fewer clusters, smaller cluster sizes, and larger ICCs. Specifically, the percentage of conditions with substantial bias decreased considerably from 34.4% to 2.8% as number of clusters increased from 10 to 50 and from 22.2% to 7.4% as cluster size increased from five to 30. However, the percentage increased considerably from 6.5% to 27.8% as ICC increased.

Across all outcomes, relative bias of estimates of the individual-level variances ($var(r_h)$) ranged from –0.061 to 0.009 (M = –0.005, SD = 0.011).The variances were substantially underestimated (RPB = –0.061 to –0.051) in 2.8% of the conditions and only with the fewest number of clusters (10), the smallest cluster size (five), and the smallest ICC (.10) .

A larger proportion of estimates of the cluster-level variances ($var(u_{h0})$) was substantially biased compared to the variances at the individual-level. The overall mean relative bias was 0.068 (SD = 0.167), which exceeds the acceptable upper bound of 0.05. Relative bias values ranged from –0.032 to 0.996, with substantial positive bias (RPB = 0.051 to 0.996) in 26.2% of the conditions. Substantial bias was more likely to occur with fewer clusters, smaller cluster sizes, and smaller ICCs, as both the mean relative bias and the percentage of conditions with substantial bias decreased considerably as number of clusters, cluster size, and ICC increased, as shown in Table 3.

Relative standard error bias (RSEB). Relative bias of the standard error estimates of the individual-level non-null fixed effects ranged from –0.078 to 0.060 (M = –0.004, SD = 0.024) for γ_{210}, from –0.088 to 0.074 (M = –0.005, SD = 0.025) for γ_{310}, and from –0.070 to 0.062 (M = 0.000, SD = 0.026) for γ_{410}. These results indicate that the standard error estimates were all unbiased.

At the cluster-level, relative bias of the standard error estimates ranged from –0.082 to 0.137 (M = 0.000, SD = 0.033) for γ_{201}, –0.086 to 0.132 (M = 0.001, SD = 0.036) for γ_{301}, and –0.068 to 0.121 (M = 0.006, SD = 0.036) for γ_{401}. The standard errors for γ_{401} were substantially overestimated (RSEB = 0.121) in only one condition, while the standard errors for γ_{201} were substantially overestimated (RSEB = 0.111 to 0.137) in 1.5% of the conditions, and the standard errors for γ_{301} were substantially overestimated (RSEB = 0.105 to 0.132) in 2.2% of the conditions. All of the biased standard error estimates occurred in the conditions with the fewest number of clusters (10), smallest cluster size (five), and smallest ICC (.10) across the range of effect sizes.

The standard error estimates of the individual-level variances were all unbiased, with the exception of one condition that substantially overestimated the residual variance of the second outcome (RSEB = 0.102). Relative bias ranged from –0.069 to 0.102 (M = 0.004, SD = 0.024). In contrast, the relative bias

Table 3. Mean Relative Parameter Bias (RPB) of the Cluster-Level Variances, *var*(u_{h0}), and the Percentage of Conditions with Substantial Bias at Each Level of a Factor, Collapsed Across All Other Factors

		MLM		MVMM	
		Mean (SD)	%	Mean (SD)	%
Outcomes	3	**0.064** (0.159)	25.9%	**0.064** (0.159)	25.9%
	4	**0.072** (0.172)	27.2%	**0.072** (0.172)	27.2%
Clusters	10	**0.172** (0.251)	57.4%	**0.172** (0.251)	57.4%
	30	0.024 (0.058)	11.1%	0.024 (0.058)	11.1%
	50	0.009 (0.025)	11.1%	0.009 (0.025)	11.1%
Cluster size	5	**0.157** (0.250)	55.6%	**0.157** (0.250)	55.6%
	10	0.044 (0.091)	22.2%	0.044 (0.091)	22.2%
	30	0.004 (0.015)	1.9%	0.004 (0.014)	1.9%
ICC	.10	**0.156** (0.254)	46.3%	**0.156** (0.254)	46.3%
	.20	0.037 (0.077)	21.3%	0.037 (0.077)	21.3%
	.30	0.012 (0.031)	12.0%	0.012 (0.031)	12.0%
Correlation	.40	**0.059** (0.146)	25.9%	**0.059** (0.146)	25.9%
	.60	**0.071** (0.172)	26.9%	**0.071** (0.172)	26.9%
	.80	**0.075** (0.180)	26.9%	**0.075** (0.180)	26.9%
Imbalance	50/50	**0.069** (0.168)	26.5%	**0.069** (0.168)	26.5%
	70/30	**0.068** (0.166)	26.5%	**0.068** (0.166)	26.5%

Note. ICC = intraclass correlation coefficient; bolded values indicate substantial bias (mean RPB > 0.05).

of the standard error estimates ranged from –0.060 to 0.472 (*M* = 0.049, *SD* = 0.088) for the cluster-level variances. The standard errors were substantially overestimated (RSEB = .102 to .472) in 21.6% of the conditions. As with the results for relative parameter bias, substantial standard error bias was more likely to occur with fewer clusters, smaller cluster sizes, and smaller ICCs. The percentage of conditions with substantial bias decreased considerably from 46.3% to 7.4% as number of clusters increased from 10 to 50, from 48.1% to 0.0% as cluster size increased from five to 30, and from 40.7% to 7.4% as ICC increased from .10 to .30.

Power. At the individual-level, empirical power rates of tests of γ_{310} and γ_{410} were all high, ranging from .880 to 1.000 (*M* = .991, *SD* = .027) for γ_{310} and .997 to 1.000 (*M* = 1.000, *SD* = .001) for γ_{410}. For the parameter with the smallest effect size, γ_{210}, empirical power rates ranged from .255 to 1.000 (*M* = .814, *SD* = .238), with power of less than .80 in 33.3% of the conditions. As expected, power improved as sample sizes increased, with power exceeding .80 in all of the conditions with 50 clusters and a cluster size of 30.

Tests of the cluster-level fixed effects were considerably less powerful than tests of the individual-level fixed effects. Empirical power rates of tests of γ_{201} (*d* = .20) were all less than .80, ranging from .030 to .515 (*M* = .165, *SD* = .099). Power rates for γ_{301} (*d* = .50) ranged from .124 to .999 (*M* = .588, *SD* = .286) and were less than .80 in 68.8% of the conditions. Power rates for γ_{401} (*d* = .80) ranged from .241 to 1.000 (*M* = .817, *SD* = .214) and were smaller than .80 in 30.2% of the conditions. Not surprisingly, power improved as effect size, number of clusters, and cluster size increased, as shown in Table 4. However, power tended to decrease as ICC increased, holding all other factors constant. Note that the power values displayed in Table 4 are collapsed across MLM and MVMM analyses as the values are identical to the third decimal place.

Table 4. Mean Power of Tests of the Cluster-Level Fixed Effects Across the MLM and MVMM Analyses and the Percentage of Conditions With Power Rates Below .80 for Each Factor Level, Collapsing Across All Other Factors

		γ_{201} (d = 0.20)		γ_{301} (d = 0.50)		γ_{401} (d = 0.80)	
		Mean (SD)	%	Mean (SD)	%	Mean (SD)	%
Outcomes	3	**.166** (.099)	100.0%	**.588** (.285)	69.8%		
	4	**.163** (.098)	100.0%	**.589** (.287)	67.9%	.817 (.241)	30.2%
Clusters	10	**.067** (.022)	100.0%	**.233** (.090)	100.0%	**.504** (.154)	90.7%
	30	**.169** (.054)	100.0%	**.670** (.143)	78.7%	.951 (.045)	0.0%
	50	**.258** (.088)	100.0%	.861 (.097)	27.8%	.995 (.008)	0.0%
Cluster size	5	**.126** (.067)	100.0%	**.500** (.270)	84.3%	**.759** (.281)	33.3%
	10	**.164** (.090)	100.0%	**.596** (.286)	65.7%	.822 (.231)	33.3%
	30	**.204** (.117)	100.0%	**.669** (.279)	56.5%	.869 (.194)	24.1%
ICC	.10	**.208** (.131)	100.0%	**.679** (.301)	48.1%	.869 (.206)	24.1%
	.20	**.156** (.078)	100.0%	**.583** (.280)	70.4%	.815 (.242)	33.3%
	.30	**.130** (.055)	100.0%	**.502** (.248)	88.0%	**.766** (.265)	33.3%
Correlation	.40	**.165** (.098)	100.0%	**.589** (.286)	66.7%	.818 (.240)	29.6%
	.60	**.165** (.101)	100.0%	**.587** (.286)	69.4%	.817 (.243)	29.6%
	.80	**.163** (.099)	100.0%	**.588** (.288)	70.4%	.815 (.244)	31.5%
Imbalance	50/50	**.176** (.107)	100.0%	**.613** (.289)	62.3%	.832 (.230)	29.6%
	70/30	**.153** (.089)	100.0%	**.563** (.281)	75.3%	.801 (.252)	30.9%

Note. d = Cohen's d effect size; ICC = intraclass correlation coefficient; bolded values indicate mean power < .80.

Type I Error. Type I error rates for the test of the individual-level null fixed effect were all within acceptable bounds, ranging from .034 to .072 (M = .050, SD = .007). At the cluster-level, Type I error rates ranged from .010 to .070 (M = .045, SD = .010) and were too conservative (i.e., less than .025) in 4.6% of the conditions. All of the unacceptable Type I error rates occurred in the conditions with 10 clusters, a cluster size of five or 10, and an ICC of .10 or .20. With 30 or more clusters, the largest cluster size (30), and the largest ICC (.30), Type I error rates were all close to the nominal value.

Discussion

The goal of this study was to determine if there are differences in the quality of parameter estimates obtained from random intercept univariate and multivariate multilevel models. The context in which this study was implemented was a cluster-randomized design with a dummy-coded treatment indicator variable at the cluster level, a continuous predictor at the individual level, and correlated continuous outcomes at the individual level, although the results apply generally to multilevel cross-sectional studies with these same variable configurations. Further, the context of our study assumes that researchers are interested in estimating fixed effects and variances at the individual and cluster level, but not covariances (or correlations) for each of several correlated outcomes. We further assumed that data were complete for all variables and that researchers were not interested in assessing whether the effects of the treatment differed across the multiple outcomes nor in estimating the within-cluster and between-cluster correlations among outcomes.

In this context, our simulation study estimated parameter and standard error bias associated with the within cluster and between cluster variances for the multiple outcomes as well as the fixed effects associated with the individual and cluster level predictors. We also estimated the power and Type I error rate associated with the test of each of the fixed effects. While other studies (Baldwin et al., 2014; Hauck & Street, 2006) analyzed an existing data set or simulated data for a single condition, we examined the performance of the MLM and MVMM across 324 conditions. The simulated conditions varied by number of outcomes, number of clusters, cluster size, intraclass correlation, outcome correlation, and degree of

imbalance. We also included a range of effect size values for the fixed effects by assigning small, moderate, and large effect sizes to the different outcomes.

For these conditions, we found, consistent particularly with the results from Baldwin et al. (2014), that there was remarkable similarity in the performance of MLM and MVMM, as there were no real differences in parameter estimation, standard error bias, power, and Type I error accuracy. In addition, we found no difference in convergence problems. For both models, parameters and standard errors were more accurately estimated, tests of the non-null fixed effects were more powerful, and Type I error rates were more accurate at the individual-level than at the cluster-level. Furthermore, in most cases, estimation accuracy, power, and Type I error rates improved considerably as number of clusters, cluster size, and ICC increased. However, for number of outcomes, outcome correlation, and degree of imbalance, any differences in results across the levels of each of these factors were regarded as trivial.

The findings from our current study suggest that if the restrictive conditions we implemented are present for an applied study and if the applied researcher were interested only in estimating fixed effects and variances, and not the covariances, the MLM and the MVMM may be used interchangeably. This similarity of performance assumes that the models of interest are simple random-intercept models, imbalance occurs only between two treatment groups and not across cluster sizes, there is no missing data, and researchers are not interested testing the equivalence of a predictor's impact across multiple outcomes.

The conditions we implemented here are important to keep in mind because the literature suggests that the MVMM may have better performance than the MLM when other conditions are present. For example, Snijders and Bosker (2012) and Hox (2010) point out that the MVMM approach is expected to yield more power and more accurate Type I error rates than the univariate MLM approach when data are missing and outcomes are correlated. While we affirmed in this study that the MLM and MVMM perform similarly in certain conditions when data are complete, the work of Park et al. (2015) showed that, with non-clustered data, MVMM provides greater power than a univariate analysis approach when outcome data are incomplete.

Taken as a whole, the results of this study provide support both to those wishing to use the simpler MLM and the more complex MVMM. That is, our results indicated that in the conditions we examined there is no benefit to using the more complex MVMM procedure, in that the parameter estimates, power, and Type I error accuracy from a specific multivariate multilevel design are virtually identical for the two statistical models. On the other hand, given research conditions that generally favor use of the univariate MLM, as implemented in this study, our results suggest that the MVMM could be used in place of the MLM without sacrificing power, Type I error accuracy, or the quality of parameter estimates. This is important for applied researchers, who, working in a similar multivariate multilevel context as implemented in this study, wish to use the MVMM, for example, to test for the equivalence of the effect of a predictor across multiple outcomes. Further, note that the MVMM performed just as well as the MLM under the smaller sample and effect size conditions implemented in this study. As a whole, then, our research supports the continued use of MVMM in multivariate multilevel designs. It is important, though that future research continues to assess the performance of the MVMM to learn of its limitations and potential benefits compared to using univariate MLMs.

Finally, as a practical note, in an applied study, it is generally important to use a Bonferroni-adjusted alpha for the tests of each outcome variable to preserve the family-wise error rate. We used an unadjusted alpha in our study because we simply wished to compare the performance of the two statistical models, and the use of the unadjusted alpha level has no bearing on this comparison. Also, we remind readers that we did not use a protected testing strategy that is often used in traditional MANOVA. The work of Frane (2015) suggests that the Bonferonni procedure as applied to the tests of specific outcomes often performs as well as or better than use of an omnibus testing strategy, particularly when the number of outcome variables is larger than two.

References

Baldwin, S. A., Imel, Z. E., Braithwaite, S. R., & Atkins, D. C. (2014). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Counseling and Clinical Psychology*. Advance online publication. doi: doi.org/10.1037/a0035628

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

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Loudermilk, B., L., Kromrey, J. D., & Ferron, J. M. (2010). *Dancing the sample size limbo with mixed models: How low can you go?* SAS Global Forum 2010 Paper 197-2010.

Biskin, B. H. (1980). Multivariate analysis in experimental counseling research. *The Counseling Psychologist, 8*, 69–72. doi: 10.1177/001100008000800422

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*(3), 391-420.

Butler, E. A., Gross, J. J., & Barnard, K. (2014). Testing the effects of suppression and reappraisal on emotional concordance using a multivariate multilevel model. *Biological Psychology, 98*, 6-18.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*(3), 378-398.

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education, 42*(2), 127-166.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115*(3), 465-474.

Cools, W., Van den Noortgate, W., & Onghena, P. (2009). Design efficiency for imbalanced multilevel data. *Behavior Research Methods, 41*(1), 192-203.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138.

Frane, A. V. (2015). Power and type I error control for univariate comparisons in multivariate two-group designs. *Multivariate Behavioral Research, 50*(2), 233-247.

Freund, P. A., Holling, H., & Preckel, F. (2007). A multivariate, multilevel analysis of the relationship between cognitive abilities and scholastic achievement. *Journal of Individual Differences, 28*(4), 188-197.

Goldstein, H. (2011). *Multilevel statistical models.* Hoboken, NJ: Wiley.

Hauck, K., & Street, A. (2006). Performance assessment in the context of multiple objectives: A multivariate multilevel analysis. *Journal of Health Economics, 25*(6), 1029-1048.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341-370.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*(1), 101-117.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Hoboken, NJ: Taylor & Francis.

Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness, 3*(2), 157-198.

Kelcey, B., & Phelps, G. (2013). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis, 35*(3), 370-390.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983-997.

Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education,78*(3), 291-317.

Korpershoek, H., Kuyper, H., & Van Der Werf, G. (2015). The relationship between students' math and reading ability and their mathematics, physics, and chemistry examination grades in secondary education. *International Journal of Science and Mathematics Education, 13*(5), 1013-1037.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

McCoach, B. D., Gubbins, E. J., Foreman, J., Rubenstein, L. D., & Rambo-Hernandez, K. E. (2014). Evaluating the efficacy of using predifferentiated and enriched mathematics curricula for grade 3 students: A multisite cluster-randomized trial. *Gifted Child Quarterly, 58*(4), 272-286.

Olaizola, J. H., Diaz, F. J. R., & Ochoa, G. M. (2014). Comparing intergroup contact effects on blatant and subtle prejudice in adolescents: A multivariate multilevel model. *Psicothema, 26*(1), 33-38.

Park, R., Pituch, K. A., Kim, J., & Chung, H., & Dodd, B. G. (2015). Comparing the performance of multivariate multilevel modeling to traditional analyses with complete and incomplete data. *Methodology, 11*(3), 100-109.

Paterson, L. (1998). Multilevel multivariate regression: An illlustration concerning school teachers' perceptions of their pupils. *Educational Research and Evaluation, 4*(2), 126-142.

Pierewan, A. C., & Tampubolon, G. (2015). Happiness and health in Europe: A multivariate multilevel model. *Applied Research in Quality of Life, 10*(2), 237-252.

Pituch, K. A., & Stevens, J. P. (2016). *Applied Multivariate Statistics for the Social Sciences* (6th ed.). New York: Routledge.

Pituch, K. A., Whittaker, T. A., Chang, W. (2016). Multivariate models for normal and binary responses in intervention studies. *American Journal of Evaluation,* doi: 10.1177/ 10982 14015626297.

Plewis, I. (2005). Modeling behaviour with multivariate multilevel growth curves. *Methodology, 1*(2), 71-80.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis.* New York: Routledge.SAS (Version 9.4) [Computer software]. Cary, NC: SAS.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2001). *Approximations to distributions of test statistics in complex mixed linear models using SAS Proc MIXED*. SAS Global Forum Paper 262-26.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, England: Sage Publications.

Snyder, F. J., Vuchinich, S., Acock, A., Washburn, I. J., & Flay, B. R. (2012). Improving elementary school quality through the use of a social-emotional and character development program: A matched-pair, cluster-randomized controlled trial in Hawai`i. *Journal of School Health, 82*(1), 11-20.

Steele, B. J., Mundfrom, D. J., & Perrett, J. (2011). Unbalanced sampling effect on the power at level-1 in the random coefficient model. *Multiple Linear Regression Viewpoints, 37*(1), 22-35.

Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin, 88*(3), 728-737.

Tate, R. L., & Pituch, K. A. (2007). Multivariate hierarchical linear modeling in randomized field experiments. *The Journal of Experimental Education, 75*(4), 317-337.

Timm, N. H. (2002). *Applied Multivariate Analysis*. New York: Springer.

Turner, R. M., Omar, R. A., & Thompson, S. G. (2006). Modelling multivariate outcomes in hierarchical data, with application to cluster randomized trials. *Biometrical Journal, 48*(3), 333-345.

Veiga, A., Smith, P. W. F., & Brown, J. J. (2014). The use of sample weights in multivariate multilevel models with an application to income data collected by using a rotating panel survey. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 63*(1), 65-84.

Xu, Z., & Nichols, A. (2010). *New estimates of design parameters for clustered randomization studies: Findings from North Carolina and Florida.* Washington, DC: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

Yang, M., Goldstein, H., Browne, W., & Woodhouse, G. (2002). Multivariate multilevel analyses of examination results. *Journal of the Royal Statistical Society: Series A, 165*(1), 137-153.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57*(298), 348-368.

Send correspondence to: Wanchen Chang
Boise State University
Email: wanchenchang@boisestate.edu