The Impact of Community-Level Variables on Individual-Level Outcomes

Theoretical Results and Applications

GUSTAVO ANGELES
DAVID K. GUILKEY
THOMAS A. MROZ
University of North Carolina at Chapel Hill

The authors study alternative estimators of the impacts of higher level variables in multilevel models. This is important since many of the important variables in social science research are higher level factors having impacts on many lower level outcomes such as school achievement and contraceptive use. While the large sample properties of alternative estimators for these models are well known, there is little evidence about the relative performance of these estimators in the sample sizes typical in social science research. The authors attempt to fill this gap by presenting evidence about point estimation and standard error estimation for both two- and three-level models. A major conclusion of the article is that readily available commercial software can be used to obtain both reliable point estimates and coefficient standard errors in models with two or more levels as long as appropriate corrections are made for possible error correlations at the highest level.

Keywords: *multilevel models; hierarchical models; multilevel error structure; Monte Carlo simulations*

1. INTRODUCTION

Multilevel models have found widespread use in social science research in recent years. These types of models are used when the outcome of interest and its observed and unobserved determinants

AUTHORS' NOTE: *Funding support for this project was provided by the MEASURE Evaluation Project under a cooperative agreement between the U.S. Agency for International Development (USAID) and the Carolina Population Center* (*no. HRN-A*-00-97-0018-00)*. The views expressed herein are those of the authors and not the sponsoring agency.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 34, No. 1, August 2005 76-121
DOI: 10.1177/0049124104273069

have a hierarchical structure. By a hierarchical structure, we mean that there are important factors influencing decisions and outcomes that arise from a variety of levels of aggregation or observation. Kreft and de Leeuw (1998) provide an excellent introduction to multilevel models, and Goldstein (1995, 2003), Bryk and Raudenbush (1992), and Raudenbush and Bryk (2002) present more advanced treatments of these modeling approaches.

In multilevel models, the outcome of interest is typically measured at an individual level, and this usually is referred to as the lower or micro-level outcome. In analyses with more than one level, this is called the Level 1 outcome. These lower level, individual outcomes are usually influenced in part by individual, micro-level characteristics. For example, a woman's age and education have been shown to have important effects on her use of contraceptives (Gertler and Molyneaux 1994; Guilkey and Cochrane 1995; Guilkey and Jayne 1997). What distinguishes the hierarchy in these types of analyses is the fact that some characteristics from a higher level also influence the lower level outcomes. For example, it has been found that the presence of family planning clinics providing contraceptives within communities can affect whether individuals adopt family planning (Tsui 1985; Bollen, Guilkey, and Mroz 1995; Thomas and Maluccio 1995; Guilkey and Jayne 1997; Angeles et al. 2001), and school quality can affect student achievement (Guo and Zhao 2000). In fact, the impact of these higher level variables (community level or school level) is often of primary interest, while the Level 1 (individual level) variables are often simply control variables.

One can readily incorporate observed community-level characteristics along with observed individual-level characteristics as determinants of individual-level behaviors in regression models. The fact that these observed higher level characteristics do not differ within groups of individuals does not complicate estimation procedures. However, there can also be unobserved or unmeasured factors at the higher level that influence the lower level outcomes. Such unmeasured factors give rise to a multilevel error structure that has important statistical implications discussed extensively in the following sections.

There are several alternative estimation methods that can be used to estimate the parameters of multilevel models with continuous dependent variables, with ordinary least squares, maximum likelihood,

and feasible generalized least squares being the most common. Since these estimators are identical to methods for use with panel data, the large sample statistical properties of these estimators are well known and can be found in standard textbooks. However, there has been surprisingly little work done on the finite sample properties of the alternative estimators (see Mátyás [1992] for a review and Maas and Hox [2002] for recent work), and the work that has been done has focused on the properties of the estimated coefficients of individual-level (Level 1) variables in models with only a two-level error structure. The impacts of community-level (Level 2) variables, however, are frequently of considerable substantive interest because these are often policy variables that can be modified to affect individual-level outcomes. In addition, data sets frequently have more than two levels, and there is no Monte Carlo research that we are aware of that has examined the finite sample performance of alternative estimators when there are more than two levels.

The purpose of this article is to fill these gaps in the literature by focusing on the correct measurement and statistical testing of the expected impact of community-level variables on individual-level outcomes in multilevel models. We do so by presenting new theoretical results based on analytical derivations and Monte Carlo simulations for three estimators: ordinary least squares, ordinary least squares with corrected standard errors (as defined in the next section), and maximum likelihood. The simulated data in the Monte Carlo work are constructed to mimic the type of data that one would find in large survey data sets. Our results clearly demonstrate the usefulness of the ordinary least squares estimator with corrected standard errors for either two- or three-level models, an estimation method that is readily available in popular, commercially available software. We illustrate the methods in an analysis of children's weight using a data set from Cebu, Philippines, with three levels of observed and unobserved determinants: the community, the individual, and over time for the individual.

While the theoretical and simulation results we present are specific to models with continuous dependent variables, there is some analytical evidence (Robinson 1982) and simulation evidence (Guilkey and Murphy 1993) suggesting that the major conclusions of the study should carry over to models with limited dependent variables. We demonstrate

this in an empirical model of the determinants of contraception at the province, municipality, and individual levels. The data for this second application come from a different data set covering all of the Philippines.

The plan of this article is as follows. The next section lays out the estimation methods and presents analytical results on the relative efficiency of the alternative estimators. Since it is not possible to determine analytically how well the alternative estimators work for hypothesis testing, section 3 presents the results of Monte Carlo simulations. Sections 4 and 5 present the empirical applications, and we conclude in section 6.

2. THEORETICAL RESULTS

This section summarizes the results presented in Angeles and Mroz (2001). We present analytical results for the two-level model followed by Monte Carlo evidence for both two- and three-level models.

ANALYTICAL RESULTS

Consider the following model:

$$Y_{ic} = \beta_0 + \beta_1 X_c^C + \beta_2 X_{ic}^{IC} + \beta_2 X_{ic}^I + \mu_c + \epsilon_{ic}, \tag{1}$$

where Y_{ic} is a continuous outcome variable for respondent i ($i = 1, 2, \ldots, N_c$) from community c ($c = 1, 2, \ldots, C$). The βs are unknown regression coefficients, X_c^C is a community-level variable, X_{ic}^{IC} is an individual-level variable having covariance τ with the community-level variable ($\text{Cov}(X_c^C, X_{ic}^{IC}) = \tau$), and X_{ic}^I is an individual-level variable that is not correlated with either the community-level variable or the other individual-level variable. All the explanatory variables are independent of the error terms.

The error term is assumed to have two components. The first is a term that is specific to each individual, ϵ_{ic}; it is assumed to have mean zero and variance σ_ϵ^2 and to be independent for all i and c. The second unobserved component, μ_c affects the outcome Y for all individuals within each community; it is assumed to have mean zero, variance σ_μ^2 and to be independent across communities. This error components specification for the error term is the same as the standard

specification for a panel data model. If we define $\beta_{0c} = \beta_0 + \mu_c$, we see also that the model is a widely used random intercept/constant slope model (see, e.g., Raudenbush and Bryk 2002 and the review article by Guo and Zhao 2000).

Given the error term assumptions, we can define the total error variance as $\sigma^2 = \sigma^2_\epsilon + \sigma^2_\mu$. The fraction of the total error variance due to the community-level component of the error term is frequently referred to as the intraclass error correlation, and it is defined by $\rho = \sigma^2_\mu / \sigma^2$. Let $N_{TOT} = \sum_{c=1}^{C} N_c$ be the total number of individual-level (Level 1) outcomes. After ordering observations by communities, the $(N_{TOT} \times N_{TOT})$ error covariance matrix, Ω, is block diagonal with zeros on the off-diagonal blocks (blocks corresponding to covariances of error terms for different communities) and $N_c \times N_c$ diagonal blocks $\sigma^2 A_c$. The $A_c (c = 1, 2, \ldots, C)$ represent the covariance matrix for community c, where all elements of A_c are equal to ρ except the elements on the main diagonal, which are equal to 1.

Three estimators for the βs and their standard errors are commonly used: ordinary least squares (OLS), ordinary least squares with a corrected covariance matrix (OLS-C), and the maximum likelihood estimator (MLE). The MLE is identical to the generalized least squares (GLS) estimator if Ω is known and the errors are normally distributed. To define these estimators, let $X_c = [1 X^C_c X^{IC}_{ic} X^I_{ic}]$ be the $(N_c \times 4)$ matrix of explanatory variables for community c, and let X be the $(N_{TOT} \times 4)$ matrix formed by stacking all the community observations. Define similar vectors Y and β. The OLS estimator is

$$\hat{\beta} = (X'X)^{-1} X'Y.$$

The naive estimate of the covariance matrix generated by a standard statistical package would be

$$\mathrm{Cov}(\hat{\beta})_{OLS} = \sigma^2 (X'X)^{-1}. \tag{2}$$

The correct covariance matrix of the OLS estimator, however, would take account of the multilevel error structure. It takes on the following form:

$$\mathrm{Cov}(\hat{\beta})_{OLS-C} = (X'X)^{-1} X'\Omega X (X'X)^{-1}. \tag{3}$$

The MLE or GLS estimator is

$$\tilde{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

with covariance matrix

$$\mathrm{Cov}(\tilde{\beta}) = (X'\Omega^{-1}X)^{-1}. \tag{4}$$

To make comparisons among (2), (3), and (4), we assume that there is no constant in the model, all the Xs have mean zero, there are exactly N individuals in each community, and all the correlation in the correlated individual-level variable is due to the community-level variable (i.e., $\mathrm{Cov}(X^{IC}_{i'c}, X^{IC}_{ic}) = \tau^2$). After some tedious algebra (see Angeles and Mroz 2001), one can show that

$$\mathrm{Cov}(\hat{\beta})_{OLS} = \frac{\sigma^2}{N}\begin{bmatrix} \frac{1}{(1-\tau^2)} & -\frac{\tau}{(1-\tau^2)} & 0 \\ -\frac{\tau}{(1-\tau^2)} & \frac{1}{(1-\tau^2)} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathrm{Cov}(\hat{\beta})_{OLS-C} = \frac{\sigma^2}{N}\begin{bmatrix} \frac{1+\rho(N-1)(1-\tau^2)}{(1-\tau^2)} & -\frac{\tau}{(1-\tau^2)} & 0 \\ -\frac{\tau}{(1-\tau^2)} & \frac{1}{(1-\tau^2)} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$\mathrm{Cov}(\hat{\beta}) = \frac{\sigma^2}{N}\begin{bmatrix} \frac{[1+\rho(N-1)][1+\rho((N-2)-\tau^2(N-1))]}{[1+\rho(N-2)](1-\tau^2)} & \frac{-\tau(1-\rho)[1+\rho(N-1)]}{[1+\rho(N-2)](1-\tau^2)} & 0 \\ \frac{-\tau(1-\rho)[1+\rho(N-1)]}{[1+\rho(N-2)](1-\tau^2)} & \frac{(1-\rho)[1+\rho(N-1)]}{[1+\rho(N-2)](1-\tau^2)} & 0 \\ 0 & 0 & \frac{(1-\rho)[1+\rho(N-1)]}{1+\rho(N-2)} \end{bmatrix}$$

The variances of the estimators for the coefficients of the community-level variables are given by the (1, 1) elements of these matrices, the variances of the correlated individual-level variable coefficients are given by the (2, 2) elements, and the variances of the independent individual-level variable coefficients are given by the (3, 3) elements. A comparison of the three matrices yields several interesting results:

1. The naive OLS estimated variance for the effect of the community-level variable understates the true variance for $\rho \neq 0$, and the degree of understatement increases as ρ or N (individuals per community) increases. (Compare the (1, 1) elements of $\mathrm{Cov}(\hat{\beta})_{OLS}$ and $\mathrm{Cov}(\hat{\beta})_{OLS-C}$).

2. The OLS estimator for the coefficient of the community-level variable's effect is almost always inefficient relative to the MLE whenever $\rho \neq 0$. (Compare the (1, 1) elements of $\text{Cov}(\hat{\beta})_{OLS-C}$ and $\text{Cov}(\tilde{\beta})$.)
3. The important exception to the preceding result is when the community-level covariates are uncorrelated with all of the individual-level covariates (i.e., $\tau = 0$). In this instance, there is no efficiency loss from using OLS instead of MLE for the estimation of the impact of the community-level covariate. A pertinent example when there would be a zero correlation is for the case where particular treatments (e.g., facilities or programs) are assigned randomly across communities. [Compare the (1, 1) elements of $\text{Cov}(\hat{\beta})_{OLS-C}$ and $\text{Cov}(\tilde{\beta})$ evaluated at $\tau^2 = 0$.)
4. The standard errors reported by naive OLS for the impacts of both individual-level covariates are correct even when there is a multilevel structure. (Compare the (2, 2) and (3, 3) elements across $\text{Cov}(\hat{\beta})_{OLS}$ and $\text{Cov}(\hat{\beta})_{OLS-C}$.) This result, however, is an artifact of our simple example. If there were correlations of the individual-level covariates across individuals within the same community that were not completely captured by the observed community-level covariates, then the naive OLS standard errors would be too small.
5. There can be significant improvements in the precision of the estimators of the impacts of individual-level covariates by using MLE instead of OLS. (Compare the (2, 2) and (3, 3) elements between $\text{Cov}(\hat{\beta})_{OLS-C}$ and $\text{Cov}(\hat{\beta})$.)

GRAPHICAL ILLUSTRATIONS

The OLS estimator with corrected standard errors is much simpler computationally than the MLE. In addition, in more complex situations, it will produce unbiased estimators when the maximum likelihood estimators do not because it does not depend on auxiliary and often arbitrary assumptions such as homoskedastic normally distributed error components. Consequently, it is useful to see how the efficiency gain from MLE varies by ρ and N (the number of individuals per community).

We use the analytic formulas for the two covariance matrices to generate this comparison. Figures 1A and 1B show the standard deviations of the OLS estimators of the impacts of the three variables

vary by the number of observations per community and the level of correlation between the community variable and the individual-level variable. Figure 1A examines the case when the intraclass correlation, ρ, is 0.25, and Figure 1B examines the case for $\rho = 0.75$. Each of the four graphs within these figures refers to a different value of the correlation between the community-level explanatory variable and the correlated individual-level explanatory variable ($\tau = 0.00$, 0.25, 0.50, and 0.95). The horizontal axis measures the number of observations per community.

The vertical axis measures the standard deviation of the OLS estimator as a fraction of the standard deviation of the efficient maximum likelihood estimator.[1] This provides a measure of how much efficiency loss one can expect by using the less precise OLS estimator instead of the maximum likelihood estimator. The relative efficiencies for the estimators of the impacts of the two individual-level variables (indicated by the diamonds and plus signs) are identical analytically, and they do not depend on the degree of correlation between the individual-level variables and the community-level variable.

At the lower value of the intraclass correlation in Figure 1A, $\rho = 0.25$, there would be little efficiency gain from using the maximum likelihood estimator instead of the OLS estimator. Only when the correlation of the community-level variable and the individual-level variable is quite high (τ well above 0.50) is there any discernable efficiency gain for the estimator of the impact of the community-level variable from using maximum likelihood estimation. The efficiency gain for the estimation of the impact of the community-level variable initially increases as one adds more observations per community, but then it falls. But with $\rho = 0.25$, even when the correlation of the regressors is as high as 0.95, the standard error falls by less than 10 percent when using maximum likelihood instead of OLS.

Figure 1B examines the case where there is a high level of intraclass correlation, $\rho = 0.75$. As above, there is little efficiency gain from using maximum likelihood to estimate the impact of the community-level variable, unless the correlation of the community-level variable and the individual-level variable (τ) is quite high. But even when there can be substantial efficiency gains in estimating the community-level variable's impact by maximum likelihood, the gains

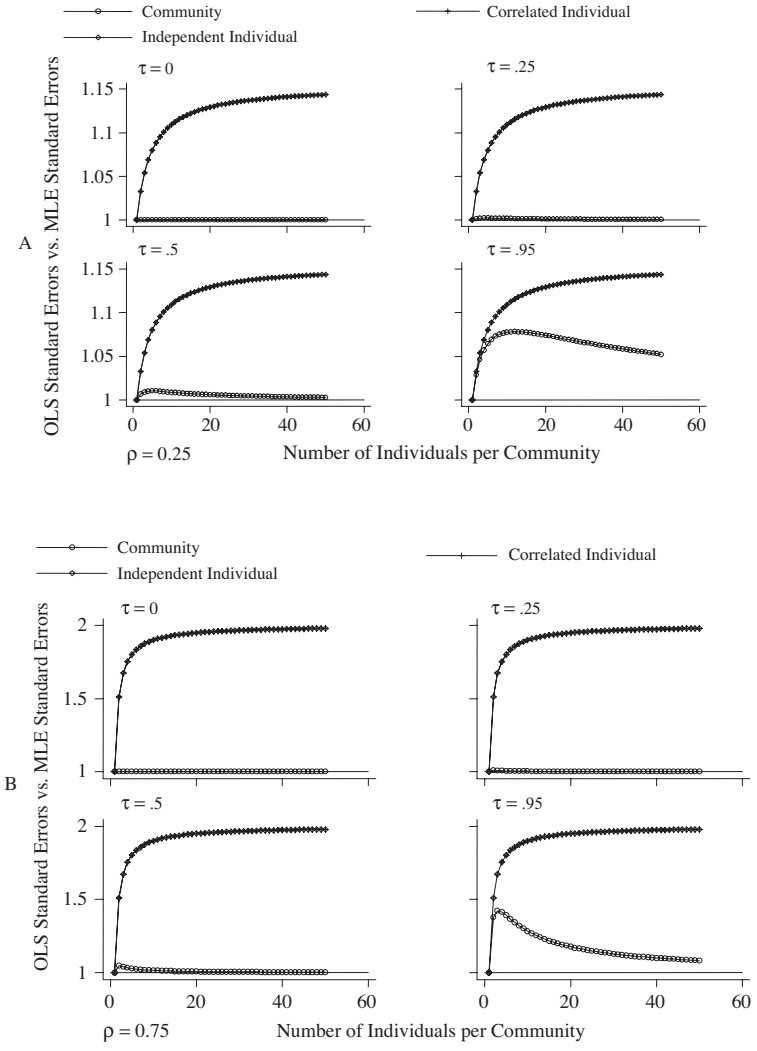

Figure 1: Standard Deviations of Ordinary Least Squares (OLS) Estimators as a Fraction of the Standard Deviations of the Maximum Likelihood Estimators as a Function of the Number of Individuals Per Community

NOTE: The graphs present relative OLS standard errors of the community-level, correlated individual-level, and independent individual-level coefficient estimators for intraclass correlation (ρ) of (A) 0.25 and (B) 0.75 and four different correlations of the individual-level and the community-level variables (τ).

diminish rapidly with increases in the number of observations per community.

The interactions among the number of observations per community, the intraclass correlation, and the correlation of the community-level regressor with the individual-level regressor appears to be the key determinants of efficiency gains from maximum likelihood estimation when estimating the impact of the community-level covariate. In Figures 2A and 2B, we examine this relationship in finer detail along the dimension of the correlation of the community-level variable and the individual-level variable. As in Figures 1A and 1B, Figure 2A is for $\rho = 0.25$, and Figure 2B is for $\rho = 0.75$. The graphs in each figure are for different numbers of individuals per community (NIPC = 2, 5, 25, and 50). The horizontal axis measures the level of correlation between the community-level and individual-level explanatory variables (τ). In Figures 2A and 2B, we only examine the relative efficiency for the estimators of the impact of the community-level variable.

For the estimation of the impact of the community-level variable, there appears to be almost no efficiency gain from using maximum likelihood estimators instead of OLS estimators for values of the regressor correlation being less than 0.50. For the moderate level of the intraclass correlation, 0.25, there never are efficiency gains over 15 percent for all values of the regressor correlation at 0.99 or lower. When the intraclass correlation is high, $\rho = 0.75$, there can be some substantial gains in efficiency, with the larger gains happening when there are only several individuals per community. These gains are quite small unless the regressor correlation is well over 0.50.

3. SIMULATION RESULTS

In section 2, we showed, for the estimation of the impacts of higher level covariates, that the efficiency gain from using the more complex MLE estimator instead of the OLS estimator was not large for a wide range of plausible correlation structures. We also showed that the naive OLS standard error estimators could seriously understate the true coefficient standard errors, which means that statistical hypothesis tests about the true impact of the explanatory variables would be incorrect. Equation (3) presents the correct formula for

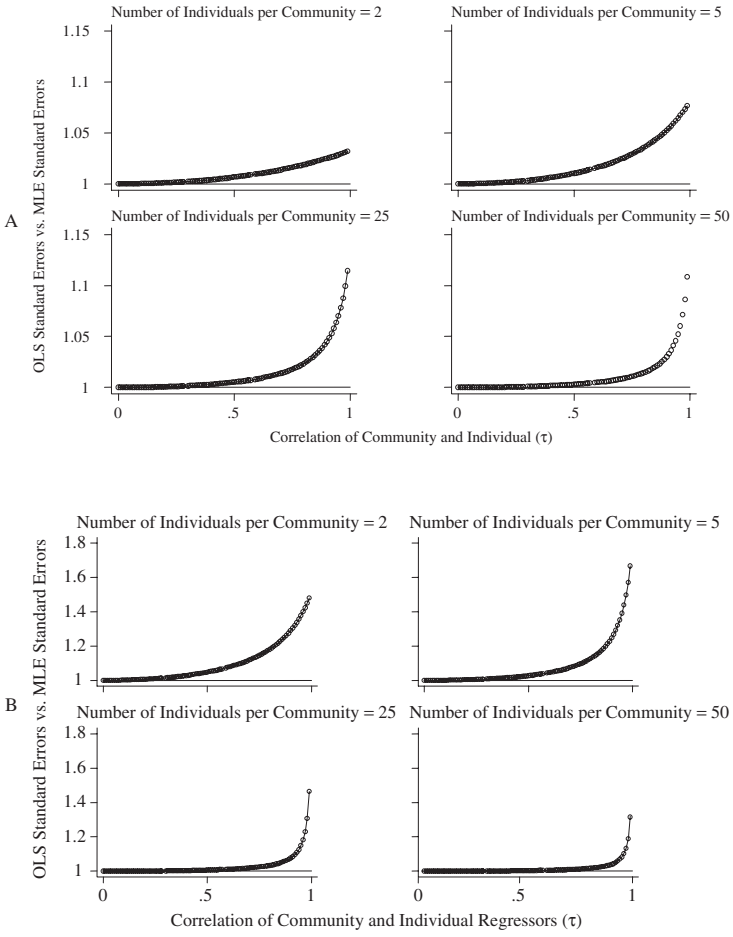

Figure 2: Standard Deviations of Ordinary Least Squares (OLS) Estimators as a Fraction of the Standard Deviations of the Maximum Likelihood Estimators of the Impact of the Community-Level Variable, as a Function of the Correlation of the Community and Individual-Level Regressors (τ)

NOTE: The graphs present relative OLS standard errors of the community-level coefficient estimators with an intraclass correlation (ρ) of (A) 0.25 and (B) 0.75 and four specifications of the number of individuals per community.

the covariance matrix for the OLS estimator (i.e., $\text{Cov}(\hat{\beta})_{OLS-C}$) given the homoskedastic error components structure specified above. Robust standard error estimators developed by Eicker (1963, 1967),

Huber (1967), and White (1980), however, provide for more general corrections of the standard errors. They allow for arbitrary forms of error correlation within communities and for error variances that are arbitrary functions of the explanatory variables. In addition, this more robust correction yields consistent standard errors even if the true regression coefficients are random, if the error structure is nonconstant or has more than two levels, or if the errors are heteroskedastic. Finally, these standard error estimators can be implemented using the cluster option in Stata, a widely available commercial statistical software package.

Unfortunately, the robust standard errors are only asymptotically valid (i.e., as the number of communities increases to infinity), and so it is important to see how well they work in sample sizes typically encountered in standard social science data sets. To do this, we use Monte Carlo simulation methods to assess their accuracy in comparison to the naive OLS standard errors and the standard errors for the MLE. We focus on statistical inferences about the community-level variable since it has not been studied in the past and since it is often the policy variable of interest. Results for individual-level variables are discussed in Angeles and Mroz (2001). We also present results for random intercept models for both a two-level error structure and a three-level error. The section ends with a set of simulation results for the random intercept/random slope model.

TWO-LEVEL ERROR SIMULATION RESULTS (RANDOM INTERCEPT MODEL)

Equation (1) specifies the data-generating process (DGP) for the Monte Carlo simulations. We set the true values of the constant term to zero and the other coefficients to 1. Even though the true constant is zero, we do estimate a constant in all models. The rest of the specification of the DGP is as follows:

1. All three independent variables (community, individual, and individual correlated with community) are generated from the normal distribution with mean zero and variance 1. We set the squared correlation coefficients between X_c^C, the community-level variable, and X_{ic}^{IC}, the correlated individual-level variable, to 0.5 (τ^2 in the notation of the previous section).

2. We adjust the variances of the community- and individual-level error components to generate 21 values for ρ (0, 0.05, 0.10, 0.15,. . ., 0.95, 1.00). Both error components are generated from independent normal distributions.
3. We focus on sample sizes with approximately 20,000 individual-level observations with 800 communities each containing between 1 and 50 individual observations, with a mean of 25 persons per community.[2] Angeles and Mroz (2001) also present results for 400 communities each containing exactly 50 individual-level observations, which are quite similar to the results presented here.
4. We choose an $R^2 = 0.10$. This restriction has almost no substantive impact on our comparisons of estimators.

All calculations were done using Stata. For each specification of the data-generating process, we draw 1,000 independent samples, each with 800 communities. Each community contains, on average, 25 individuals (standard deviation 9.5). We simulate community- and individual-level explanatory variables, community-level disturbances, and individual-level disturbances according to fixed, specific rules. For each of these 1,000 replications of the DGP, we estimate the model specified in equation (1) by OLS and by a maximum likelihood procedure that allows for the two-level hierarchical error structure. For the OLS estimates, we calculate estimates of the standard errors of the point estimates by using standard, naive OLS formulas and by using the robust (Eicker-Huber-White) formulas that adjust for possible heteroskedasticity and clustering within communities (i.e., $\rho \neq 0$). For the maximum likelihood procedure, we use Stata's report of the square root of the diagonal elements of the inverse of the Hessian matrix as the standard error estimator.

We assess the accuracy of alternative standard error estimators by determining whether hypothesis tests that use the standard error estimator yield accurate probabilities under the null hypothesis. In particular, a standard error estimator would be considered accurate if hypothesis tests that use this estimator reject a true null hypothesis with a frequency given by the specified size of a test. If one tests at the 5 percent level, for example, then one should reject correct null hypotheses 5 percent of the time. Otherwise, the standard error estimator does not allow one to carry out meaningful tests.

A standard, simple hypothesis test is often of the following form: $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. One typically undertakes a hypothesis test of this form by using a two-tailed test under the assumption that the estimator of β follows an approximate Student t or normal distribution. To carry out such a test, one sets a size of the test, α, to be the desired probability of rejecting the null hypothesis when the null hypothesis is actually true. Our evaluation of the accuracy of the estimator for the standard error of the estimate is an assessment of how closely the observed frequency of rejecting a true null hypothesis in the Monte Carlo experiments matches the specified size α.

In the Monte Carlo experiments, we know exactly the true parameter value (all βs equal to 1), so we can examine how frequently a true null hypothesis is rejected when we use various standard error estimators for the different point estimation procedures. We examine the size of the tests for three configurations of the test. If $\rho = 0$, all testing configurations should provide close to identical results.

Two configurations use the ordinary least squares point estimators for the parameter estimates. The first of these uses the standard error of the estimate as reported by the simple ordinary least squares procedure to evaluate the hypothesis test. This procedure corresponds to using a standard OLS procedure and using the "default" estimators of the standard errors that assume uncorrelated, homoskedastic disturbances, as defined in equation (2). This will usually provide biased tests when the intraclass correlation is nonzero for the DGPs examined here.

The second testing configuration uses a robust standard error estimator that accounts for the fact that there could be arbitrary correlations of disturbances within communities along with the simple OLS parameter point estimator (Eicker-Huber-White). It approximates the standard error formula given in equation (3) for this random intercept model with homoskedastic errors, but it will provide consistent standard error estimates under much more general conditions. The third testing configuration we examine uses the maximum likelihood estimator. We use the point estimates and the standard error estimates from the maximum likelihood procedure to carry out hypothesis tests. For each of these three testing configurations, we examine whether the standard error estimators used with the point estimators provide tests of the correct size.

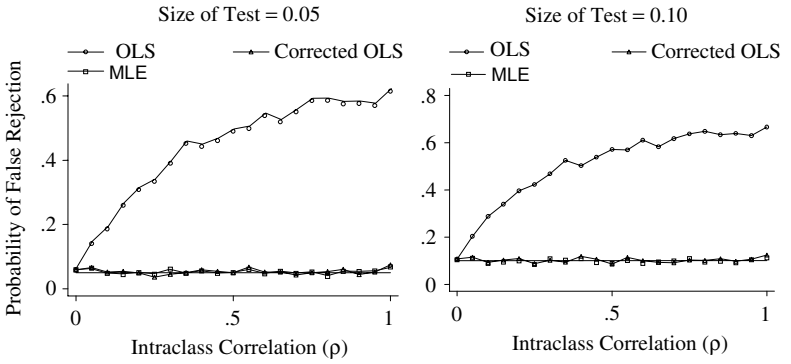

Figure 3: Performance of Standard Error Estimators of the Community-Level Variable

NOTE: The graphs present the probability of rejecting a true null hypothesis of the community-level variable as a function of the intraclass correlation (ρ) for two sizes of the test.

Figure 3 provides evidence on the probability that each of the testing procedures incorrectly rejects a true null hypothesis about the impact of the community-level variable (i.e., H_0: $\beta_C = 1$ vs. H_1: $\beta_C \neq 1$). The vertical axis measures the fraction of times out of the 1,000 replications that the hypothesis test rejects a true null hypothesis for a particular testing procedure. The horizontal axis measures the level of the intraclass correlation coefficient. The left side of the graph presents tests where the desired size of the test is 5 percent (0.05), and the right side of the graph presents 10 percent (0.10) results. The vertical scales vary within Figure 3.

When the intraclass correlation coefficient (ρ) equals 0.00, all three of the testing procedures yield approximately the correct size of 0.05. As the intraclass correlation rises, the procedure using the ordinary least squares point estimate with the simple OLS standard error estimate (labeled OLS, with circles) has an empirical probability of false rejection that greatly exceeds the specified 5 percent size. For all intraclass correlations above 0.10, the empirical size of this testing procedure exceeds 20 percent when one specifies a probability of false rejection of only 5 percent. With fewer communities and the same number of total individual-level observations, the empirical size from this approach can be much greater than 50 percent for a specified size of 5 percent.

From the same graph in Figure 3, tests that use the same OLS point estimate for the coefficient on the community-level variable but with the standard error adjusted to correct for arbitrary forms of correlation within communities (labeled "Corrected OLS," with triangles) yield approximately the correct size for all values of the intraclass correlation coefficient. Similarly, the tests based on the maximum likelihood point estimates and the corresponding maximum likelihood estimates of the standard errors of the estimates appear to have approximately the correct size. The 10 percent size results are very similar.

Given that there are estimators and testing procedures that have the correct size for the forms of multilevel models that we have examined, we can now examine the ability of these procedures to reject null hypotheses that are incorrect (i.e., statistical power). A graph displaying the probability of rejecting H_0: $\beta = \beta_p$ versus H_1: $\beta \neq \beta_p$ for a possible set of values β_p when the true parameter $\beta = \beta_0$ is one way of displaying the power of the test. For each null hypothesis examined, we present the fraction of times (out of 1,000 replications) that the testing procedure rejects each specified null hypothesis when test size is set at 0.05. Figure 4 contains graphs of the power functions corresponding to tests involving the coefficient of the community-level variable of the form $H_0 : \beta = \beta_p$ versus $H_1 : \beta \neq \beta_p$. The definition of the power function displayed here is the probability of rejecting the null hypothesis $\beta = \beta_p$ (against the alternative $\beta \neq \beta_p$) as a function of the value of β_p when the true parameter value is 1.0.[3]

Only two testing approaches had the correct size: OLS point estimates with standard errors adjusted for possible clustering of disturbances within communities and maximum likelihood point estimates and standard errors. These are displayed in Figure 4 as "Corrected OLS" and "MLE." When the intraclass correlation is 0.10 and the true value of $\beta_c = 1$, then one would reject the (false) null hypothesis that $\beta_c = 0.75$ (or $\beta_c = 1.25$)[4] about 76 percent of the time with OLS and 79 percent of the time with the maximum likelihood procedure. As the intraclass correlation rises to 0.25, the power to reject H_0: $\beta_c = 0.75$ (or $\beta_c = 1.25$) when the true value is 1 falls to 50 percent for the OLS-based procedure and 55 percent for the maximum likelihood procedure; when the intraclass correlation is a high 0.75, the

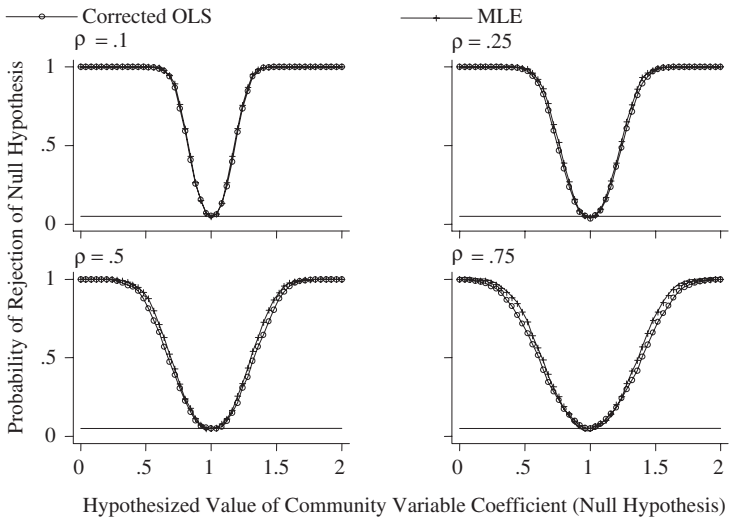

Figure 4: Power to Reject Null Hypotheses as a Function of the Intraclass Error Correlation

NOTE: The graphs present ordinary least squares estimators with Eicker-Huber-White standard errors and two-level maximum likelihood estimators, for the true value of the community-level coefficient equal to 1.0, size of test equal to 0.05, and four different intraclass correlations (ρ).

power for the same test is only 25 percent for the OLS-based test and 27 percent for the maximum likelihood–based test. In all cases examined here, the largest discrepancy in size between the two testing procedures is only 8 percentage points. In over half of the tests displayed in Figure 4, the probability of rejection using the maximum likelihood estimator is less than 1 percentage point larger than the probability of rejection from using the OLS point estimate with a robust standard error estimator.

Overall, for the coefficient on the community-level variable, one would conclude that there is little difference between the performance of tests based on the OLS point estimators with standard error adjustments for arbitrary within-community correlation and the tests based on the correctly specified maximum likelihood estimators. This should not be surprising; the comparisons of the asymptotic standard deviations of these two estimators discussed above

indicate that there would only be sizable differences if the intraclass correlation were exceptionally high with only a few observations per community.

A concern with Monte Carlo experiments is whether the results can be generalized since they could be contaminated by the choice of artificially generated data. This could limit the ability of these experiments to provide sound guidance in real-world situations. To partially remedy this concern, we use data from the National Longitudinal Study of Adolescent Health (Add Health) to calibrate an additional set of experiments for the two-level model. Add Health is a nationally representative survey of adolescents in Grades 7 through 12 in the United States in 1995 who had been followed with multiple interview waves into young adulthood. The study used a multistage, stratified, school-based cluster sampling design (see Harris et al. 2003).

To obtain the Monte Carlo data, we first estimate a simple model where we relate a 16-year-old's grade point average (GPA) to several individual-level characteristics and to several school-level characteristics. For the individual-level characteristics, we include a measure of the youth's aptitude and dummy variables for whether the youth was living with both biological parents, whether the youth's race was white, and whether the youth's father had graduated from high school and attended at least some college. For the school-level characteristics, we include the percentage of nonwhite students in the school and dummy variables for whether the school was small (less than 400 students), whether the school was large (more than 1,000 students), whether the school was a private school, and whether the school was located in an urban area. For our sample of 16-year-olds, there are 13,647 students with complete data for this analysis (Level 1 observations) in a total of 133 Level 2 units (schools). On average, schools contain 102 students, but this number ranges from 1 to 1,277 students per school. The estimated coefficients from this regression were all of the expected sign and were then used to calibrate the Monte Carlo for 51 values of ρ, the intraclass correlation.

To conserve space, we do not display these results. However, as in the earlier Monte Carlo models, as soon as the intraclass correlation exceeds zero, the OLS standard error estimators lead to excessive

false rejections. The maximum likelihood estimators appear to have an empirical size just slightly higher on average than the requested 5 percent size. The OLS point estimates in conjunction with the robust approximations to the standard errors have an empirical size of about 7 to 10 percent as opposed to 5 percent. However, when we used 400 Level 2 observations instead of 133, with the same nine explanatory variables, the empirical size closely matched the requested size just as it did for the simulated data. Thus, the results are supportive of those already presented, but they provide a note of caution. One should recognize that the adjusted standard error estimators could be a bit too small when there are a large number of Level 2 characteristics relative to the number of Level 2 observations. In the next section, we return to the use of simulated data in the examination of models with three levels.

THREE-LEVEL ERROR SIMULATION RESULTS (RANDOM INTERCEPT MODEL)

The analysis we reported on above indicated that one could obtain correct inferences from ordinary least squares model estimates that ignored the multilevel error structure, provided one adjusts the standard errors to ex post account for the within-community error correlation, but we only examined the performance of the standard error estimators when there were two levels in the analysis. In this section, we consider the performance of standard error estimators when the error term has up to three levels. Extending the descriptive notation used above, the three levels in this model are the individual level, the family level, and the community level. The DGPs we consider have the individual-level outcome being influenced by one explanatory variable from each level. Let X_c^C be the community-level explanatory variable ($c = 1, 2, \ldots, C$), X_{fc}^F be the family-level variable ($f = 1, 2, \ldots, F_c$), and X_{ifc}^{IC} be the individual-level variable ($i = 1, 2, \ldots, N_{fc}$). We allow these explanatory variables to be correlated within communities and families, and we set $\text{Cor}[X_c^C, X_{fc}^F] = \text{Cor}[X_{fc}^F, X_{ifc}^{IC}] = 0.5$ and $\text{Cor}[X_c^C, X_{ifc}^{IC}] = 0.667$. We permit there to be unobservable determinants of the individual-level outcomes associated with each of these three levels.

The linear regression model we examine takes the following form:

$$Y_{ifc} = \beta_0 + \beta_1 X_c^C + \beta_2 X_{fc}^F + \beta_3 X_{ifc}^{IC} + \mu_c + \lambda_{fc} + \epsilon_{ifc}, \quad (5)$$

where μ_c gives rise to the within–Level 3 error correlation (community), λ_{fc} gives rise to Level 2 error correlation (family), and ϵ_{ifc} is the Level 1 error term (individual). The μ_c are independent across different communities (Level 3 observations) with variance σ_μ^2, the λ_{fc} are independent across families (Level 2 observations) with variance σ_λ^2, and the ϵ_{ifc} are independent across all individuals with variance σ_ϵ^2. If σ^2 is defined to be the sum of the three variances, we can define $\rho_C = \sigma_\mu^2/\sigma^2$ and $\rho_F = \sigma_\lambda^2/\sigma^2$ as the proportions of the overall error variance due to Level 3 (community) and Level 2 (family) unobserved factors. These calculations assume that the covariance among the community-, family-, and individual-level error terms is equal to zero.

In the simulations, the community and family error components are distributed as independent normal random variables. For each of 21 values of ρ, we consider three sets of error correlations: $\rho_C = \rho_F = \rho$; $\rho_C = 0$ with $\rho_F = \rho$; and $\rho_C = \rho$ with $\rho_F = 0$. We set the three regression coefficients equal to 1.0, and we specify four Level 1 units (individuals) within each of 25 Level 2 units (families) for each of 200 Level 3 units (communities), for a total of 20,000 individual-level observations. We set the R^2 to always equal 0.10.

Our primary concern here is how one can carry out unbiased tests on the coefficient of the community-level variable in these three-level models. Figures 5, 6, and 7 contain pertinent information about the size performance of various estimators of standard errors for different configurations of the multilevel error correlations. The left-hand side graphs examine various ways to estimate standard errors for the OLS point estimators. We consider three standard error estimators for these OLS estimates. The first is the naive standard errors as reported by standard OLS procedures assuming completely uncorrelated disturbances (labeled "OLS"). The second is a robust standard error estimator assuming that only observations within the second level are correlated (labeled "OLS Family Correction"). These standard error estimators would be appropriate, for example, if there could be nonzero error correlation among individuals within the same family

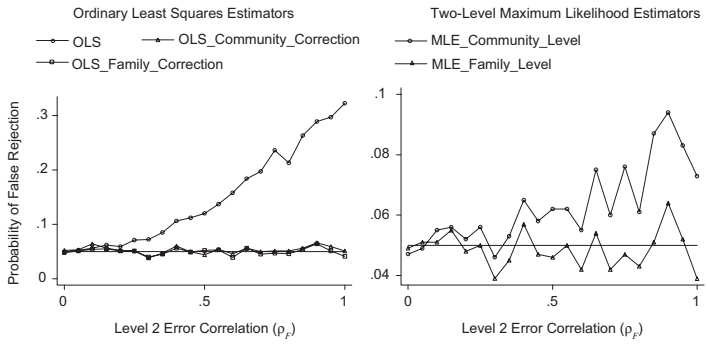

Figure 5: Performance of Standard Error Estimators for Three-Level Models With Only Level 2 Error Correlation

NOTE: The graphs present the probability of rejecting a true null hypothesis as a function of the Level 2 error correlation for size of test .05.

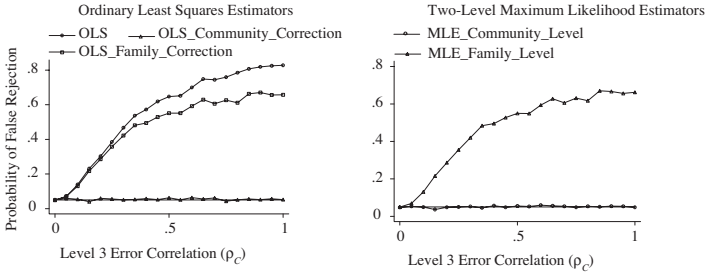

Figure 6: Performance of Standard Error Estimators for Three-Level Models With Only Level 3 Error Correlation

NOTE: The graphs present the probability of rejecting a true null hypothesis as a function of the Level 3 error correlation for size of test .05.

($\rho_F \neq 0$) but no correlation of disturbances across families living within the same community ($\rho_C = 0$). The third standard error estimator is similar to the second, except that it allows for possible error correlation at up to the third level among Level 1 units (e.g., error correlation among families and individuals living within the same community, labeled "OLS Community Correction").

The right-hand side graphs are based on maximum likelihood point and standard error estimators that naively assume a two-level

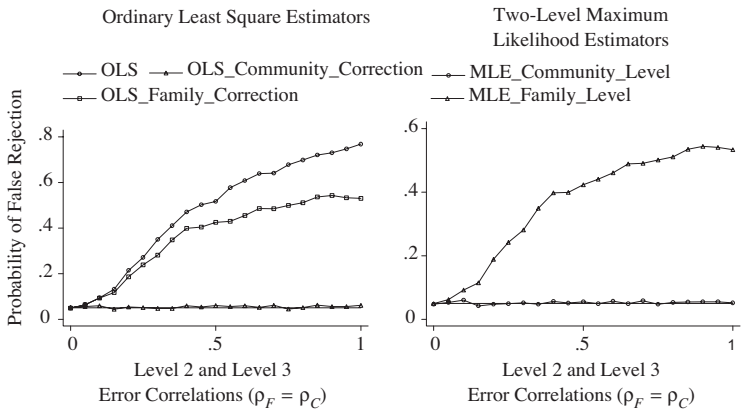

Figure 7: Performance of Standard Error Estimators for Three-Level Models With Equal Level 2 and Level 3 Error Correlations

error hierarchy.[5] The first assumes that all Level 1 observations are equally correlated within the Level 3 units (labeled "MLE Community Level"). This would be the case, for example, if community-level unobserved factors could influence an individual's outcomes ($\rho_C \neq 0$), but there are no unobserved family-level factors influencing the individual-level outcome ($\rho_F = 0$). The second set of maximum likelihood point and standard error estimators again assumes that there is only error correlation among Level 1 units within the same Level 2 unit (e.g., only disturbances for individuals within the same family are correlated, i.e., $\rho_C = 0$ and $\rho_F \neq 0$, labeled "MLE Family Level").

The graphs display the empirical Type I error (size) for null hypotheses of the form $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$, where β_0 is the true value of the parameter in the DGP (i.e., 1.00 for all parameters examined), as a function of the intraclass correlation coefficient among individuals at Level 1 within each Level 2 unit.[6] Each of these tests take place at a 5 percent level, and we carry out each test for each of the 1,000 Monte Carlo replications. As in the analysis of standard error estimators in the simpler models, a point on the graph represents the fraction of times the true null hypothesis is rejected using that particular point and standard error estimator at the specified level of the intraclass correlation. An accurate standard

error estimator for a particular point estimator would exhibit a straight, horizontal line at 0.05 for all values of the intraclass error correlation. Note that the vertical scales vary across graphs within these figures.[7]

Figure 5 considers the case where there is only error correlation among Level 1 units within the same Level 2 unit (e.g., only error correlation among individuals within the same family). In particular, $\rho_C = 0$, while $\rho_F \neq 0$. We see that the naive standard error estimator for the OLS point estimator performs quite poorly (''OLS''). The empirical size exceeds twice the specified size even for some intraclass correlations below 0.50, with the empirical size rising to about 0.30 at the highest levels of intraclass correlation. Both of the robust standard error estimators yield tests of the correct size. It is important to recognize that for these robust standard error estimators to perform correctly, one only needs to specify the highest level at which there could be error correlations.[8] Hence, the estimator allowing there to be correlations among all individuals within the same community (''OLS Community Correction'') provides unbiased hypothesis tests, even though there is no community-level (Level 3) error correlation. Its assumption of clustering up to as high as the community level (Level 3) incorporates as a special case clustering only within families (Level 2). For this standard error estimator to perform well, there need not be the same form of error correlation for all observations within the level specified as being the highest level with correlation.

The MLE does not generalize this way. The right-hand graph in Figure 5 indicates that the MLE that models the within Level 2 (family) correlation does provide unbiased tests; this estimation procedure coincides with this specification of the DGP (only Level 2 correlation). The MLE assuming only the higher level, community-level error correlations, however, appears to be increasingly biased as the level of the intraclass correlation rises; this estimation method does not contain Figure 5's DGP as a special case. However, even at the highest levels of ρ, the incorrectly specified maximum likelihood estimator provides tests that reject at most about 8 percent of the time when the requested size is 5 percent.[9]

Figure 6 provides the same information as Figure 5, but the DGP used for Figure 6 has all of the error correlation taking place at

the community (third) level. Here $\rho_C > 0$, while $\rho_F = 0$. After controlling for the Level 3 correlation (e.g., community), there is no additional correlation among Level 1 units (e.g., individuals) within the same Level 2 unit (e.g., families). The performances of the testing procedures in this instance are somewhat different than those discussed for Figure 5. The naive OLS standard error estimator continues to provide biased tests. The "robust" standard error estimator with only family-level (Level 2) error correlation and the maximum likelihood procedure that allows for error correlation only at the family level (Level 2) now provide quite biased tests; this is because these procedures do not recognize the Level 3 error correlation. The two-level maximum likelihood model that allows there to be error correlation at the community level (Level 3), not surprisingly, provides unbiased tests for the impact of the community-level covariate because it is correctly specified. Tests using the OLS point estimates along with the robust standard error estimators allowing for up to Level 3 error correlation also provide unbiased tests for the impact of the community-level covariate.

Figure 7 presents the perhaps more realistic case when there are error correlations at Level 2 (within the family) and at Level 3 (within the community). The OLS robust standard error estimator with control for community error (Level 3) correlations and the maximum likelihood approach that allows for error correlation at the community level (Level 3) provide unbiased tests even though error correlation is present at the family level also. It is surprising that this maximum likelihood estimator performs correctly here. It performed somewhat poorly when there was only family-level (Level 2) error correlation as in Figure 5, while here there is family error correlation as well as community error correlation. For the community-level variable effect, any approach that fails to recognize that there can be correlated errors at the community level provides biased hypothesis tests.

THREE-LEVEL ERROR SIMULATION RESULTS (RANDOM INTERCEPT/RANDOM SLOPE MODEL)

OLS with corrected standard errors has done extremely well in the simulations reported thus far. As stated above, the Eicker-Huber-White correction to the OLS standard errors can also correct

the standard errors for general types of heteroskedasticity, and so it is of interest to see how well it works for the extension of the random intercept model to the widely used random intercept/random slope model (for textbook discussions of this model, see Goldstein 1995, 2003; Bryk and Raudenbush 1992; Raudenbush and Bryk 2002). We modify equation (5) as follows:

$$Y_{ifc} = \beta_{0ifc} + \beta_{1ifc} X_c^C + \beta_{2ifc} X_{fc}^F + \beta_{3ifc} X_{ifc}^{IC}, \tag{6}$$

where the random coefficients have the following specification:

$$\beta_{jifc} = \beta_j + \rho\sqrt{.5}(\mu_{jc} + \lambda_{jfc}) + \sqrt{1-\rho^2}\epsilon_{jifc}. \tag{7}$$

This is a random coefficient specification in which the intercept and the slope terms are each functions of community-, family-, and individual-level "errors." If only β_0(the intercept) is random, then equation (6) reduces to equation (5). In the standard textbook random coefficient model with two levels, the slope and intercepts are only functions of the community-level error, which would involve dropping λ and ϵ from equation (7). Thus, our specification is more general and allows for more complicated forms of heteroskedasticity and correlation.

The simulations use equations (6) and (7) for the DGP and vary ρ between zero and 1 in the same manner as was done for the three-level error model with a random intercept only. Comparisons for the five estimators discussed in the previous section are presented in Figure 8. The left-side panel of Figure 8 displays the results for the OLS-based estimators: OLS with naive standard errors, OLS with robust standard errors with correction at the family level, and OLS with robust standard errors with correction at the community level. It is clear from the graph that correcting the standard errors at only the outmost level leads to tests with the correct size. The performance of the naive OLS estimator deteriorates at the most rapid rate as error correlation increases. OLS with robust standard error estimators that recognize only family (Level 2) correlations have quadruple the nominal size of 5 percent when $\rho \geq 0.4$. As one would expect, neither of the simple, two-level MLE estimators performs well in this situation since neither likelihood function takes into account the heteroskedastic nature of the disturbance term.

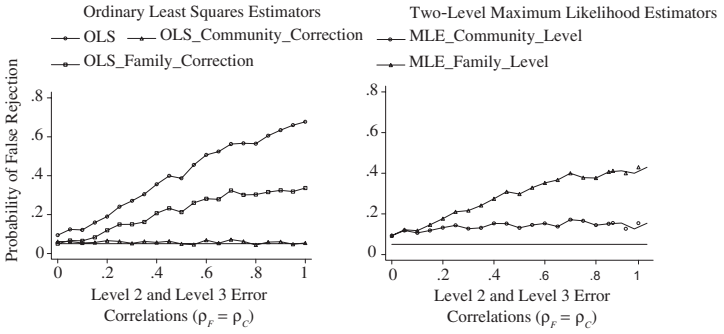

Figure 8: Performance of Standard Error Estimators for Three-Level Models When the True Data-Generating Process Contains Random Slopes and Intercepts Correlated at Levels 1 and 2

The main conclusion about the performance of standard error estimators when there are three-level models for both the random intercept and random slope models is that it is most important to control for the error correlation at the highest possible level at which it might exist. For the OLS estimates with robust standard errors (Eicker-Huber-White), it does not matter whether one "over-controls" and allows for possible error correlations at a higher level than is actually the case (see Figure 5). For these robust standard error estimators, as long as the highest level of actual error correlation is nested within the level specified in the estimation, hypothesis tests will be unbiased. For the two-level maximum likelihood estimator, it is more important to specify exactly the level at which the error correlation takes place. But, if one were going to use two-level maximum likelihood estimators in the presence of three-level error components, these results suggest that it would be better to assume that all error correlation takes place at the highest level (e.g., community level).

4. CONTINUOUS DEPENDENT VARIABLE APPLICATION: THE DETERMINANTS OF CHILD'S WEIGHT

To illustrate the methods presented above, we estimate a reduced-form model of the determinants of child weight using data from

the Cebu Longitudinal Health and Nutrition Survey (CLHNS). This data set provides an excellent illustration of the methods since there are data at three levels: community, individual, and time-varying individual. The Level 1 units are bimonthly observations on an infant's weight from birth to two years of age (up to a maximum of 13 per infant). A Level 2 unit is the child (or, equivalently, its mother or its family), and a Level 3 unit is the Barangay (community) where the family resides. There are 33 Barangays in this study, with 3,327 infants, and a total of 34,293 bimonthly recordings of the children's weights. Table 1 contains summary statistics, and Guilkey et al. (1989) provide a detailed description of the data. This data set is about 50 percent larger than the Monte Carlo data set and has more observations per community (Level 3) unit. As a result, a priori, we would expect the naive OLS standard errors to be badly biased.

Table 2 reports six sets of estimation results for the effect of community-, family-, and individual-level covariates on an infant's weight in kilograms for up to 13 points in time from birth to two years of age:

1. OLS with naive standard errors,
2. OLS with robust (Eicker-Huber-White) standard errors with correction at the person level,
3. OLS with robust (Eicker-Huber-White) standard errors with correction at the Barangay level,
4. two-level MLE at the person level,
5. two-level MLE at the Barangay level, and
6. three-level GLS with Barangay and person levels, with standard errors from the maximum likelihood procedure and robust (Eicher-Huber-White) standard errors assuming clustering at the community level.[10]

All estimation methods provide consistent point estimates of the parameters, but in general, only Method 3 and Method 6 with the robust standard errors will provide correct size hypothesis tests. We used the MLwiN program (Rasbash et al. 2000) to carry out an iterated feasible GLS estimation for the three-level model since three-level MLE cannot be easily estimated using Stata. Iterated feasible GLS is asymptotically equivalent to MLE under the assumption of error term normality, and the MLwiN procedure yielded nearly

TABLE 1: Descriptive Statistics

Cebu: Child's Weight		Philippines: Use of Family Planning	
Variable	Mean (*SD*)	Variable	Mean (*SD*)
		Community variables	
Price of formula ($n = 429$)	4.4553 (1.4513)	Public province health expenditure ($n = 75$)	0.4353 (0.4716)
Price of corn ($n = 429$)	5.9831 (1.5006)	Public municipal health expenditure ($n = 484$)	0.5094 (0.3288)
Urban Barangay ($n = 33$)	0.5151 (0.5000)		
		Individual variables	
Child's weight ($n = 34{,}297$)	7.3889 (2.1513)	Use of family planning ($n = 7{,}492$)	0.6884 (0.4631)
Mother's age ($n = 3{,}073$)	26.0680 (5.9915)	Ages 15 to 19 ($n = 7{,}492$)	0.0119 (0.1083)
Mother's education ($n = 3{,}073$)	7.1132 (3.3115)	Ages 20 to 24 ($n = 7{,}492$)	0.1481 (0.3552)
Mother's height ($n = 3{,}073$)	150.5611 (5.0042)	Ages 25 to 29 ($n = 7{,}492$)	0.2714 (0.4448)
Child is a boy ($n = 3{,}073$)	0.5307 (0.4991)	Ages 30 to 34 ($n = 7{,}492$)	0.2474 (0.4316)
		Ages 35 to 39 ($n = 7{,}492$)	0.1817 (0.3856)
		Ages 40 to 44 ($n = 7{,}492$)	0.0874 (0.2825)
		Ages 45 to 49 ($n = 7{,}492$)	0.0539 (0.2257)
		Education 0 to 5 years ($n = 7{,}492$)	0.1517 (0.3588)
		Education 6 years ($n = 7{,}492$)	0.1969 (0.3977)
		Education 7 to 10 years ($n = 7{,}492$)	0.3812 (0.4857)
		Education 11+ years ($n = 7{,}492$)	0.2603 (0.4388)
		Partner's education 0 to 5 years ($n = 7{,}492$)	0.1931 (0.3948)
		Partner's education 6 years ($n = 7{,}492$)	0.1771 (0.3818)
		Partner's education 7 to 10 years ($n = 7{,}492$)	0.3509 (0.4773)
		Partner's education 11+ years ($n = 7{,}492$)	0.2619 (0.4397)
		Respondent lives in urban area ($n = 7{,}492$)	0.3990 (0.4897)
		Asset index ($n = 7{,}492$)	2.1307 (2.3751)
		Religion is Catholic ($n = 7{,}492$)	0.7623 (0.4257)

identical estimates and standard errors as the maximum likelihood estimation procedure we developed for this study. As a check on the possible differences between MLwiN, Stata, and our programs, we estimated Methods 4 and 5 (MLE with only two levels) using each of the three methods. Nearly identical results were obtained for the three estimation procedures.

The top half of Table 2 presents results for three Barangay-level variables: the price of formula, the price of corn, and whether or not the Barangay is urban; the columns are labeled to refer, in order, to the estimation procedures just outlined. Looking at the first three columns of the table for these three variables, we see that the point estimates of the coefficients are identical, as they should be. However, we see that the naive OLS standard errors appear to be badly biased, and the OLS standard errors with the Eicker-Huber-White correction at the individual level (labeled column 2) understate the standard errors relative to those using the correction at the Barangay (community) level—the outmost level.

The last three columns of Table 2 (marked as columns 4, 5, and 6) present MLE results. The corrections to the standard errors we observed in the first three columns (columns 1, 2, and 3) suggest that there may be important correlations at both Level 2 and Level 3 for these data. This implies that columns 4 and 5 could be incorrectly specified, as column 4 ignores the Barangay-level error correlation and column 5 ignores the Level 2 (child) error correlation. The bottom of the table reports the estimated error component variances, and we see that both the Barangay-level variance and the fixed individual-level variance are highly significantly different from zero in the three-level model, regardless of which standard error estimator one uses. The fraction of error variance due to child-level unobservables, from column 6, is 0.66, while the fraction of variance due to Barangay is only 0.02. The analytical results presented in section 2 suggest that such a combination of a large intraclass correlation with a small number of observations per Level 2 unit (at most 13 observations per child) is a case where there could be substantial efficiency gains from using MLE instead of OLS with standard error corrections.

The last column in Table 2 (column 6) reports two sets of standard errors for the maximum likelihood, three-level estimation model. The first set is what one obtains directly from the maximum

TABLE 2: The Determinants of Child's Weight in the Philippines (Standard Errors in Parentheses)

Variable	1. OLS	2. OLS Individual Correction	3. OLS Community Correction	4. MLE Individual Only	5. MLE Community Only	6. MLE Both (MLE SE) (Robust SE)
			Community variables			
Price of formula	0.0108 (0.0052)	0.0108 (0.0083)	0.0108 (0.0109)	−0.0031 (0.0033)	−0.0031 (0.0057)	−0.0035 (0.0033) (0.0052)
Price of corn	−0.0138 (0.0060)	−0.0138 (0.0070)	−0.0138 (0.0077)	−0.0010 (0.0035)	−0.0084 (0.0061)	−0.0009 (0.0035) (0.0052)
Urban	−0.0301 (0.0122)	−0.0301 (0.0351)	−0.0301 (0.0635)	−0.0371 (0.0342)	−0.0799 (0.0459)	−0.06444 (0.0486) (0.0526)
			Individual variables			
Mother's age	−0.0028 (0.0008)	−0.0028 (0.0025)	−0.0028 (0.0023)	−0.0029 (0.0023)	−0.0030 (0.0008)	−0.0032 (0.0023) (0.0021)
Mother's years of education	0.0570 (0.0016)	0.0570 (0.0047)	0.0570 (0.0074)	0.0531 (0.0045)	0.0591 (0.0016)	0.0544 (0.0046) (0.0066)
Mother's height	0.0406 (0.0010)	0.0406 (0.0029)	0.0406 (0.0031)	0.0396 (0.0029)	0.0399 (0.0010)	0.0391 (0.0028) (0.0029)
Child is a boy	0.5127 (0.0099)	0.5127 (0.0292)	0.5127 (0.0252)	0.4782 (0.0280)	0.5062 (0.0099)	0.4754 (0.0279) (0.0250)
Age 2 months	1.8941 (0.0239)	1.8941 (0.0111)	1.8941 (0.0157)	1.8944 (0.0134)	1.8983 (0.0237)	1.8943 (0.0134) (0.0149)
Age 4 months	3.1415 (0.0243)	3.1415 (0.0154)	3.1415 (0.0213)	3.1393 (0.0137)	3.1506 (0.0242)	3.1394 (0.0137) (0.0189)
Age 6 months	3.8753 (0.0250)	3.8753 (0.0184)	3.8753 (0.0261)	3.8661 (0.0141)	3.8860 (0.0248)	3.8662 (0.0141) (0.0230)
Age 8 months	4.3267 (0.0259)	4.3267 (0.0236)	4.3267 (0.0285)	4.3120 (0.0146)	4.3388 (0.0257)	4.3122 (0.0146) (0.0249)
Age 10 months	4.6715 (0.0270)	4.6715 (0.0236)	4.6715 (0.0327)	4.6472 (0.0153)	4.6841 (0.0269)	4.6474 (0.0153) (0.0279)
Age 12 months	4.9814 (0.0283)	4.9814 (0.0261)	4.9814 (0.0371)	4.9511 (0.0161)	4.9952 (0.0282)	4.9513 (0.0161) (0.0311)
Age 14 months	5.2641 (0.0294)	5.2641 (0.0282)	5.2641 (0.0438)	5.2284 (0.0167)	5.2783 (0.0293)	5.2287 (0.0167) (0.0377)

			Community variables			
Age 16 months	5.5459 (0.0301)	5.5459 (0.0293)	5.5459 (0.0478)	5.5109 (0.0171)	5.5631 (0.0300)	5.5112 (0.0171) (0.0399)
Age 18 months	5.8302 (0.0301)	5.8302 (0.0301)	5.8302 (0.0483)	5.7965 (0.0172)	5.8487 (0.0300)	5.7969 (0.0172) (0.0408)
Age 20 months	6.1396 (0.0296)	6.1396 (0.0304)	6.1396 (0.0509)	6.1085 (0.0169)	6.1608 (0.0296)	6.1090 (0.0169) (0.0440)
Age 22 months	6.4593 (0.0281)	6.4593 (0.0286)	6.4593 (0.0468)	6.4327 (0.0160)	6.4776 (0.0280)	6.4331 (0.0160) (0.0411)
Age 24 months	6.8048 (0.0268)	6.8048 (0.0269)	6.8048 (0.0443)	6.7748 (0.0152)	6.8204 (0.0267)	6.7750 (0.0152) (0.0415)
Constant	−3.7223 (0.15414)	−3.7223 (0.4454)	−3.7223 (0.4221)	−3.4903 (0.4269)	−3.5361 (0.1575)	−3.3931 (0.4266) (0.4201)
			Variances			
σ^2 columns 1, 2, 3	0.83 (0.00)	0.83 (0.00)	0.83 (0.00)			
σ^2_ζ columns 4, 5, 6				0.25 (0.00)	0.82 (0.00)	0.25 (0.00) (0.01)
σ^2_λ				0.5666 (0.0153)		0.5585 (0.0151) (0.0237)
σ^2_μ					0.0156 (0.0042)	0.0083 (0.0039) (0.00278)

NOTE: OLS = ordinary least squares; MLE = maximum likelihood estimation; SE = standard error.

likelihood procedure, and the second comes from robust (sandwich in MLwiN) formulas that allow for heteroskedasticity and arbitrary error correlations within each community. In general, robust standard errors should be valid even when data do not conform to the assumptions of the maximum likelihood estimator. The fact that the robust standard error estimates in column 6 are often quite different from those derived directly from the maximum likelihood procedure suggests that there could be more general forms of error correlations within clusters or that some of the errors could be heteroskedastic; consequently, the first set of standard errors reported in column 6 is most likely invalid. While not reported in this table, we used similar robust standard errors estimators for columns 4 and 5 and found large changes in the standard errors from using these more general procedures. This suggests that the standard errors reported in columns 4 and 5 are incorrect. We also compared bootstrap standard errors, obtained by sampling with replacement at the Barangay level (Level 3), for the models in columns 3, 4, 5, and 6, and we found that they were almost always in close agreement with those from the robust (Eicker-Huber-White) standard error formulas. Given the lack of robustness for the standard errors reported by the maximum likelihood procedures, it would seem prudent for researchers to use either Eicker-Huber-White standard error estimators or bootstrap procedures with sampling at the highest level where correlation exists for carrying out hypothesis tests with the maximum likelihood point estimators.

Table 2 reveals similar impacts of the various estimation procedures on the estimates for the impact of the child's gender and the mother's age at birth, education, and height on the child's weight. Note that these maternal variables, unlike the community-level variables, do not have any time-series variation that is not captured by the child age variables. Naive OLS standard errors suggest that the mother's age is a significant determinant of the child's weight; its significance disappears after controlling appropriately for the multilevel error structure. Similarly, the naive OLS standard errors of the gender, education, and height effects understate the corrected standard errors by factors of 2 to 4. The three-level MLE model does appear to provide somewhat more accurate estimators of these effects than the OLS model, with the largest efficiency gains being for the Level 1 indicators of the child's age.

This example clearly reflects the statistical and Monte Carlo evidence presented in the previous two sections. It demonstrates the need to correct the OLS standard errors using the outmost level of clustering if one wants to make correct inferences.

5. EXTENSION TO THE PROBIT MODEL WITH AN APPLICATION TO USE OF FAMILY PLANNING IN THE PHILIPPINES

Probit and logit models are typically used for models with dichotomous dependent variables, and extensions to multilevel models have been developed for both methods. For the two-level model, the MLE estimator is the same as the panel data estimator. See Wooldridge (2002) for a textbook presentation of the methods. For work involving extensions of the logit model, see Pendergast et al. (1996) for a review of estimation methods; Guo and Zhao (2000) provide a general review of the literature. Unlike the case of the continuous outcome models, it is not possible to make simple, theoretical (analytic) statements about the relative performance of estimators. Instead, in this section, we point out some important interpretation issues that arise when one considers multilevel models with discrete outcomes and present an example illustrating the efficiency gains and properties of standard error estimators from using models that control for multilevel error structures.

Probit models are slightly more convenient for multilevel models because they depend on an underlying normal error distribution. Sums of normal random variables from different level units will remain in the class of a normal distribution, and models controlling for possible correlations across different levels can fit into a common and internally consistent statistical model. The probit extension involves modifying equation (1) as follows:

$$Y^*_{ic} = \beta_0 + \beta_1 X^C_c + \beta_2 X^{IC}_{ic} + \beta_2 X^I_{ic} + \mu_c + \epsilon_{ic}, \qquad (8)$$

where all terms are defined as above except that Y^*_{ic} is an unobserved latent variable. The observed variable, Y_{ic}, is an indicator variable that takes on the value 1 when the latent variable is positive and zero otherwise. As in the continuous dependent variable case, we assume

that the two-error components are independent. What makes this a probit model instead of a logit or other binary outcome model is the assumption that both μ_c and ϵ_{ic} follow normal distributions. We define $\sigma^2 = \sigma_\epsilon^2 + \sigma_\mu^2$ and $\rho = \sigma_\mu^2/\sigma^2$, where ρ is the fraction of the total error variance due to the community-level component of the error term.

The difference from the continuous outcome case is that we only observe the sign of the dependent variable Y_{ic}^*; we must therefore impose a normalization. The normalization used in all simple probit procedures is $\sigma^2 = 1$ (Heckman 1981). In more complex models, some computer packages instead impose $\sigma_\epsilon^2 = 1$. Because of this need to make an arbitrary normalization, only ratios of coefficients (i.e., relative effects) and significance levels can usually be identified.

Robinson (1982) has shown that simple probit applied to (8) will consistently estimate the model's coefficients. Just as in the continuous case, however, the coefficient standard errors from the simple estimation will be incorrect. Robust (Eicker-Huber-White) standard errors will be asymptotically valid as long as one allows for error correlation at the highest level. A Monte Carlo study (Guilkey and Murphy 1993) obtained results that were similar to the results of Angeles and Mroz (2001) for the continuous dependent variable case: Probit with Eicker-Huber-White standard errors performed quite well, while the naive probit model produced standard errors that were badly biased. In addition, there was only a small efficiency gain from using the more complicated MLE estimator in their experiments. Note, however, that their Monte Carlo study was designed to examine longitudinal data models with many individuals and a small number of observations per individual (either 2, 5, or 10 observations); it was not designed to examine the community-level multilevel model with more than 10 observations per community.

Hardin (1996) confirmed the results of the Guilkey and Murphy (1993) study but also only examined cases with up to 20 observations per community. Both studies stressed the need to increase the number of Hermite points used in the numerical evaluation of the likelihood function as the number of individuals per community increased. This is necessary because one must calculate a joint probability of the outcome for everyone within the same community, and accuracy problems arise as the number of observations within a community gets large. Borjas and Sueyoshi (1994) provide an alternative

two-step estimator that may be useful if the Hermite approximation does not work well because of large sample sizes within communities. This was not a problem in our numerical example.

To discuss the MLE for this two-level probit model and to see the effect of the normalization on estimated coefficients, we rewrite equation (8) with an arbitrary normalization as follows:

$$\frac{Y^*_{ic}}{\sigma_\eta} = \left[\frac{\beta_0}{\sigma_\eta}\right] + \left[\frac{\beta_1}{\sigma_\eta}\right] X^C_c + \left[\frac{\beta_2}{\sigma_\eta}\right] X^{IC}_{ic} + \left[\frac{\beta_3}{\sigma_\eta}\right] X^I_{ic} + \left[\left(\frac{\rho}{1-\rho}\right)^{1/2} \frac{\sigma_\epsilon}{\sigma_\mu} \frac{1}{\sigma_\eta}\right] \mu_c + \left[\frac{\varepsilon_{ic}}{\sigma_\eta}\right]. \quad (9)$$

The only restriction on σ_η is that it is positive. Since the μ_c are not observed, the MLE integrates out with respect to the μ_c using either a simulation method or numerical quadrature (Butler and Moffitt 1982).

Some maximum likelihood procedures impose $\sigma_\eta = \sqrt{\sigma^2_\epsilon + \sigma^2_\mu}$, and so the overall error variance is 1. These procedures estimate as coefficients the quantities $\beta/\sqrt{\sigma^2_\varepsilon + \sigma^2_u}$. Since this is the same normalization as a simple probit procedure ($\sigma^2 = 1$), one can directly compare the coefficient estimates of these MLE to simple probit. Other maximum likelihood procedures, such as Stata's xtprobit, impose the normalization so that $\sigma_\eta = \sigma_\epsilon$, so the estimated coefficients equal $\beta/\sqrt{\sigma^2_\varepsilon}$. This means that direct comparisons of the point estimates of simple probit and the MLE do not make sense unless $\rho = 0$. A study by Mroz and Zayats (2003) suggests that the failure to account for different arbitrary normalizations across estimation procedures has led some researchers (e.g., Rodríguez and Goldman 1995) to conclude incorrectly that simple probit models yield biased estimators in multilevel settings. Note, however, that significance levels are not affected by the choice of normalization and that it is possible to compare ratios of coefficients across estimation procedures.

We can also extend the three-level continuous dependent variable model to probit:

$$Y^*_{ifc} = \beta_0 + \beta_1 X^C_c + \beta_2 X^F_{fc} + \beta_3 X^{IC}_{ifc} + \mu_c + \lambda_{fc} + \epsilon_{ifc}, \quad (10)$$

where all terms are as defined above. As in the two-level model, numerical methods can be used to integrate out with respect to μ_c and λ_{fc}. Again, it is important to recognize that one needs to impose some arbitrary normalization. A useful approach is to normalize by $\sqrt{\sigma_\epsilon^2 + \sigma_\mu^2 + \sigma_\lambda^2}$; this makes the coefficient estimates directly comparable to those estimated by simple probit. Often, however, researchers normalize by σ_ϵ, which requires one to adjust the coefficients before comparing them to those obtained from a simple probit estimation. Again, ratios of coefficients and significance levels do not depend on the chosen normalization. Just as in the two-level model, Eicker-Huber-White standard errors are asymptotically correct as long as one corrects at the outmost level—community, in this example. However, there is no Monte Carlo evidence on the finite sample performance of these standard errors or the MLE.

Our application for the probit model uses data with three levels from the Philippines, where the outcome of interest is whether a woman of reproductive age uses family planning. The data are cross-sectional with province (Level 3), municipality (Level 2), and individual (Level 1) observations. We use province- and municipality-level health care expenditure data from 1998 matched to individual-level women from the 1998 Philippines National Demographic and Health Survey (DHS). The higher level data come from a Commission on Audit survey (see Schwartz, Guilkey, and Racelis 2001). There are 7,492 Level 1 observations residing in 484 communities contained in 75 provinces. Descriptive statistics for all variables are presented on the right-hand side of Table 1. In the multilevel estimation models, we impose the restriction that the total error variance is 1.00; that allows us to compare coefficients across models without needing to renormalize coefficients.

The original reason for gathering the expenditure data was to measure the effect of devolution of public health expenditures on health outcomes. In discussing the results, we focus on the effects of public health expenditures at the province and municipality levels on contraceptive use that are reported at the top of Table 3. A comparison of the standard errors of the three simple probit estimators shows that the standard errors almost triple for both variables when we correct for possible error correlation at the municipality level and further

increase when we correct at the province level. Province-level public expenditures are insignificantly different from zero at any reasonable level of significance. The point estimate of the impact of the public municipality health expenditures is much larger, with a 2–standard deviation increase in municipal health expenditures, implying, at the mean probability, a 6 percentage point increase in the fraction of women using contraception. This effect, however, is not significantly different from zero at the 5 percent level once one recognizes that there could be error correlations at up to the municipal ($p = .10$, in column 2) or province level ($p = .13$, in column 3).

We now turn to the MLE results reported in the last three columns of Table 3.[11] As in Table 2, the standard errors reported in columns 4 and 5 are those obtained from the maximum likelihood procedure, while column 6 contains two sets of standard errors. The first standard errors come from the maximum likelihood procedure and are valid when the statistical model is the true population model. The second set uses robust standard errors allowing for clustering at up to the province level. We also calculated bootstrap standard errors for all three of these maximum likelihood models (but not included in Table 3 for columns 4 and 5), where we sampled with replacement at the province level. For all coefficients but one, the Eicker-Huber-White standard errors were within 10 percent of the bootstrap standard errors. In addition, for each of these three maximum likelihood models, there was considerable evidence that the standard errors as reported by the maximum likelihood procedure differed appreciably from those obtained by using the two robust approaches. The standard errors reported for columns 4 and 5 are thus likely to be incorrect.

When one only estimates two-level models, the intraclass correlations are .45 and .21 for municipality and province Level 2 components. When the three-level model is estimated, the municipality-level ρ is .40, while the province-level ρ falls to .06. Using the standard errors obtained from the maximum likelihood procedure, the p values for the two ρs in the three-level model are .00 and .05, respectively. The province-level error correlation, however, is only significant at about the 25 percent level if one uses the robust standard error. A chi-square test for twice the increase of the likelihood function when moving from column 4 to column 6, however, is significant at below

TABLE 3: Determinants of Contraceptive Use in the Philippines (Probit With Standard Errors in Parentheses)

Variable	1. Probit	2. Probit Municipality Correction	3. Probit Province Correction	4. MLE Municipality Only	5. MLE Province Only	6. MLE Both (MLE SE) (Robust SE)
			Province and municipality			
Public province health expenditures	0.0006 (0.0438)	0.0006 (0.1324)	0.0006 (0.1325)	−0.0597 (0.0948)	0.0257 (0.1270)	−0.0411 (0.1125) (0.1013)
Public municipality health expenditures	0.2622 (0.0559)	0.2622 (0.1614)	0.2622 (0.1733)	0.1423 (0.1075)	−0.0230 (0.0709)	0.0717 (0.1198) (0.1524)
			Individual			
Ages 15 to 19	0.9492 (0.1533)	0.9492 (0.2900)	0.9492 (0.2906)	0.6959 (0.1349)	0.8551 (0.1404)	0.6904 (0.1345) (0.2722)
Ages 20 to 24	1.3329 (0.0823)	1.3329 (0.1572)	1.3329 (0.1612)	1.0283 (0.0760)	1.2182 (0.0794)	1.0242 (0.0769) (0.1180)
Ages 25 to 29	1.3994 (0.0773)	1.3994 (0.1439)	1.3994 (0.1398)	1.2037 (0.0722)	1.2773 (0.0762)	1.1961 (0.0725) (0.1132)
Ages 30 to 34	1.2681 (0.0773)	1.2681 (0.1403)	1.2681 (0.1259)	1.1763 (0.0721)	1.1636 (0.0748)	1.1679 (0.0723) (0.0923)
Ages 35 to 39	1.1715 (0.0795)	1.1715 (0.1544)	1.1715 (0.1613)	1.0471 (0.0735)	1.0945 (0.0764)	1.0406 (0.0736) (0.1148)
Ages 40 to 44	0.7975 (0.0867)	0.7975 (0.1898)	0.7975 (0.1918)	0.6539 (0.0799)	0.7760 (0.0818)	0.6494 (0.0798) (0.1255)
Education 0 to 5 years	−0.4444 (0.0657)	−0.4444 (0.1644)	−0.4444 (0.1508)	−0.4000 (0.0659)	−0.3861 (0.0637)	−0.3945 (0.0659) (0.1232)
Education 6 years	−0.1519 (0.0577)	−0.1519 (0.1444)	−0.1519 (0.1574)	−0.2350 (0.0560)	−0.2034 (0.0553)	−0.2374 (0.0559) (0.1074)
Education 7 to 10 years	0.1269 (0.0475)	0.1269 (0.1095)	0.1269 (0.1242)	0.1157 (0.0449)	0.0530 (0.0449)	0.1108 (0.0448) (0.0761)
Partner's education 0 to 5 years	0.1206 (0.0608)	0.1206 (0.1683)	0.1206 (0.1611)	0.0163 (0.0584)	0.0502 (0.0577)	0.0156 (0.0582) (0.1464)
Partner's education 6 years	0.3344 (0.0596)	0.3344 (0.1548)	3344 (0.1529)	0.2090 (0.0573)	0.2941 (0.0573)	0.2081 (0.0572) (0.1374)
Partner's education 7 to 10 years	0.2607 (0.0462)	0.2607 (0.1174)	0.2607 (0.1007)	0.1520 (0.0436)	0.2538 (0.0441)	0.1555 (0.0435) (0.0888)
Urban	−0.0484 (0.0352)	−0.0484 (0.0894)	−0.0484 (0.0945)	−0.1142 (0.0500)	−0.1050 (0.0359)	−0.1187 (0.0504) (0.0826)
Asset index	0.0082 (0.0081)	0.0082 (0.0251)	0.0082 (0.0242)	−0.0057 (0.0081)	0.0110 (0.0079)	−0.0049 (0.0081) (0.0189)
Religion is Catholic	0.0675 (0.0377)	0.0675 (0.0929)	0.0675 (0.1140)	0.0176 (0.0398)	−0.0304 (0.0390)	0.0073 (0.0402) (0.0830)
Constant	−0.9751 (0.0964)	−0.9751 (0.2163)	−0.9751 (0.2649)	−0.5731 (0.1111)	−0.6980 (0.1285)	−0.5534 (0.1207) (0.1927)
ρ_λ				0.4493 (0.0241)		0.3958 (0.0315) (0.0384)
ρ_μ					0.2215 (0.0336)	0.0561 (0.0290) (0.0505)

NOTE: MLE = maximum likelihood estimate; SE = standard error.

the 2 percent level, so we focus on the model controlling for both Level 2 and Level 3 error correlation.

For the individual-level coefficients reported in columns 3 and 6 of the lower panel of Table 3, the changes in point estimates and standard errors follow the same patterns as seen in Table 2 for the continuous outcome. The point estimates change by relatively small amounts from using multilevel models, and one can obtain somewhat more efficient estimators by using multilevel models instead of simple models.

A much different situation arises for the higher level explanatory variables. The estimated impact of municipal health expenditures, for example, falls by over 70 percent from the probit estimate (in columns 1, 2, and 3) to the multilevel estimates with controls for province- and municipal-level correlations. With the multilevel model that has error correlation controls only at the province level, the estimated effect actually becomes negative. Such changes should not be statistically significant if the underlying model is correct.

Even though neither estimate of the impact of municipal-level expenditures in columns 3 and 6 is individually significantly different from zero, the point estimate does change significantly between the two columns. To see this, note that maximum likelihood estimation of the three-level model is an efficient estimator when the model is correctly specified. The simple probit estimator should also be consistent in this situation, so one can carry out a Hausman (1978) specification test. The change in the point estimates is 0.1905 between columns 3 and 6. The estimate of the standard error of this difference, calculated from the differences in the estimated variances, is 0.083. With a t statistic of over 2.3, this change indicates a potential model specification problem. This significant change might be due to the fact that government health expenditures are not allocated independent of perceived need, resulting in the endogeneity of the health expenditures. Schwartz et al. (2001) tested the endogeneity of health expenditures using both the 1993 and 1998 data, and they found strong evidence that health expenditures at the municipality level are endogenous determinants of contraceptive use. When important point estimates of effects do change significantly when one uses a more detailed multilevel model instead of a simple model, it is an indication that the researcher should consider spending time

examining the underlying assumptions of the model instead of trying to obtain more "precise" estimates from more refined multilevel models.

6. CONCLUSION

This article presents both analytical and simulation evidence on the finite sample performance of the OLS estimator in multilevel models with up to three levels of data in comparison to the MLE. We focus on the correct measurement of the impacts of community-level variables. These variables are often the variables of primary policy interest, but the performance of alternative estimators for the impacts of these important variables appears to have been neglected in the literature. In fact, we could find no other work that focused on the finite sample performance of the estimators of the impacts of Level 2 variables in two-level models; previous studies focused on the correct measurement of the impact of Level 1 variables, which is where one finds the largest efficiency gains from using multilevel models (Angeles and Mroz 2001). In addition, some of our most interesting results deal with three-level models whose finite sample performances have not been studied previously, to our knowledge.

Even though the OLS point estimators appear to perform quite well relative to the maximum likelihood estimators in most applied situations, the standard error estimators provided by standard OLS formulas are incorrect in the presence of multilevel-error correlations. For two-level models, we find that the robust asymptotic approximations to the standard errors of the OLS model due to Eicker, Huber, and White provide approximately unbiased tests for all parameter estimators when one uses formulas that allow error correlations at the higher level. The two-level maximum likelihood standard error estimators perform flawlessly for these two-level models with homoskedastic errors.

We also examined random intercept and random intercept/random slope versions of three-level models. In both cases, OLS point estimates with robust standard error estimators allowing for error correlations within the highest level continue to perform quite well, even with the complex forms of heteroskedasticity that result from random slopes. The maximum likelihood estimators that assume

only two error levels and fixed slopes, on the other hand, often perform poorly. This failure of the maximum likelihood estimators is due to the fact that they are incorrectly specified for the three-level models we examine. It is important to note that one will usually obtain biased tests with these "maximum likelihood" estimators even when they control for correlations at the highest level. This could be an important factor to consider when using maximum likelihood estimators if there could be a missing "middle" level in a researcher's empirical model. Ordinary least squares estimators with the robust, Eicker-Huber-White standard error estimators do not have this limitation. Note, however, that robust standard error estimators could be used with these incorrectly specified maximum likelihood estimators to resolve these problems, but we have not examined experimentally such procedures in extensive detail.

If one is primarily interested in estimating the average impacts of community-level variables (Level 2) on individual-level (Level 1) outcomes, then the results of this study provide some important guidelines. First, unless both the intraclass correlations and the correlations of the regressors across levels are large, there are typically small efficiency gains for the estimators of the impacts of community-level factors on the individual-level behaviors from using maximum likelihood procedures instead of simple ordinary least squares estimation. In fact, for estimators of the impacts of the higher level variables, our analytic results reveal that with a random assignment of covariates at the higher level (e.g., an experimental assignment), there would be no efficiency gain by using maximum likelihood estimators instead of OLS estimators. Second, it is crucial to adjust the estimated standard errors of the ordinary least squares estimators to reflect the fact that there can be correlated error terms at higher levels; the robust standard error estimators appear to provide adequate adjustments. Third, even if there are complex multilevel error correlations in the data and heteroskedasticity caused by random slopes, the robust standard error adjustments always provide unbiased tests, as long as one allows for error correlation at the highest level. Simple two-level maximum likelihood models do not provide unbiased tests when lower level error correlations or heteroskedasticity are present.

In addition to the theoretical results, we also present empirical illustrations of the methods for the continuous dependent variable

case and present extensions of the probit model to three-level error structures along with an empirical example. The empirical examples dovetail nicely with the theoretical results; they demonstrate how the standard error estimates and hypothesis tests change when one allows for error correlations at higher levels. Our final example highlights what might be a most important concern: All estimation approaches could yield potentially misleading results if the model is not correctly specified. Given that it is straightforward to obtain robust standard error estimators for simple estimation approaches and that there appear to be only slight efficiency gains for the estimators of the effects of higher level covariates, a researcher's time might be spent better by evaluating key maintained assumptions in a model rather than by trying to incorporate multilevel error structures into the estimation of point estimates.

NOTES

1. We evaluate the analytic formulas for the variances of the two estimators of each regression coefficient, calculate their ratio, and take the square root. This provides a ratio of the standard deviations of the estimators. A value of 1.10, for example, would mean that the ordinary least squares (OLS) estimator would have a standard error of estimate 10 percent higher than the maximum likelihood estimator of the coefficient for the same data-generating process (DGP); heuristically, t statistics would tend to be about 10 percent smaller for the OLS estimator than they would be for the maximum likelihood estimator.

2. We used this range of individuals to represent roughly the distribution of the number of adult women per community in the Demographic and Health Survey (DHS). For each community in each replication of each data-generating process, we selected the number of individuals per community by drawing from a truncated normal distribution with mean 25.5 and standard deviation 10, with the truncation points set at 1 and 50. We then took the integer portion of this truncated normal random variable as the choice of the number of individual-level observations per community. This yields a mean number of individuals per community of 25 and a standard deviation of 9.5. Using this procedure, 91 percent of the time, the number of individuals per community lies in the range [9,41].

3. This definition differs from the usual definition of a power function. For the more standard definition of the power function, one tests an identical hypothesis (e.g., $H_0 : \alpha = 2$ vs. $H_a : \alpha \neq 2$) and graphs the probability of rejection as a function of a varying true value of the parameter. Here, we graph the probability of rejecting a varying null hypothesis, when the true parameter value is 1.0, as a function of hypothesized values specified in the null and alternative hypotheses.

4. The alternative hypothesis for all power and size tests discussed in this study is the complement of the null hypothesis under examination.

5. What we attempt to evaluate here are simple-to-use estimators that are available in many multipurpose statistical packages. Consequently, we do not examine correctly specified

maximum likelihood estimators that recognize the possible three-level error structure. Such models should provide accurate estimates and unbiased hypothesis tests for the DGPs we examine.

6. If individuals are members of families located within communities, then the intraclass correlation we consider is the correlation of the disturbances among individuals within the same family.

7. Figures 5, 6, and 7 only examine tests at size .05. We obtained quite similar results for size .10.

8. There could be a cost of specifying the "clustering" level higher than is necessary. It is important for there to be enough independent higher level observations for this estimator to work well. In fact, the estimator will not provide a positive definite covariance matrix unless there are at least as many independent higher level units as parameters being estimated. Typically, one would like to have many more than this number of observations to obtain accurate estimators of the standard errors of the parameter estimates. If there is no community-level error correlation but one specifies that there could be error correlation within communities, this approach will yield valid standard error estimators as long as the number of communities is large.

9. We obtained approximately the same probabilities of false rejections for all approaches and for all data-generating processes when we examined R^2 values of 0.90 instead of the 0.10 examined in these figures.

10. These robust standard errors were calculated by using the sandwich standard error estimators in MlwiN. We also calculated bootstrap standard errors by drawing 250 samples at the Barangay level (the highest level at which we allow there to be error correlation) with replacement from the original data. We used the empirical standard deviations of the bootstrap point estimates as standard error estimates, and they were always within 10 percent of the robust standard errors for this example.

11. We estimated these maximum likelihood models using our own set of programs. Instead of using a Taylor expansion to approximate the likelihood function, we used 51 Gauss-Hermite points of support for each of the Level 2 and Level 3 errors and integrated over these points of support.

REFERENCES

Angeles, Gustavo, Jason Dietrich, David K. Guilkey, Dominic Mancini, Thomas A. Mroz, Amy O. Tsui, and Zhang Fengyu. 2001. "A Meta-Analysis of the Impact of Family Planning Programs on Fertility Preferences, Contraceptive Method Choice and Fertility." Working Paper WP-01-30, MEASURE Evaluation, Chapel Hill, NC.

Angeles, Gustavo and Thomas A. Mroz. 2001. "A Guide to Using Multilevel Models for the Evaluation of Program Impacts." Working Paper WP-01-33, MEASURE Evaluation, Chapel Hill, NC.

Bollen, Kenneth, David K. Guilkey, and Thomas A. Mroz. 1995. "Binary Outcomes and Endogenous Explanatory Variables: Tests and Solutions With an Application to the Demand for Contraceptive Use in Tunisia." *Demography* 32:111-31.

Borjas, George J. and Glenn T. Sueyoshi. 1994. "A Two-Stage Estimator for Probit Models with Structural Group Effects." *Journal of Econometrics* 64:165-82.

Bryk, Anthony and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Butler, John S. and Robert Moffitt. 1982. "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model." *Econometrica* 50:761-4.

Eicker, F. 1963. "Asymptotic Normality and Consistency of Least Squares Estimators for Families of Linear Regressions." *The Annals of Mathematical Statistics* 34:447-56.

———. 1967. "Limit Theorems for Regressions With Unequal and Dependent Errors." Pp. 59-82 in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics*, vol. 1. Berkeley: University of California Press.

Gertler, Paul J. and John W. Molyneaux. 1994. "How Economic Development and Family Planning Programs Combined to Reduce Indonesian Fertility." *Demography* 31(1):33-63.

Goldstein, Harvey. 1995. *Multilevel Statistical Models*. London: Arnold.

———. 2003. *Multilevel Statistical Models*. 3d ed. London: Kendall's Library of Statistics.

Guilkey, David and Susan Cochrane. 1995. "The Effects of Fertility Intentions and Access to Services on Contraceptive Use in Tunisia." *Economic Development and Cultural Change* 43:779-80.

Guilkey, David and Susan Jayne. 1997. "Zimbabwe: Determinants of Contraceptive Use at the Leading Edge of Fertility Transitions in Sub-Saharan Africa." *Population Studies* 51:173-89.

Guilkey, David and James Murphy. 1993. "Estimation and Testing in the Random Effects Probit Model." *Journal of Econometrics* 59:301-18.

Guilkey, David, Barry Popkin, John Akin, and Emelita Wong. 1989. "Prenatal Care and Pregnancy Outcomes in the Philippines." *Journal of Development Economics* 30:241-72.

Guo, Guang and Zhao Hongxin. 2000. "Multilevel Modeling for Binary Data." *Annual Reviews of Sociology* 26:441-62.

Hardin, James. 1997. "Panel Data Estimators." Paper presented at the 3rd meeting of the Stata U.K. Users' Group.

Harris, Katherine, Francesca Florey, Joyce Tabor, and J. Richard Udry. 2003. *The National Longitudinal Study of Adolescent Health: Research Design*. http://www.cpc.unc.edu/projects/addhealth/design.html

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46:1251-71.

Heckman, James J. 1981. "Statistical Models for Discrete Panel Data." In *Structural Analysis of Discrete Data With Econometric Applications*, edited by C. Manski and D. McFadden. Cambridge, MA: MIT Press.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions." Pp. 221-33 in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics*, vol. 1. Berkeley: University of California Press.

Kreft, Ita G. and Jan de Leeuw. 1998. *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage.

Maas, Cora J. M. and Joop J. Hox. 2002. "Sample Sizes for Multilevel Modeling." In *Social Science Methodology in the New Millennium: Proceedings of the Fifth International Conference on Logic and Methodology: Second Expanded Edition*, edited by J. Blasius, J. Hox, E. de Leeuw, and P. Schmidt. Opladen, Germany: Leske + Budrich Verlag.

Mátyás, László. 1992. "Error Component Models." Pp. 46-71 in *The Econometrics of Panel Data*, edited by L. Mátyás and P. Sevestre. Boston: Kluwer.

Mroz, Thomas A. and Yaraslau V. Zayats. 2003. "Estimated Coefficients, Information Sets, and 'Biases' in Nonlinear Models." Mimeo. Department of Economics and the Carolina Population Center, University of North Carolina at Chapel Hill.

Pendergast, Jane F., Stephen J. Gange, Michael A. Newton, Mary J. Lindstrom, Mari Palta, and Marian R. Fisher. 1996. "A Survey of Methods for Analyzing Clustered Binary Response Data." *International Statistics Review* 64:89-118.

Rasbash, Jon, William Browne, Harvey Goldstein, Min Yang, Ian Plewis, Michael Healy, Geoff Woodhouse, David Draper, Ian Langford, and Toby Lewis. 2000. *A User's Guide to MlwiN*. London: Center for Multilevel Modelling, University of London.

Raudenbush, Stephen and Anthony Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2 ed. Thousand Oaks, CA: Sage.

Robinson, Peter M. 1982. "On the Asymptotic Properties of Models Containing Limited Dependent Variables." *Econometrica* 50:27-41.

Rodriguez, German and Noreen Goldman. 1995. "An Assessment of Estimation Procedures for Multilevel Models With Binary Response." *Journal of the Royal Statistical Society: Series A* 158:73-89.

Schwartz, Brad, David K. Guilkey, and Rachel Racelis. 2001. "Decentralization, Allocative Efficiency, and Health Service Outcomes in the Philippines." Working Paper WP-02-44, MEASURE Evaluation, Chapel Hill, NC.

Thomas, Duncan and John Maluccio. 1995. "Contraceptive Choice, Fertility, and Public Policy in Zimbabwe." Working Paper No. 109, Living Standards Measurement Study, The World Bank, Washington, DC.

Tsui, Amy O. 1985. "Community Effects on Contraceptive Use." In *The Collection and Analysis of Community Data*, edited by B. J. Casterline. Voorburg, the Netherlands: International Statistical Institute, World Fertility Survey.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48:817-30.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Gustavo Angeles is an assistant professor in the Department of Maternal and Child Health at the University of North Carolina at Chapel Hill. He is a health economist specializing in methods for monitoring and evaluating the impact of population and health programs. His work includes program evaluations in Bangladesh, Nicaragua, Tanzania, and Indonesia, as well as methods for collecting health service supply data. Articles include "Purposive Program Placement and the Estimation of Family Planning Program Effects in Tanzania" (with D. Guilkey and T. Mroz, Journal of the American Statistical Association, *1998).*

David K. Guilkey is the Cary C. Boshamer Professor of Economics at the University of North Carolina at Chapel Hill. He is an applied econometrician specializing in estimation methods that are useful with large survey data sets. Through his work with the MEASURE Evaluation Project, much of his recent research has involved evaluations of USAID's health and population programs. Recent articles include "Effect of Childbearing on Filipino Women's Work Hours and Earnings" (with L. Adair, E. Bisgrove, and S. Gultiano, Journal of Population Economics, *2002) and "Determinants of Contraceptive Method Choice in Rural Tanzania Between 1991 and 1999" (with S. Chen,* Studies in Family Planning, *2003).*

Thomas A. Mroz is a professor of economics at the University of North Carolina at Chapel Hill. He is a labor economist and economic demographer specializing in econometric evaluations of behavioral life cycle models. He received his Ph.D. from Stanford University. His research ranges from health, nutrition, and transition

economics to program evaluation and applied econometrics. Recent articles include "A Flexible Approach for Estimating the Effects of Covariates on Health Expenditures" (with D. Gilleskie, Journal of Health Economics, *forthcoming) and "The Gender Gap in Wages in Russia From 1992 to 1995" (with E. Glinskaya,* Journal of Population Economics, *2000).*