Multilevel Modeling with Correlated Effects *

Jee–Seon Kim
Edward W. Frees
University of Wisconsin–Madison

2006

Psychometrika, accepted

*This research was supported by National Academy of Education/Spencer Foundation and the National Science Foundation, Grant Number SES-0436274. Correspondence concerning this paper should be addressed to Jee–Seon Kim, Department of Educational Psychology, University of Wisconsin at Madison, 1025 Johnson Street, Madison, WI 53706. Electronic mail may be sent to jeeseonkim@wisc.edu.
Multilevel Modeling with Correlated Effects

Abstract

When there exist omitted effects, measurement error, and/or simultaneity in multilevel models, explanatory variables can be correlated with random components and standard estimation methods do not provide consistent estimates of model parameters. This paper introduces a new statistical technique to obtain consistent estimators under such conditions. By employing generalized method of moments (GMM) estimation techniques to multilevel models, the authors present a series of estimators along a robust to efficient continuum. This continuum depends on the nature of the assumptions that the analyst makes regarding the extent of the correlated effects. It is shown that the GMM approach provides an overarching framework that unifies well-known estimators such as fixed effects estimators and random effects estimators and also provides more options. It is also shown that GMM estimators can be expressed as instrumental variable (IV) estimators which enhances the interpretability of the estimates. Moreover, unlike traditional IV methods, the current technique can obtain robust IV estimators without requiring additional variables by exploiting the hierarchical structure of the data. Further, statistical tests are developed to compare the different estimators. A simulation study examines the finite sample properties of the estimators and tests and confirms the theoretical order of the estimators with respect to their robustness and efficiency. It further shows that not only are regression coefficients biased, but variance components can be severely underestimated in the presence of correlated effects. Empirical standard errors are introduced as they are less sensitive to correlated effects when compared to model-based standard errors. An example using student achievement data shows that GMM estimators can be effectively used in a search for the most efficient among unbiased estimators.

Key words: generalized method of moments, omitted variable bias, model specification test, robust estimation, consistency, empirical standard errors, hierarchical linear models
1 Introduction

The omission of important variables, possibilities of measurement errors and issues of simultaneity bias are critical model misspecification problems in multilevel analysis. Because multilevel models are often used to analyze observational data, researchers may not have opportunities to design a study that is appropriate for addressing these issues. To assist these researchers, this paper presents statistical methodology that can be used to obtain consistent regression coefficient estimators even in the presence of these difficulties.

1.1 Correlated effects

In regression modeling, it is customary to assume that predictors in a model are uncorrelated with the error term. As in single-level regression, in multilevel models the consistency of standard estimation methods depends on the assumption that the predictors are uncorrelated with any random component (Raudenbush & Bryk, 2002; Goldstein, 2003; Snijders & Bosker, 1999). Unlike single-level regression, however, there are multiple independence assumptions in multilevel models involving various random components at different levels. The most stringent assumption is that all predictors are uncorrelated with all random components. Although this strict assumption is typically assumed without validation, even a moderate correlation between some predictors and a random component may result in severe bias in the estimation of model parameters (Ebbes et al., 2004; Kim & Frees, in press).

There are three common sources of correlated effects: First, some unobserved effects can be attributed to omitted variable bias. Second, predictors might be measured imprecisely and result in measurement error or error-in-variables bias. Third, some predictors may not only cause but also be influenced by the outcome variable, yielding simultaneity bias. If any variable in a model suffers from one of these problems, all regression coefficients and variance component estimates in the model may be biased (Hayashi, 2000; Wooldridge, 2002; Frees, 2004).

Correlated effects arising from these common sources have been recurrent issues in the social sciences. For example, in educational research, Griliches (1977) demonstrated problems in estimating the effects of schooling when unobserved ability bias and/or errors of measurement exist. Boardman and Murnane (1979) promoted the use of panel data to minimize omitted variable bias in estimating the determinants of educational achievement. Ecob and Goldstein (1983) suggested instrumental variable approaches for estimating the reliability of test scores to reduce errors in variables. In 1990s, several well-known educational multilevel datasets were reanalyzed—for example, High School and Beyond (HS&B: Ehrenberg & Brewer, 1994), National Education Longitudinal Study of 1988 (NELS88: Goldhaber & Brewer, 1995; Ludwig & Bassi, 1997), and the Coleman
report data (Ehrenberg & Brewer, 1995)—so as to allow for the possibility that school and/or teacher characteristics were correlated with unobserved variables.

Recently, a number of researchers have examined the issues of correlated effects in multilevel models in the context of model specification, diagnostics, and consistent estimation. For example, Ebbes et al. (2004) conducted a simulation study to demonstrate the effects of level-1 and level-2 dependencies between predictors and random components on parameter estimation. McCaffrey et al. (2004) showed that omitted variable bias becomes especially hazardous when schools are heterogeneous in multilevel value-added models. Snijders and Berkhof (2005) discussed diagnostics for multilevel model specification. Also, several statistical techniques aiming at consistent estimation have been proposed, such as instrumental variable approaches (Spencer & Fielding, 2002), the conditional iterative generalized least squares (CIGLS) algorithm (Rice et al., 2002), multiprocess multilevel modeling (Steele, 2003), an omitted variable estimator method (Rothstein, 2004), and fixed effects transformation (Kim & Frees, 2005, in press). Fielding (2004) provided a review of the literature from Hausman (1978) to Skrondal and Rabe-Hesketh (2004) on these issues.

1.2 An illustration

Consider a two-level, random intercepts model of the form:

\[ y_{ct} = \alpha_c + x'_{ct} \beta + \epsilon_{ct}. \] (1)

Here, \( y_{ct} \) denotes the response variable (achievement test score) for child \( c \) at time \( t \); the predictors \( x_{ct} \) may also depend on child and time. To illustrate, consider four predictors: gender, ethnicity, family income, and a binary variable indicating whether or not a student switched schools in a particular year. The first two variables are taken to be constant over time whereas the latter two are time varying. The parameter \( \alpha_c \) represents a latent variable that is constant over time; it accounts for the correlation among a child’s test scores over time as well as individual differences in test scores among children. It can be viewed as a proxy for unobserved child characteristics, for example, the ability, proficiency, or aptitude of the respondent.

When \( \alpha_c \) is uncorrelated with the predictors, as is assumed in conventional multilevel analysis, the estimation of the coefficient \( \beta \) is straightforward and the generalized least squares (GLS) method is perhaps the most common estimation algorithm. The GLS estimators are equivalent to empirical Bayes estimators in this context and are often referred to as random effects estimators. Random effects estimators are unbiased and efficient when the independence assumption holds. However, children’s aptitude will likely be correlated with variables like parent aptitude and/or parent occupation, while parent occupation may be correlated with variables like family income.
and the likelihood of moving, for example. As a result, $\alpha_c$ may not be independent of certain predictors in the model, and GLS estimators may consequently provide biased and inconsistent estimators of the regression coefficients in the model.

To illustrate the consequences of the correlated effects, consider a stacked version of (1) for each child

$$y_c = 1_c \alpha_c + X_c \beta + \epsilon_c,$$

where $1_c$ is a vector of ones. Suppose that the joint distribution of $\alpha_c$ and $X_c$ can be defined by a multivariate normal distribution

$$\begin{pmatrix} \alpha_c \\ x_c \end{pmatrix} \sim N \left( \begin{bmatrix} 0 \\ \mu_X \end{bmatrix}, \begin{bmatrix} \sigma^2_{\alpha} & \Sigma_{\alpha X} \\ \Sigma_{X \alpha} & \Sigma_{XX} \end{bmatrix} \right),$$

where $x_c$ is the column vector associated with $X_c$. Then, standard calculations show that expected value of the outcome variable $y_c$ given the set of predictors is

$$E(y_c|x_c) = X_c \beta + 1_c E(\alpha_c|x_c),$$

where $E(\alpha_c|x_c) = \Sigma_{\alpha X} \Sigma_{XX}^{-1} (x_c - \mu_X)$. Thus, $E(\alpha_c|x_c)$ is nonzero unless $\Sigma_{\alpha X} = 0$. Therefore, $E(y_c|x_c) \neq X_c \beta$ so that the random effects estimators provide biased estimates of $\beta$ if $\alpha_c$ and $X_c$ are correlated.

### 1.3 Exogeneity, endogeneity, and instruments

Collecting the random effects terms on the right hand side, we may write the linear model

$$y = X \beta + \delta$$

where $\delta$ is a mean zero random variable. Following standard economics literature (Hayashi, 2000; Wooldridge, 2002; Frees, 2004), a set of variables that comprise the columns of the matrix $H$ is said to be predetermined if $E(H\delta) = 0$, that is, they are uncorrelated with $\delta$. Because this paper focuses on linear models, we will primarily use $E(H\delta) = 0$ as our condition for exogeneity; a variable that does not satisfy this condition is said to be endogenous. However, in Section 3.4, we find it useful to examine the more restrictive assumption $E(\delta|H) = 0$. Exogeneous variables in $H$ that are useful for estimating $\beta$ are said to be instruments. In most applications, there is a large overlap between model-based variables $X$ and instrumental variables $H$.

The estimators presented in this paper have natural interpretations in terms of instrumental variable (IV) estimation. In a linear model context, an IV estimator is

$$b_{IV} = (X'PX)^{-1} X'Py,$$  (2)
where $\mathbf{P} = \mathbf{P}(\mathbf{H}) = \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$ is the projection onto the linear space spanned by the columns of $\mathbf{H}$ and "\ldots" denotes a generalized inverse. Recall that in a classic regression problem one decomposes model variables $\mathbf{X}$ as $(\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ represents exogenous variables and $\mathbf{X}_2$ represents endogenous variables. The instruments $\mathbf{H}$ are decomposed as $(\mathbf{H}_1, \mathbf{H}_2)$ where $\mathbf{H}_1 = \mathbf{X}_1$ represents exogenous model variables included as instruments and $\mathbf{H}_2$ represents non-model variables. For consistent estimation, one requires that the number of instruments not from the model ($\text{rank}(\mathbf{H}_2)$) be at least as large as the number of endogenous model variables ($\text{rank}(\mathbf{X}_2)$). Thus, IV estimators readily accommodate the presence of endogenous model-based variables. Moreover, because both the responses and model-based variables are projected onto the space of instruments ($\mathbf{P}_y$ and $\mathbf{P}_x$, respectively), they could be computed in “two stages,” which is helpful in interpreting results.

Although endogeneity and correlated effects have been studied in the context of two-level panel data models (e.g., Arellano, 1993; Arellano & Bover, 1995; Arellano, 2003; Breusch et al., 1989), their implications for higher-level multilevel models have not been examined closely with a few exceptions (see section 1.1). It is well known that multilevel models can be written as linear mixed-effects models. However, this paper demonstrates that it is not straightforward to apply techniques designed for panel data models to multilevel models.

In contrast to panel data models that include only one random intercept of “individual effects” ($\alpha_c$ in equation 1), general multilevel models consist of multiple random intercepts and random slopes and require methodology that can handle higher levels and more general error structures. Consequently, there are many more possibilities for relationships between random components and predictors in multilevel models than in panel data models.

On the positive side, the hierarchical structure of multilevel data can be a great advantage for creating various internal instrumental variables, and we can exploit the multilevel nature of the data for robust estimation, as will be shown in the following sections. Given the growing popularity of multilevel models, the need for a better understanding of these issues seems obvious.

1.4 Bias in random effects estimators and omitted variable tests

To illustrate the issue of omitted effects in multilevel models, we revisit the example in Kim and Frees (2006) where they fit a three-level ‘time-child-school’ model and demonstrated bias in the random effects estimator. In their analysis, they examined the effects of retention and mobility on student outcomes, and faced correlated effects due to the lack of relevant variables including family characteristics such as family size, parent education and occupation, family income, and family involvement. While previous studies have shown these to be related to the retention and mobility of a student, this information was not available in the data.
The dataset was a subset of a much larger dataset from the Texas Assessment of Academic Skills (TAAS). It contained student achievement scores on a statewide mathematics test between 1994 and 2000 in Dallas, Texas; the subset consisted of 60 schools, 1,200 students, and 2,485 test scores. The level 1 time-varying variables consisted of the TAAS mathematics test score (dependent variable), grade level, retained (whether a student was retained in the same grade), switching schools (whether a student switched schools during the previous year), and the proportion of students at each school who were eligible for a free or reduced-price lunch program. The level 2 child-specific variables were gender, ethnicity, and cohort. Female and Caucasian were used as the reference categories for the gender and ethnicity variables, respectively. There were ten cohorts from the elementary class that graduated from 1994 to 2004. The level-3 school variables were percentage of minority students, average years of experience of teachers, and average class size.

Kim and Frees (2006) fit a three-level random slope model, where the grade effect varies across students. Using the SAS PROC MIXED procedure, they obtained the random effects estimator (GLS) which is the default and often the sole option in most multilevel modeling software. However, because of unobservable variables and subsequent correlated effects, they questioned the reliability of some of these efficient estimators. Therefore, they tested exogeneity hypotheses using their own procedure on the basis of a fixed child effects estimator and a fixed school effects estimator (see section 3.1 for the explanation of the procedure).

Although Kim and Frees obtained consistent estimators using the fixed child effects estimator, the estimated effects were available only for the time-varying variables and not for any child- and school-level variables. Sweeping out higher-level effects is a critical limitation especially for those who are interested in the hierarchical dynamics of educational systems such as, for example, researchers in school effectiveness and intervention studies or policy makers in educational programs. This might be one of the main reasons why efficient random effects estimators have been used routinely despite the danger of omitted variable bias.

This paper introduces a statistical technique that overcomes the limitations of fixed effects estimators and provides unbiased estimates for the effects of predictors at all levels in the presence of correlated effects. Specifically, this paper provides generalized method of moments (GMM) extensions of instrumental variable estimators that are useful in multilevel modeling. Using a GMM framework in Section 2, we extend IV estimators by (1) incorporating weights to accommodate the variance structure of a multilevel model and (2) specifying more general projections to take advantage of the hierarchical structure of multilevel models. We do not restrict our projections to those generated by a list of instruments. Moreover, Section 2.3 shows how a theory of hypothesis
testing based on nested projections can be used for specification of multilevel models.

As will be shown in Section 3, this new GMM estimation method includes the random and fixed effects estimators as special cases and, more importantly, provides “in-between” estimators between the two extremes. By considering multiple small steps between the two extremes, the current method helps us to find the most efficient among unbiased estimators. Section 4 describes the comparison among multilevel GMM estimators using the hypothesis testing theory. Section 5 summarizes a simulation study of finite sample properties and Section 6 illustrates an application using educational testing data. Section 7 describes some extensions of the theory and Section 8 concludes. The statement and proofs of all propositions appear in an Appendix.

2 GMM Inference

This section outlines some GMM results with an emphasis on features needed for multilevel models. Section 2.1 shows how to write a multilevel model as a mixed linear effects model, a simpler framework in which we express our results. Section 2.2 gives results for weighted regression coefficient estimators, where the weights are suggested by the variance structure of a multilevel model. Section 2.3 describes comparisons of estimators; in most GMM treatments, one compares estimators by examining additional conditions that are imposed by a hypothesis of interest. In multilevel modeling, it is of more interest to compare estimators that are projections onto linear subspaces.

2.1 Multilevel model specification

Consider three levels of nesting, where the subscript $s$ identifies a school, the subscript $c$ identifies a child within school $s$, and the subscript $t$ denotes time. We adopt the notation system for high order multilevel models as introduced in Frees and Kim (2006). With this set-up, we explicitly allow for unbalanced designs so, for example, we do not require the number of children per school be identical over schools. However, we do not explicitly model patterns of missingness such as drop-out, switchers and so forth. Following standard conventions, we assume these patterns arise due to an exogenous random mechanism (see, for example, Little and Rubin, 1987).

The level-1 model is then written as $y_{sct} = Z_{sct}^{(1)}\beta_{sc}^{(1)} + X_{sct}^{(1)}\beta_1 + \epsilon_{sct}^{(1)}$, where $y_{sct}$ denotes the response variable; further, the explanatory variables are $Z_{sct}^{(1)}$ and $X_{sct}^{(1)}$, and may be related to time (grade level, year and so on), the child (gender, family income and so on), or the school (organization, structure, location and so on). Parameters that are constant appear in the $\beta_1$ vector and so we interpret $X_{sct}^{(1)}\beta_1$ to be part of the “fixed effects” portion of the model. The term $\beta_{sc}^{(1)}$ captures latent, unobserved characteristics that are school and child specific (such as “ability”).

We wish to allow for, and test, the possibility that these latent characteristics are related to our
explanatory variables $Z_{sc}^{(1)}$ and $X_{sc}^{(1)}$.

For identification purposes, we adopt the usual convention and assume that the disturbance term, $\epsilon_{sc}$, is independent of the other right-hand variables, $Z_{sc}^{(1)}, X_{sc}^{(1)}$, and $\beta_1$. This assumption is maintained throughout Sections 2 to 6; in Section 7, we will discuss techniques that allows one to address situations where this assumption may be a concern.

The level-2 model describes the variability at the child level and is written as $\beta_{sc}^{(1)} = Z_{sc}^{(2)} \beta_2 + X_{sc}^{(2)} \beta_2 + \epsilon_{sc}^{(2)}$. Analogous to the level-1 model, the explanatory variables $Z_{sc}^{(2)}$ and $X_{sc}^{(2)}$ relate to the child or school and do not vary over time. Constant parameters appear in the $\beta$ term, $\epsilon_{sc}$, $Z_{sc}$ be related to explanatory variables captures latent, unobserved characteristics that are school specific; these latent characteristics may be related to explanatory variables $Z_{sc}^{(2)}$ and $X_{sc}^{(2)}$.

Finally, the level-3 model describes variability at the school level, and is written as $\beta_{s}^{(2)} = X_{s}^{(3)} \beta_3 + \epsilon_{s}^{(3)}$. The variables $X_{s}^{(3)}$ may depend on the school. We let $\epsilon_{s}^{(3)}$ represent omitted characteristics of the school that are not explained by the fixed effects portion, $X_{s}^{(3)} \beta_3$.

In our analysis, we combine the three levels into a single equation, linear mixed effects model as follows:

$$y_{sc} = X_{sc}^{(1)} \beta_1 + \epsilon_{sc}^{(1)} + Z_{sc}^{(1)} (X_{sc}^{(2)} \beta_2 + \epsilon_{sc}^{(2)}) + Z_{sc}^{(1)} Z_{sc}^{(2)} (X_{sc}^{(3)} \beta_3 + \epsilon_{sc}^{(3)}).$$

Next, define $Z_{2,sc} = Z_{sc}^{(1)}$ and $Z_{3,sc} = Z_{sc}^{(1)} Z_{sc}^{(2)}$. With this notation, we may summarize the disturbance term as $\delta_{sc} = \epsilon_{sc}^{(1)} + Z_{2,sc} \epsilon_{sc}^{(2)} + Z_{3,sc} \epsilon_{s}^{(3)}$. Further, define the $K \times 1$ vector $\beta = (\beta_1^{'}, \beta_2^{'}, \beta_3^{'})$ and the $1 \times K$ vector $X_{sc} = (X_{sc}^{(1)}, Z_{2,sc} X_{sc}^{(2)}, Z_{3,sc} X_{s}^{(3)})$. With this notation and stacking, we may express the model as

$$y_{s} = X_{s} \beta + \delta_{s}. \tag{3}$$

### 2.2 Mixed linear model estimation

We now work with mixed linear effects models as given in equation (3). Assume that $y_{s}$ is independent over $s$ with $E \delta_{s} = 0$, $\text{Var} \delta_{s} = V_{s}$ and $V = \text{blkdiag}(V_1, \ldots, V_n)$. The variance-covariance matrix $V$ is indexed by parameters $\tau$, known as variance components. Let $W_{s}$ be a known weight matrix associated with $\delta_{s}$. We use matrices of instruments $H_{s}$ for the estimation and identification of $\beta$ and assume

$$E H_{s}^{'} W_{s} (y_{s} - X_{s} \beta) = E H_{s}^{'} W_{s} \delta_{s} = 0, \quad s = 1, \ldots, n. \tag{4}$$

To simplify the notation, we define by stacking $y = (y_1^{'}, \ldots, y_n^{'})^{'}, X = (X_1^{'}, \ldots, X_n^{'})^{'}, H = (H_1^{'}, \ldots, H_n^{'})^{'}, \delta = (\delta_1^{'}, \ldots, \delta_n^{'})^{'}. \text{ With the matrix of weights } W = \text{blkdiag}(W_1, \ldots, W_n), \text{ we}
have the GMM estimator

\[ \hat{b}_{GMM} = (X'W_PX) - X'W_PY, \]

where \( P = P(H) = H(H'H)^{-1}H' \). This is also an instrumental variable estimator, and hence is simple to implement and interpret. For example, an intuitively pleasing feature of instrumental variable estimators is that they can be computed using two-stage least squares.

For this GMM estimator, we have introduced weights \( W_s \) to allow for a variance structure. Here, one may use the identity for the weight matrix in the case of ordinary least squares. Alternatively, the weight can be the inverse of the square root of the variance-covariance matrix of the disturbance term \( (V^{-1/2}) \), thus producing a generalized least squares estimate. As another option, in mixed linear effects modeling (see, for example, Diggle et al., 2002), it is customary for analysts to use a weight matrix that approximates \( V^{-1/2} \) and then use a robust estimate of standard errors to correct for misspecifications. Proposition 2 provides a robust estimator of the asymptotic variance,

\[ \widehat{\text{Var}} \hat{b}_{GMM} = (X'W_PX)^{-1}X'W_PW \text{blkdiag}(e_1e_1', \ldots, e_ne_n') WPWX' (X'W_PX)^{-1}, \]

determined from the residuals \( e_s = y_s - X_s \hat{b}(P) \).

2.3 Comparing estimators

Different GMM estimators arise by varying the set of instruments \( H \), and thus the projections \( P \), used. In standard economics, the practice is to use a set of instruments, say \( H_1 \), compute the corresponding estimator \( \hat{b}_1 \), and compare this to estimator \( \hat{b}_2 \) that is based on a second set of instruments \( H_2 = (H_1 : H) \). Comparison of \( \hat{b}_1 \) and \( \hat{b}_2 \) allows one to test the hypothesis that \( E H_1 \delta = 0 \) (under the maintained hypothesis that \( E H_1 \delta = 0 \)).

We show that this test procedure is available assuming only that the projections are nested. Two projections, say \( P_1 \) and \( P_2 \), are said to be nested if \( P_1P_2 = P_2 \), that is, the space spanned by \( P_2 \) is contained in \( P_1 \). We will compare two estimators, \( \hat{b}_1 \) and \( \hat{b}_2 \), of the form in equation (5), that use the same weights \( W \) and differ only based on the projections used, \( P_1 \) and \( P_2 \), respectively. We assume that the projections are nested with the former being the “larger” one, hence serving as our null hypothesis. We consider the test statistic

\[ TS = (\hat{b}_2 - \hat{b}_1)' (\text{Var} (\hat{b}_2 - \hat{b}_1))^{-1} (\hat{b}_2 - \hat{b}_1). \]

Test statistics of this form were introduced by Hausman (1978). However, Hausman required that estimators corresponding to the null hypothesis be efficient whereas we relax this requirement via the nesting of projections. Proposition 5 establishes conditions under which the test statistic \( TS \) has an asymptotic chi-square distribution. This proposition shows that the degrees of freedom
associated with the asymptotic distribution is \( df = \text{rank}(\Omega) \), where the matrix \( \Omega \) is first defined in Proposition 3. A robust estimator of \( \Omega \) is given in Proposition 6, allowing us to compute \( p \)-values to quantify the significance of the test statistic. The conditions needed for these results are essentially those implicitly given by Newey (1985).

Thus, each estimator can be computed using a projection and, where the projections are nested, the estimators are directly comparable. This is the same idea as with the likelihood ratio test that is used in compare nested hypotheses. We now demonstrate the usefulness of this weaker assumption in multilevel modeling.

3 Multilevel Model Estimation

3.1 Robust estimation using fixed effects transformations

We return now to the Section 2.1 multilevel set-up. A primary concern of this paper is to address situations in which the higher level disturbance terms, \( \epsilon^{(2)}_{sc} \) and \( \epsilon^{(3)}_{s} \), may be correlated with explanatory variables. One way to address these potential difficulties is to use a transformation to “sweep out” their effects; see, for example, Kim and Frees (2006) for extensive discussions. To this end, let \( Z_{2,sc} \) be the stacked version of \( Z_{2,sc}^{(1)} \) and consider the transformation matrix

\[
Q^{(1)}_{sc} = I_{sc} - P(Z_{2,sc}).
\]  

The transform \( Q^{(1)}_{sc} \) is orthogonal to \( Z_{2,sc} \) in the sense that \( Q^{(1)}_{sc} Z_{2,sc} = 0 \). Thus, \( Q^{(1)}_{sc} \delta_{sc} = Q^{(1)}_{sc} \epsilon^{(2)}_{sc} \), so the transformation removes the unobserved variables \( \epsilon^{(2)}_{sc} \) and \( \epsilon^{(3)}_{s} \) from \( \delta_{sc} \). Applying this transform yields

\[
Q^{(1)}_{sc} y_{sc} = Q^{(1)}_{sc} x_{sc}^{(1)} \beta_1 + Q^{(1)}_{sc} \epsilon^{(1)}_{sc}.
\]  

This expression underscores an important drawback of this approach; the parameters \( \beta_2 \) and \( \beta_3 \) are no longer estimable based only on this transformed model.

Nonetheless, desirable estimators of \( \beta_1 \) are straightforward to construct. Using ordinary least squares, the transformed model yields the robust estimator

\[
b_{1,FE} = \left(X'^{(1)}Q^{(1)}X^{(1)}\right)^{-1}X'^{(1)}Q^{(1)}y,
\]

where \( Q^{(1)} = \text{blkdiag}(Q^{(1)}_{sc}) \) and \( X^{(1)} \) is the stacked version of \( X_{sc}^{(1)} \). So that the \( \beta_1 \) parameters are estimable, we assume

\[
E X'^{(1)}Q^{(1)}_{sc} \delta_{sc} = 0,
\]  

that we interpret to mean \( E X'^{(1)}Q^{(1)}_{sc} \epsilon^{(1)}_{sc} = 0 \). In particular, we can allow correlations among \( \epsilon^{(2)}_{sc} \), \( \epsilon^{(3)}_{s} \), and \( X_{sc} \) and still achieve desirable, robust estimators of \( \beta_1 \). It will be convenient to write this estimator as

\[
b_{1,FE} = F_1 \left(X'Q^{(1)}X\right)^{-1}X'Q^{(1)}y,
\]  

\[\text{(10)}\]
where $F_1 = (I_1 : 0_1)$.

For another type of robust fixed effects estimator, consider the transformation matrix

$$Q_s^{(2)} = I_s - P(Z_{3,s}),$$

(11)

where $Z_{3,s}$ is the stacked version of $Z_{3,sc}$. Because the transform $Q_s^{(2)}$ is orthogonal to $Z_{3,s}$, this transformation removes the unobserved variables $\epsilon_s^{(3)}$ from the disturbance term $\delta_s$. This suggests an estimator of the form

$$b_{2,FE} = F_2 \left( X'Q_s^{(2)}X \right)^{-1} X'Q_s^{(2)}y,$$

(12)

where $Q_s^{(2)} = \text{blkdiag}(Q_s^{(2)}).$

### 3.1.1 Special case: Random intercepts model

For developing intuition and interpretation, we now consider the special case of the random intercepts model. Here, there are no slopes associated with the random coefficients so that $Z_s^{(1)} = 1$ and $Z_s^{(2)} = 1$. The child level combined model is

$$y_{sc} = X_{sc}^{(1)}\beta_1 + X_{sc}^{(2)}\beta_2 + X_{sc}^{(3)}\beta_3 + \epsilon_{sc}^{(1)} + \epsilon_{sc}^{(2)} + \epsilon_{sc}^{(3)},$$

where there are $t = 1, \ldots, T_{sc}$ observations for the $cth$ child in the $sth$ school, $c = 1, \ldots, n_s$ children in the $sth$ school and $s = 1, \ldots, n$ schools. From equation (8), the transformation matrix that removes both child and school effects is $Q_s^{(1)} = I_{sc} - \frac{1}{T_{sc}}1_{sc}1'_{sc}$, where $1_{sc}$ is a vector of ones of dimension $T_{sc} \times 1$, and $I_{sc}$ is an identity matrix of dimension $T_{sc} \times T_{sc}$. This matrix transforms a vector into deviations from its time-series average. For example, we have

$$Q_s^{(1)}y_{sc} = y_{sc} - \frac{1}{T_{sc}}1_{sc}1'_{sc}y_{sc} = y_{sc} - 1_{sc}\overline{y}_{sc} = \begin{pmatrix} y_{sc,1} - \overline{y}_{sc} \\ \vdots \\ y_{sc,T_{sc}} - \overline{y}_{sc} \end{pmatrix},$$

because $\overline{y}_{sc} = \frac{1}{T_{sc}}1'_{sc}y_{sc} = \frac{1}{T_{sc}} \sum_{t=1}^{T_{sc}} y_{sc,t}$. Note that this transformation sweeps out all time-constant terms, that is,

$$Q_s^{(1)} \left( X_{sc}^{(2)}\beta_2 + X_{s}^{(3)}\beta_3 + \epsilon_{sc}^{(2)} + \epsilon_{s}^{(3)} \right) = 0.$$

Similarly, the matrix $P_s^{(1)} = I_{sc} - Q_s^{(1)}$ transforms vectors into time-series averages, that is, $P_s^{(1)}y_{sc} = 1_{sc}\overline{y}_{sc}$. With these notations, it is straightforward to use the expression for the robust estimator in equation (10).

At the school level, the corresponding fixed effects transformation matrix is $Q_{1,s}^{(2)} = I_s - \frac{1}{N_s}1_s1'_s$. Here, $1_s$ is a vector of ones of dimension $N_s \times 1$, $I_s$ is an identity matrix of dimension $N_s \times N_s$, ...
and \( N_s = T_{s,1} + \ldots + T_{s,n_s} \) is the number of observations in the \( s \)th school. For the \( s \)th school, the comparable level-1 fixed effects transformation matrix is \( Q_{1,s}^{(1)} = \text{blkdiag} \left( Q_{1,sc}^{(1)} \right) \). Now, note that

\[
Q_{1,s}^{(1)} 1_s = \begin{pmatrix}
Q_{1,s,1}^{(1)} & \cdots & 0 \\
0 & \ddots & \vdots \\
0 & \cdots & Q_{1,s,n(s)}^{(1)}
\end{pmatrix} 1_s = \begin{pmatrix}
\left( I_{s,1} - \frac{1}{T_{s,1}} 1_s 1_s' \right) 1_s,1 & \cdots & 0 \\
0 & \ddots & \vdots \\
0 & \cdots & \left( I_{s,n(s)} - \frac{1}{T_{s,n(s)}} 1_s,n(s) 1_s,n(s)' \right) 1_s,n(s)
\end{pmatrix} = 0,
\]

where \( n(s) = n_s \). Thus, we have

\[
Q_{1,s}^{(1)} Q_{1,s}^{(2)} = Q_{1,s}^{(1)} \left( I_s - \frac{1}{N_s} 1_s 1_s' \right) = Q_{1,s}^{(1)} - \frac{1}{N_s} Q_{1,s}^{(1)} 1_s 1_s' = Q_{1,s}^{(1)},
\]

establishing that the projection matrices are \textit{nested}. This is a key property for this paper.

### 3.2 One-step GMM estimators

This section provides estimators of \( \beta_2 \) and \( \beta_3 \), allowing for correlated disturbance terms yet introducing alternative assumptions to permit consistent estimation. The estimators are “one-step” in the sense that they do not rely on estimates of the variability to improve efficiency and so use \( W = I \). Analysts that are uncertain about the estimation of the first moment portion of the model through \( \beta_2 \) and \( \beta_3 \) may be unwilling to rely on estimates of the variability through second moment assumptions.

Suppose that we are concerned with the child level random component \( (\epsilon_{sc}^{(2)}) \) but desire estimates of the parameters \( \beta_2 \) and \( \beta_3 \). To accommodate this, we now assume that certain variables contained in \( X_{sc} \) are uncorrelated with \( \delta_{sc} \). Define \( X_{1,sc} \) to be those columns belonging to \( X_{sc} \) that are exogenous to \textit{all} of the disturbance terms in the sense that \( E X_{1,sc}' \delta_{sc} = 0 \). With equation (9), we note this requirement is equivalent to

\[
E X_{1,sc}' P_{sc}^{(1)} \delta_{sc} = 0,
\]

(13)

where \( P_{sc}^{(1)} = I_{sc} - Q_{sc}^{(1)} \).

Motivated by these assumptions, we introduce a GMM estimator, which can be viewed as an extended Hausman-Taylor estimator, given by

\[
b_{1,GMM} = \left( X' Q^{(1)} X + X' P^{(1)} X_1 \left( X_1' P^{(1)} X_1 \right)^{-1} X_1' P^{(1)} X \right)^{-1} \left( X' Q^{(1)} y + X' P^{(1)} X_1 \left( X_1' P^{(1)} X_1 \right)^{-1} X_1' P^{(1)} y \right),
\]

(14)
where $\mathbf{P}^{(1)} = \text{blkdiag} \left( \mathbf{P}^{(1)}_{sc} \right)$. This extends the Hausman and Taylor (1981) estimator in the sense that we allow slopes, as well as intercepts, to vary by individual. The expression for $\mathbf{b}_{1,GMM}$ in equation (14) underscores the way in which it extends the $\mathbf{b}_{1,FE}$ estimator in equation (10).

To show that $\mathbf{b}_{1,GMM}$ can be expressed as an instrumental variable estimator, we now define a set of explanatory variables

$$\mathbf{H}_{1,sc} = \left( \mathbf{Q}^{(1)}_{sc} \mathbf{X}_{sc} : \mathbf{P}^{(1)}_{sc} \mathbf{X}_{1,sc} \right),$$

that can be viewed as instruments. Let $\mathbf{H}_1$ be the stacked version of $\{\mathbf{H}_{1,sc}\}$. With this notation, equations (9) and (13) are equivalent to

$$\mathbf{E} \mathbf{H}_1' \mathbf{\delta} = \mathbf{0}. \quad (15)$$

We refer to the equation (15) conditions as the exogeneity assumptions. These conditions suggest the following simpler expression for $\mathbf{b}_{1,GMM}$ as an instrumental variable estimator,

$$\mathbf{b}_{1,GMM} = \left( \mathbf{X}' \mathbf{P}(\mathbf{H}_1) \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{P}(\mathbf{H}_1) \mathbf{y}. \quad (16)$$

Note that these instruments are not external to the data set but rather are transforms of model variables ($\mathbf{X}$). That is, although we express $\mathbf{b}_{1,GMM}$ as an instrumental variable and calculate it using familiar projection operators, it does not require variables that are outside of, or external to, the model variables contained in $\mathbf{X}$.

For consistent estimation of $\beta$, we require the asymptotic rank of $\mathbf{X}' \mathbf{P}(\mathbf{H}_1) \mathbf{X}$ be equal to $K$, the number of regression coefficients. A necessary condition for this is that the number of columns of $\mathbf{X}^{(1)}$ that are not collinear to $\mathbf{Z}^{(1)}$ plus the number of columns of $\mathbf{X}_1$ should be at least as large as the number of regression coefficients.

For another case of interest, we assume that child level effects are uncorrelated with model variables but that school level effects are not. We have seen that the fixed effects estimator $\mathbf{b}_{2,FE}$ is robust to school level effects but does not estimate $\beta_3$. Analogous to the arguments given above, we can estimate $\beta_1$ and $\beta_2$ through the requirement that

$$\mathbf{E} \mathbf{X}_s' \mathbf{Q}_s^{(2)} \mathbf{\delta}_s = \mathbf{0}. \quad (17)$$

To estimate the parameters $\beta_3$, define $\mathbf{X}_{2,s}$ to be those columns belonging to $\mathbf{X}_s$ that are exogenous to the disturbance terms in the sense that $\mathbf{E} \mathbf{X}_{2,s}' \mathbf{\delta}_s = \mathbf{0}$. With equation (17), this is equivalent to

$$\mathbf{E} \mathbf{X}_{2,s}' \mathbf{P}_s^{(2)} \mathbf{\delta}_s = \mathbf{0}, \quad (18)$$

where $\mathbf{P}_s^{(2)} = \mathbf{I}_s - \mathbf{Q}_s^{(2)}$. From these arguments, we use the set of instruments

$$\mathbf{H}_{2,s} = \left( \mathbf{Q}_s^{(2)} \mathbf{X}_s : \mathbf{P}_s^{(2)} \mathbf{X}_{2,s} \right).$$
With this notation, the school level robust estimator is

$$b_{2,GMM} = (X'P(H_2)X)^{-1}X'P(H_2)y.$$  \hspace{1cm} (19)

For a third case, suppose again that we are concerned with child as well as school level effects. We now consider the case where we have sufficient information to estimate parameters associated with children, \(\beta_2\), although not schools, \(\beta_3\). Thus, we first use the child level transform \(Q^{(1)}\) to sweep out children and school effects and use equation (9). Next, we define \(X_{3,s}\) to be those columns belonging to \(X_s\) that are exogenous to the transformed disturbance terms in the sense that \(E X_{3,s}' Q^{(2)} s \delta_s = 0\). We use the set of instruments \(H_3 = (Q^{(1)} X : P^{(1)} P^{(2)} X_3)\). With this notation, we propose an intermediate robust estimator

$$b_{3,GMM} = (X'P(H_3)X)^{-1}X'P(H_3)y.$$  \hspace{1cm} (20)

### 3.3 Regression coefficient standard errors

As noted above, the estimators \(b_{j,GMM}\), for \(j = 1, 2, 3\), are ordinary least squares estimators in the sense that they do not rely on variance component estimators to improve their efficiency. As noted in Section 2.2, they are asymptotically normal. A consistent estimate of its asymptotic variance is of the form

$$\hat{\text{Var}}(b_{GMM}) = (X'P(H)X)^{-1}X'H(H'H)^{-1}\text{Var} (H'\delta)(H'H)^{-1}H'X(X'P(H)X)^{-1}.$$  

Thus, the key term to approximate is \(\text{Var} (H'\delta) = E H'\delta\delta'H\). Following the linear mixed effects literature (see, for example, Diggle et al., 2002), we outline two approaches, a *model-based* and an *empirical* approach.

The model-based approach uses estimates of \(\tau\), the parameters that summarize the variance of the disturbance terms, known as *variance components*. Further, assume the strict exogeneity of the slope variables, \(Z_{2,s}\) and \(Z_{3,s}\), so that one can condition on these variables and write

$$V_s = \text{Var} (\epsilon_s^{(1)} + Z_{2,s} \epsilon_s^{(2)} + Z_{3,s} \epsilon_s^{(3)}) = R_s + Z_{2,s}' (\text{Var} \epsilon_s^{(2)}) Z_{2,s} + Z_{3,s}' (\text{Var} \epsilon_s^{(3)}) Z_{3,s}'. $$

Here, we denote \(R_s = E (\epsilon_s^{(1)} \epsilon_s^{(1)'} = \text{Var} \epsilon_s^{(1)}\) and \(V_s = E (\delta_s \delta_s') = \text{Var} \delta_s\). Thus, with the model-based approach, one estimates the asymptotic variance of \(b_{GMM}\) by estimating the variance components of \(R_s\) and \(V_s\).

We illustrate the model-based approach using the first instrumental variable estimator. In this
case, the \( s \)th term is

\[
E \mathbf{H}'_{1,s} \delta_s \delta'_s \mathbf{H}_{1,s} = E \left( \begin{pmatrix} Q_s^{(1)} \\ X_{1,s}^{(1)} \end{pmatrix} \begin{pmatrix} \epsilon_s^{(1)} \\ \delta_s^{(1)} \end{pmatrix} \right) \left( \begin{pmatrix} \epsilon'_s^{(1)} & \delta'_s P_s^{(1)} X_{1,s}^{(1)} \end{pmatrix} \right) \]

\[
= E \left( \begin{pmatrix} Q_s^{(1)} R_s Q_s^{(1)} \\ X_{1,s}^{(1)} P_s^{(1)} R_s Q_s^{(1)} \end{pmatrix} \right) \left( \begin{pmatrix} Q_s^{(1)} R_s P_s^{(1)} X_{1,s}^{(1)} \\ X_{1,s}^{(1)} P_s^{(1)} V_s P_s^{(1)} X_{1,s}^{(1)} \end{pmatrix} \right).
\]

To estimate \( \tau \), first project the model onto the linear space spanned by \( \mathbf{H} \)

\[
y^* = P(\mathbf{H})y = P(\mathbf{H})X\beta + P(\mathbf{H})\delta = X^*\beta + \delta^*.
\]

With the transformed data, we may apply any of the usual procedures for estimating variance components, including maximum likelihood, restricted maximum likelihood (REML) and unbiased moment estimators such as MIVQUE (minimum variance quadratic unbiased estimation). For example, for REML one employs a transformation of the form \( Q^* = I - P(X^*) \) and uses the assumption that \( Q^*y = Q^*\delta \) is normally distributed. This transformation removes the regression coefficients in \( \beta \), and so the likelihood for only \( Q^*y \) involves variance components. For this calculation, one implicitly assumes that \( \mathbf{H} \) is independent of \( \delta \) (not just uncorrelated).

However, it is well-known that all of the variance components may not be identifiable for a given transformed equation. Kiefer (1980) pointed this out by observing that identification of variance components for an unstructured form of \( \mathbf{R} \) is impossible for a two-level random intercept model. Moreover, an analyst concerned with correlated effects may be unwilling to make assumptions about the behavior of (second moment based) variance component estimators. The simulation study in Section 5 demonstrates this difficulty. For these reasons, one may also be interested in an empirical estimator of \( \text{Var} \ \mathbf{b}_{\text{GMM}} \). To this end, define

\[
\hat{\text{Var}} \ \mathbf{b}_{\text{GMM}} = (X'P(\mathbf{H})X)^{-1} X' \mathbf{H} (H'H)^{-1} \left( \sum_s H'_s \mathbf{e}_s \mathbf{e}'_s H_s \right) (H'H)^{-1} H'X (X'P(\mathbf{H})X)^{-1},
\]

where we use the residuals \( \mathbf{e}_s = y_s - X_s \mathbf{b}_{\text{GMM}} \). This estimator is robust to incorrect specifications of the variance structure. In multilevel systems where covariance structures are complex, it can be especially comforting to use such a robust estimator.

We remark that, in applications, one would typically assume strict exogeneity of the slope variables \( \mathbf{Z} \). Formally, we only require that the instruments in \( \mathbf{H} \) be uncorrelated with the disturbance term \( \delta \). Nonetheless, because the fixed effects transform is a nonlinear function of the slope variables, it is difficult to think of situations wherein \( \mathbf{Z} \) is endogenous yet \( \mathbf{H} \) is exogenous.
3.4 Optimal GMM estimation

When consistent estimates of the variance components are not available, optimal GMM estimators can be constructed. However, as is well-known (see, for example, Hayashi, 2000), this type of estimator essentially relies on approximations of fourth moments and is not always reliable.

Instead, this section focuses on situations when consistent estimators of the variance components $\tau$ are available and considers robust estimators of $\beta$ that are more efficient than those introduced in Section 3.2. To this end, suppose for the moment that this matrix $V_s$ is known. Then, one can simply pre-multiply all variables, including the instrumental variables, by $V_s^{-1/2}$. To illustrate, from equation (8), the new transformation matrix is

$$Q^{(1)}_{s,GLS} = I_{sc} - P(V_s^{-1/2}Z_{2,sc}).$$

(21)

Following Section 3.2, the new set of instruments is

$$H_{1,s,GLS} = (Q^{(1)}_{s,GLS}V_s^{-1/2}X_s : P^{(1)}_{s,GLS}V_s^{-1/2}X_{1,s}),$$

where $P^{(1)}_{s,GLS} = I_{sc} - Q^{(1)}_{s,GLS}$. With this notation, an expression for the optimal GMM estimator is

$$b_{1,GMM,GLS} = \left(X'V_s^{-1/2}P(H_{1,GLS})V_s^{-1/2}X_s \right)^{-1}X'V_s^{-1/2}P(H_{1,GLS})V_s^{-1/2}y,$$

(22)

where $H_{1,GLS}$ is the stacked version of $\{H_{1,s,GLS}\}$ and $V = \text{blkdiag}(V_s)$.

The exogeneity condition on which this robust efficient estimator is based differs from the corresponding Section 3.2 condition. Specifically, the moment condition for the transformed system is

$$E H'_{1,s,GLS}V_s^{-1/2}\delta_s = 0$$

(23)

that can be expressed as

$$E X_s'V_s^{-1/2}Q^{(1)}_{s,GLS}V_s^{-1/2}\delta_s = 0$$

and

$$E X_{1,s}'V_s^{-1/2}P^{(1)}_{s,GLS}V_s^{-1/2}\delta_s = 0.$$

Stronger assumptions such as $E \left(\epsilon^{(1)}_s | X, V \right) = 0$ and $E (\delta_s | X, V) = 0$ are sufficient for both equations (15) and (23). In this paper, we assume that such stronger assumptions are available and do not dwell on this distinction. We note, however, that the distinction is important for some applications, such as dynamic panel data models with lagged dependent variables as explanatory variables.
It is also possible to use a “three-stage” least squares estimator such as
\[ b_{3SLS} = \left( X' P_V (H_{1, GLS}) X \right)^{-1} X' P_V (H_{1, GLS}) y, \]
where \( P_V (H) = H (H' V H)^{-1} H' \) is a weighted projection of \( H \). See, for example, the discussion by Im et al. (1999) of three-stage least squares estimators in the context of Hausman-Taylor estimators. Because of our interest in comparing robust estimators to the usual GLS estimators, we restrict attention to the robust efficient estimators described in this section.

The estimator \( b_{1, GMM, GLS} \) is a generalized least squares (GLS) version of \( b_{1, GMM} \). As is customary, when we say “efficient” we actually refer to asymptotic efficiency. In finite samples, there are no clear guidelines as to which estimate is preferred.

When all model variables are assumed to be exogenous (\( X = X_1 \)), then the IV estimator is the usual GLS estimator that we denote as \( b_{RE} \). Whereas \( b_{RE} \) assumes all explanatory variables are uncorrelated with the random intercepts and slopes in the model, \( b_{FE} \) allows for endogeneity of all effects, then sweeps out the random components along with explanatory variables at the same levels. The more general estimator \( b_{GMM} \) allows for some of the explanatory variables to be endogenous, and uses this information to build internal instruments. Intuitively, to obtain \( b_{GMM} \), one uses both the between and within variations of the exogenous variables, but only the within variation of the potentially endogenous variables. This is done by pre-multiplying the “deviation” projection matrix \( Q \) by all explanatory variables \( X \) and the “mean” projection matrix \( P \) by exogenous explanatory variables \( X_1 \) (see section 3.2 for details). Consequently, the instruments \( H \) consist of \( QX \) and \( PX_1 \).

Further, the instruments \( H \) help us to understand the relationship among \( b_{FE} \), \( b_{GMM} \), and \( b_{RE} \). As noted above, \( b_{GMM} \) equals \( b_{RE} \) when \( X = X_1 \), where all variables are assumed to be exogenous. On the opposite end where \( X_1 \) is empty, we obtain \( b_{FE} \) by using only the \( QX \) part of the instruments (see section 3.1). Therefore, \( X_1 \) can be as small as empty and as large as \( X \). In between, we obtain different \( b_{GMM} \)'s for different sets of \( X_1 \). Depending on the choice of instruments, \( b_{GMM} \) can be as robust as the corresponding \( b_{FE} \).

4 Comparing Estimators
There are many types of hypothesis tests that arise as special cases within this broad framework. The focus of this paper concerns the following five types.

1. **Fixed effects versus random effects estimators**

One may wish to test for omitted effects by comparing a fixed effects estimator to the random effects estimator. This test is an extension of the standard Hausman (1978) test and was considered earlier by Frees (2001) and Kim and Frees (2005). One computes the random effects estimator
(b_{GLS}) using weights \( W = V^{-1/2} \) and equation (14) with \( P_1 = P_{GLS} = I \), the identity matrix. This may be compared to the fixed effects estimator (b_{j,FE,GLS}) using equation (14) with \( P_2 = Q_G^{(j)} \), for either \( j = 1 \) or 2. In our framework, this is a valid test because \( P_1 \) is nested in \( P_2 \).

For the independence case, the degrees of freedom of this test is \( rank \left( X' \left( I - Q^{(j)} \right) \right) \). From the discussion in Section 3, this corresponds to the number of regression coefficients in level 1 (for \( j=1 \)) or levels 1 and 2 (for \( j=2 \)).

2. Fixed effects versus fixed effects estimators

One may wish to test for omitted effects at the second level, allowing for the fact that there exists omitted effects at the third level. This test was introduced earlier in a paper by Kim and Frees (2005). For this test, one compares the fixed effects estimator \( b_{1,FE,GLS} \) that provides protection against omitted variables at the second and third level to the relevant components of \( b_{2,FE,GLS} \) that only provides protection against omitted variables at the third level. This is possible because the projections used \( P_1 = Q_{GLS}^{(1)} \) and \( P_2 = Q_{GLS}^{(2)} \) are nested.

For the independence case, the degrees of freedom of this test is \( rank \left( X' \left( I - Q^{(2)} \right) Q^{(1)} \right) \). From the discussion in Section 3, this corresponds to the number of regression coefficients in level 3 (for \( j=2 \)).

3. GMM versus random effects estimators

The framework introduced in this paper also allows us to extend the Hausman and Taylor (1981) test to the multilevel model. For the GMM estimator (b_{j,GMM,GLS}), one uses the fact that only a subset of the explanatory variables are exogenous. In contrast, the random effects estimator assumes that all explanatory variables are exogenous. The assumption that the extra variables are exogenous can be tested by comparing \( b_{j,GMM,GLS} \) to \( b_{GLS} \). This test is possible because the projection \( P_2 = P(H_{1,GLS}) \) (for the case \( j = 1 \)) is nested within \( P_1 = I \).

For the independence case, the degrees of freedom of this test is \( rank \left( X' \left( I - P_{j,GMM} \right) \right) \).

4. GMM versus GMM estimators

Just as with fixed effects, one may wish to compare a GMM estimator to another GMM estimator that uses more instruments. To see this, recall that the projection is

\[
H_{1,s,GLS} = \left( Q_{s,GLS}^{(1)} V_s^{-1/2} X_s : P_{s,GLS}^{(1)} V_s^{-1/2} X_{1,s} \right).
\]

This test is possible because the first part of the projection is common to the two estimators whereas the second part simply contains a larger linear space, and hence is nested.

5. Fixed effects versus GMM estimators

The fixed effects estimator provides protection against omitted effects at the price of losing estimability of components of \( \beta \). The corresponding GMM estimator recovers these estimates of
\( \beta \) at the price of additional exogeneity assumptions. One may wish to test whether these moment assumptions are valid. This test is possible because the fixed effects projection, \( P_{j,FE} = Q_{GLS}^{(1)} \), is nested within the instrumental variables projection \( H_{1,GLS} = (Q_{GLS}^{(1)} V^{-1/2} X : P_{GLS}^{(1)} V^{-1/2} X_1) \).

More explicit implementation of these test procedures may be found in the simulation and data sections. Section 7 describes additional types of hypothesis tests that are not included in these basic five types.

5 Simulation Study

5.1 Designs and factors

This section investigates the finite sample properties of the estimators and tests proposed in this paper. The underlying model for the simulation study was the three-level school-child-time random intercept model considered in Section 3.1.1. The simulation focuses on the following four factors hypothesized to influence the performance of the estimators and tests: (1) the type of correlated effects and (2) the size of correlated effects, (3) the strength of the instrumental variables and (4) the sample size.

Specifically, the outcome variable was generated as:

\[
y_{sct} = \beta_0 + \beta_{11}X_{sct1}^{(1)} + \beta_{12}X_{sct2}^{(1)} + \beta_{13}X_{sct3}^{(1)} + \beta_{21}X_{sc1}^{(2)} + \beta_{22}X_{sc2}^{(2)} + \beta_{23}X_{sc3}^{(2)} + \beta_{31}X_{s1}^{(3)} + \beta_{32}X_{s2}^{(3)} + \epsilon_{sct}^{(1)} + \epsilon_{sc}^{(2)} + \epsilon_{s}^{(3)}.
\]

There were eight predictors and three random components in the model: \( X_{sct1}^{(1)} \), \( X_{sct2}^{(1)} \), and \( X_{sct3}^{(1)} \) were level-1 predictors, \( X_{sc1}^{(2)} \), \( X_{sc2}^{(2)} \), and \( X_{sc3}^{(2)} \) were level-2 predictors, and \( X_{s1}^{(3)} \) and \( X_{s2}^{(3)} \) were level-3 predictors. The random components at levels 1, 2, and 3 are denoted as \( \epsilon_{sct}^{(1)} \), \( \epsilon_{sc}^{(2)} \), and \( \epsilon_{s}^{(3)} \), respectively. Several model parameters were held constant to keep the simulation manageable. The coefficients \( \beta \) were fixed at negative two, the variances of the random components \( \epsilon \) were all fixed at one, and the constant term \( \beta_0 \) was set to 70.

The variables \( X_{sct3}^{(1)} \) and \( X_{sc3}^{(2)} \) were potentially endogenous whereas the other six variables were kept exogenous and so comprised the set of instruments denoted as \( X_{1,sct} \) (see section 3.2). For the potentially endogenous variables, we first created two latent variables:

\[
\tilde{X}_{sc3}^{(2)} = \gamma(X_{sc1}^{(2)} + X_{sc2}^{(2)} + X_{sc3}^{(2)} + X_{s1}^{(3)}) + \zeta_{sc},
\]

\[
\tilde{X}_{sct3}^{(1)} = \gamma((X_{sct1}^{(1)} - \tilde{X}_{sct1}^{(2)}) + (X_{sct2}^{(1)} - \tilde{X}_{sct2}^{(2)}) + X_{sc1}^{(2)} + X_{sc2}^{(2)} + X_{s2}^{(3)}) + \zeta_{sct}.
\]

Here, all exogenous variables and random components were generated from independent standard normal distributions. Depending on the size of \( \gamma \)=1, 2, or 3, the collection of variables in \( X_{1,sct} \)
become weak, moderate, or strong instruments. Then, to simulate correlated effects, we set \( X_{sc3}^{(1)} = \tilde{X}_{sc3}^{(1)} + \eta_{sc} \epsilon_{sc}^{(2)} \) and \( X_{sc3}^{(2)} = \tilde{X}_{sc3}^{(2)} + \eta_{s} \epsilon_{s}^{(3)} \). For each potential correlated effect, the corresponding coefficient \( \eta \) was set to be 0, 1, 2, or 3, allowing us to vary the size of the correlated effects. The values of 0, 1, 2, and 3 for \( \eta \) and \( \gamma \) indicate that the size of correlated effects and the strength of the instruments are none, smaller than, the same as, and larger than the effect of the exogenous predictors in the models (\(|\beta|\leq 2\)), respectively.

The first factor in our simulation study represents the type of correlated effects. We used four designs for this factor. In Design 1, there was no endogenous variable so that \( \eta_{sc} = \eta_{s} = 0 \) and \( E(\mathbf{X}\delta) = 0 \), where \( \mathbf{X} = (\mathbf{X}_{sc3}^{(1)}|\mathbf{X}_{sc}^{(2)}|\mathbf{X}_{s}^{(3)}) \) and \( \delta = \epsilon_{sc}^{(1)} + \epsilon_{sc}^{(2)} + \epsilon_{s}^{(3)} \). In contrast, \( X_{sc3}^{(1)} \) is endogenous in Design 2 (\( \eta_{sc} > 0 \)) and so \( E(X_{sc3}^{(1)}e_{sc}^{(2)}) \neq 0 \). Similarly, \( X_{sc3}^{(2)} \) is endogenous in Design 3 (\( \eta_{s} > 0 \)), and both \( X_{sc3}^{(1)} \) and \( X_{sc3}^{(2)} \) are endogenous in Design 4. The second factor is the size of the correlated effects, considered at levels of none, small, medium, and large (\( \eta=0, 1, 2, 3 \)). The third factor is the strength of instrumental variables considered at levels of none, weak, moderate, and strong (\( \gamma=0, 1, 2, 3 \)). As a fourth factor, different combinations considered of sample sizes \( (nS, nC, nT) \) were considered, where \( nS \) indicates the number of schools, \( nC \) indicates the number of children per school, and \( nT \) indicates the number of time points per child. The values of \( nS, nC, \) and \( nT \) are chosen to cover the range where they influence bias and power most significantly. The combinations were \((15, 10, 3), (18, 10, 3), (17, 10, 4), (15, 15, 4), (18, 15, 4), (18, 20, 5), (17, 15, 5), (15, 20, 5), (17, 20, 5), (18, 20, 6), \) and \((20, 20, 6)\), resulting in ten different total sample sizes between 450 and 2,400. Therefore, there were a total of \((design\times correlated\ effects\ size\times strength\ of\ instrument\times sample\ size)=4\times 4\times 3\times 10=480 \) conditions in this study. For each condition, 1,000 datasets were generated.

We examined the following five estimators: the fixed child effects estimator (\( b_{FEc} \)), the fixed school effects estimator (\( b_{FEs} \)), the GMM child effects estimator (\( b_{GMMc} \)), the GMM school effects estimator (\( b_{GMMs} \)), and the random effects estimator (\( b_{RE} \)). Our main interest was to see how these estimators behave in the presence of different correlated effects. We focused on \( \beta_{13} \) and \( \beta_{23} \), the coefficients corresponding to the two potentially endogenous variables \( X_{sc3}^{(1)} \) and \( X_{sc3}^{(2)} \). We refer to the estimates of \( \beta_{13} \) and \( \beta_{23} \) as \( b_{13} \) and \( b_{23} \).

Theoretically, the order of robustness among the estimators from the most to least robust is \( b_{FEc}, b_{GMMc}, b_{FEs}, b_{GMMs}, \) and \( b_{RE} \), but relative distances among the estimators is unknown.\(^{1}\) Kim and Frees (2006) explored two fixed school effects estimators with and without a child-level random component \( \epsilon_{sc}^{(2)} \) (GLS vs. OLS). They found that the GLS version consistently produces much smaller bias and RMSE than the OLS version, and concluded that the former should be preferred. In this paper, \( b_{FEs} \) refers to the GLS version with \( \epsilon_{sc}^{(2)} \).

\(^{1}\)Kim and Frees (2006) explored two fixed school effects estimators with and without a child-level random component \( \epsilon_{sc}^{(2)} \) (GLS vs. OLS). They found that the GLS version consistently produces much smaller bias and RMSE than the OLS version, and concluded that the former should be preferred. In this paper, \( b_{FEs} \) refers to the GLS version with \( \epsilon_{sc}^{(2)} \).
The large-sample properties of the estimators are summarized at the top of Table 1. For example, the most robust estimator $b_{FEc}$ would be unbiased across all conditions, while the most efficient estimator $b_{RE}$ would be biased unless the exogeneity condition holds. Note that there are only four estimators at level 2, because $b_{FEc}$ cannot be estimated at level 2.

5.2 Simulation results

The average bias of the estimators is shown at the bottom of Table 1 as a function of $\eta$, the size of the correlated effects. The simulation confirmed the theoretical order among the five estimators on a robust to efficient continuum. Moreover, it was found that the first two estimators behave similarly while the other three are close to each other. As a result, the five estimators seemed to form two clusters: $(b_{FEc}, b_{GMMc})$ and $(b_{FEs}, b_{GMMs}, b_{RE})$.

As anticipated, only $b_{FEc}$ was unbiased across all types of correlated effects. Its GMM counterpart, $b_{GMMc}$, was also unbiased in most conditions but showed a small bias at level 2 in Design 3 and in Design 4 when $\eta=3$ (strong the correlated effects). The other fixed effects estimator $b_{FEs}$ was biased in Design 2, even though there were no correlated effects at the school level. This demonstrates an asymmetry between higher- and lower-level fixed effects estimators (Kim & Frees, 2005, 2006) in that lower-level fixed effects estimators are robust to higher-level correlated effects ($b_{FEc}$ is unbiased in Design 3), but the higher-level fixed effects estimators are not robust to lower-level correlated effects ($b_{FEs}$ is biased in Design 2).

Asymptotically, the GMM estimators share the same properties of corresponding fixed effects estimators (e.g., $b_{GMMc}$ to $b_{FEc}$ and $b_{GMMs}$ to $b_{FEs}$), provided the instruments are relevant. However, in our finite sample study, $b_{GMMs}$ was biased in Design 3, especially at level 2, whereas $b_{FEs}$ was not. Nonetheless, the bias of $b_{GMMs}$ was considerably smaller than the bias of $b_{RE}$ and a close investigation indicated that the bias in $b_{GMMs}$ was minimal when the correlated effect was small ($\eta = 1$). Moreover, as displayed in Figure 1, strong instruments substantially reduced the biases even when correlated effects were large. An increasing number of children also lessened the bias of $b_{GMMs}$ at the child level.

Also as anticipated, the most efficient estimator $b_{RE}$ was unbiased only in Design 1. Correlated effects at one level (Designs 2 & 3) resulted in bias at both levels, although the size of bias was larger at the level of the correlated effects. The size of the bias in $b_{RE}$ was comparable to $b_{FEs}$ and $b_{GMMs}$ in Design 2, because all three estimators were sensitive to child-level correlated effects. However, the bias in $b_{RE}$ was much larger than $b_{FEs}$ and $b_{GMMs}$ in Designs 3 and 4, because $b_{FEs}$ and also $b_{GMMs}$ to some degree were robust to school-level correlated effects.

Of course, bias is only important if it is large relative to the overall variability of the estimator.
To investigate the relative size, Table 2 displays the root mean square errors (RMSEs) of the estimators. In contrast to bias, where the size of correlated effects is most important, sample size $N = nS \times nC \times nT$ was the dominating factor with respect to RMSE. RMSEs were in general smaller at level 1 than level 2 (where the number of observations is smaller), and decreased as the sample size increased. When none of estimators were biased (Design 1), the two most robust estimators, $b_{FEc}$ and $b_{GMMc}$ had larger RMSEs than the other three, confirming that $b_{RE}$ and also $b_{FEs}$ and $b_{GMMs}$ are more efficient in the absence of correlated effects.

The theory of this paper develops regression coefficient estimators that are robust to the presence of correlated effects. It is also of interest to note the effect of correlated effects on estimates of second moment parameters. Specifically, Table 3 shows that the random components estimators were also affected by correlated effects. Here, the variances of the three random components, $\epsilon_{sct}^{(1)}$, $\epsilon_{sc}^{(2)}$, and $\epsilon_{s}^{(3)}$ were estimated via maximum likelihood, under the usual model of no omitted effects. When there were no correlated effects, none were biased, and the average RMSEs for the level 1, 2, and 3 random components were 0.05, 0.11, and 0.41, respectively, indicating that variability is larger at higher levels where the number of observations is smaller.

However, when there existed correlated effects, variance components were severely underestimated because they were estimated based on misspecified models. Multiple correlated effects (Design 4) worsen the estimation of the random components. Both bias and RMSE increased as the size of correlated effects became larger. An extreme case was found with respect to $\epsilon_{sc}^{(2)}$ in Designs 2 and 4 with $\eta=3$, where the estimated variance component was only 7% of its actual size.

The underestimation of variance components results in deceptively large test statistics in omitted variable tests. On the other hand, empirical standard errors (section 3.3) are more robust against model misspecification. The rejection rates of tests comparing $b_{GMMc}$ with $b_{RE}$ and $b_{GMMs}$ with $b_{RE}$ are summarized in Table 4. Note that, unlike tests between fixed and random effects estimators, the degrees of freedom is not simply the number of predictors compared, but $\text{rank}(\Omega)$ (see Proposition 3) and this rank may vary across samples. Therefore, we report the average over the 1,000 replications.

In the comparison of $b_{GMMc}$ and $b_{RE}$, some $\chi^2$'s based on model-based errors are more than ten times larger than those based on empirical standard errors. The average test statistics in Design 2 were 174.21 (model based) versus 15.67 (empirical) and 162.16 (model based) versus 15.49 (empirical) in Design 4. The power of the test $b_{GMMc}$ versus $b_{RE}$ based on model-based standard errors was always one in Designs 2 and 4, but it is difficult to interpret this extreme power because of the downward bias in variance estimates. Therefore, we focused on the corresponding power
based on empirical standard errors, which was also very high (over 0.95 on average). This indicates that \( b_{GMMc} \) performs significantly better than \( b_{REE} \) in the presence of correlated effects at the child level. Note that the power of this test reduced slightly when \( \eta = 3 \) because, as is shown in Table 1, \( b_{GMMc} \) is moderately biased when the correlated effects are strong. Although \( b_{GMMs} \) and \( b_{RE} \) are both biased in Design 4, the bias was considerably smaller for \( b_{GMMs} \) than \( b_{RE} \), resulting in 0.54 to 0.65 rejection rates for the test between the two estimators.

The bias of \( b_{RE} \) in the presence of correlated effects at the school level was found to be smaller than the bias with correlated effects at the child level (Table 1). Consequently, test statistics as well as power were smaller in Design 3 compared to Designs 2 and 4. When there exists correlated effects at the school level but not at the child level, \( b_{GMMs} \) appears to provide a more powerful test against \( b_{RE} \) than \( b_{GMMc} \), because both GMM estimators are asymptotically unbiased and \( b_{GMMs} \) is more efficient.

In summary, the simulation study confirms the order of robustness (bias) and efficiency (RMSE) among the five estimators from \( b_{FEc} \) to \( b_{RE} \). It was found that \( b_{GMMc} \) is almost as robust as \( b_{FEc} \) without loss of information, and thus would be a great alternative when correlated effects are suspected, especially at the child level. Although the other GMM estimator \( b_{GMMs} \) was not robust to child-level omitted effects, bias in \( b_{GMMs} \) was considerably smaller than the standard \( b_{RE} \), and the performance of \( b_{GMMs} \) can be substantially improved by strong instruments and large sample size. Also, when \( b_{GMMs} \) is unbiased, because it has smaller RMSE and requires a smaller number of parameters for estimation, \( b_{GMMs} \) provides a more powerful test against \( b_{RE} \) than \( b_{GMMc} \).

The simulation study also found that not only regression coefficients but also variance component estimation was affected by correlated effects. Specifically, higher level variance components were severely underestimated in the presence of correlated effects and would yield misleading inferences such as falsely significant effects of predictors and deceivingly large test statistics when model-based standard errors are used. To overcome these consequences of model misspecification, we recommend empirical standard errors instead of model-based standard errors. In the next section, our analysis will be based on empirical standard errors.

6 Analysis of Student Achievement Data

This section completes the data analysis introduced in Section 1.4. The analysis in Kim and Frees (2006) showed a large discrepancy between robust and efficient estimators, suggesting that only the fixed child effects estimator, \( b_{FEc} \), provides reliable solutions. Although \( b_{FEc} \) offered unbiased estimates for the effects of time-varying variables, the analysis did not provide estimates for the
effects of the child- or school-level variables due to the fact that fixed effects estimators sweep out all higher level effects.

In searching for unbiased estimates of the effects of all variables in the model, we used the new GMM estimators. This section illustrates how we used the GMM technique without additional variables outside the model (i.e., external instruments), instead using functional forms of variables already in the model (i.e., internal instruments).

6.1 Omitted variable tests in multilevel models

Kim and Frees (2006) conducted two omitted variable tests using two fixed effects estimators and a random effects estimator. First, they tested the existence of omitted child and school effects simultaneously by comparing the fixed child effects estimator $b_{FEc}$ and the random effects estimator $b_{RE}$. They referred to this test as a “multiple-level” test. The test indicated severe bias in the random effects estimator.

Second, they tested omitted child effects separately (regardless of omitted school effects) by comparing the fixed child effects estimator $b_{FEc}$ and the fixed school effects estimator $b_{FEs}$. When comparing the two fixed effects estimators, the test was referred to as an “intermediate” level test. The results of the test implied that $b_{FEs}$ is biased due to omitted child effects. A rejection at the intermediate level implies that further testing of omitted school effects separately (referred to as a “single-level” test) is not meaningful, as the validity of the single-level test depends on the assumption of no omitted effects at lower levels (Kim & Frees, 2006). Consequently, $b_{FEc}$ was the only unbiased estimator among the three estimators.

6.2 Robust inference using GMM estimation

To deal with omitted effects using the GMM technique, we selected exogenous instruments $X_1$ among the model variables. Prior literature in the field suggests that RETAINED and SWITCHING SCHOOLS may be correlated with omitted child and family characteristics (Anderson, Whipple, & Jimerson, 2002; Hanushek, Kain, & Rivkin, 2004). Also, there was concern that mathematics achievement scores in one period may influence values of SCHOOL FREE LUNCH RATE in subsequent periods, and thus this time-varying variable was not assumed to be exogenous. On the other hand, the variable GRADE was used for the random slope and thus was assumed to be exogenous (see the section 3.4 discussion). Also, the level-2 and level-3 variables were assumed to be exogenous.

Note that, unlike two-level models, within and between cluster variability can be defined at different levels in three or higher level models. Specifically, within and between children variability
were used to obtain \( b_{\text{GMMc}} \), and within and between school variability were used to obtain \( b_{\text{GMMs}} \). Thus, we define \( b_{\text{GMMc}} \) (\( b_{\text{GMMs}} \)) to be the estimator from equation (14) when using the child (school) level fixed effects transformation and \( X_1 \) as exogenous variables. Both \( b_{\text{GMMc}} \) and \( b_{\text{GMMs}} \) are optimal GMM estimators using \( W = V^{-1/2} \) as weights (equation 22).

### 6.3 Comparing estimators

Therefore, we consider two GMM estimators in addition to the two fixed effects estimators and the random effects estimator considered in Kim and Frees (2006). These five estimators can be ordered from most robust to most efficient (least robust) as follows: \( b_{\text{FEc}} \), \( b_{\text{GMMc}} \), \( b_{\text{FEs}} \), \( b_{\text{GMMs}} \), and \( b_{\text{RE}} \). These estimators are summarized with empirical standard errors in Table 5. Because \( b_{\text{RE}} \) and \( b_{\text{FEs}} \) both suffer from omitted variable bias (section 6.1), we base our considerations on \( b_{\text{FEc}} \) and its GMM counterpart, \( b_{\text{GMMc}} \).

Examining \( b_{\text{FEc}} \) and \( b_{\text{GMMc}} \), we observe that they each provide the same regression coefficients and standard errors for the level-1 variables RETAINED, SWITCHING SCHOOLS and SCHOOL FREE LUNCH RATE. This was anticipated as we only added GRADE, level-2 and level-3 variables as exogenous variables, so that additional instruments were orthogonal to the child-level fixed effects transformation. A limitation of making this minimalist assumption is that we do not have a sufficient number of degrees of freedom to test whether \( b_{\text{GMMc}} \) is an improvement compared to \( b_{\text{FEc}} \). A strength is that we have enough degrees of freedom to provide consistent estimates of coefficients associated with GRADE as well as the level-2 and level-3 variables.

We next compared \( b_{\text{GMMc}} \) with \( b_{\text{RE}} \). Details of this test are in Table 6. The omnibus test for differences between all variables in \( b_{\text{GMMc}} \) and \( b_{\text{RE}} \) has a \( \chi^2 \) value of 25.679 (\( p=0.019 \)) with 13 degrees of freedom. The individual coefficient tests show large discrepancies between the coefficients of RETAINED and SWITCHING SCHOOLS, suggesting the presence of some important endogeneities with regard to these variables. This is consistent with our speculation based on the previous studies (Anderson, Whipple, & Jimerson, 2002; Hanushek, Kain, & Rivkin, 2004). Although the difference in the coefficients of SCHOOL FREE LUNCH RATE was relatively large, the standard error was too large to show statistical significance. Also, the estimates for CLASS SIZE were too close to zero in both estimators to argue that there is a discrepancy.

We also conducted a test comparing two GMM estimators, \( b_{\text{GMMc}} \) and \( b_{\text{GMMs}} \). The results of the overall test (\( \chi^2=25.965 \), \( df=13 \), \( p=0.017 \)) and the individual coefficient tests (not shown) were very similar to the values in Table 6. This was anticipated because the estimates and standard errors of \( b_{\text{GMMs}} \) and \( b_{\text{RE}} \) were very similar. The discrepancy between the two GMM estimators or, more precisely, the unbiasedness of \( b_{\text{GMMc}} \) and biasedness of \( b_{\text{GMMs}} \), imply that the within person
variability of RETAINED, SWITCH SCHOOLS, and SCHOOL FREE LUNCH RATE seem to be exogenous after controlling for the child and school variables in the model, while the within school variability of these variables appears to still be correlated with child and school effects.

Finally, we examined the sensitivity of GMM estimation by comparing OLS and GLS versions of $b_{\text{GMMc}} (W = I$ and $V^{-1/2}$, respectively). We found that the values of $b_{\text{GMMc}}$ and standard errors were similar with and without weights, suggesting the stability of the GMM estimators regardless of weights.

6.4 Data analysis summary

In this analysis, we used the GMM estimators and successfully obtained unbiased estimates for the effects of all variables in the model in the presence omitted effects in multilevel models. Specifically, we questioned the exogeneity of three variables, RETAINED, SWITCHING SCHOOLS, and SCHOOL FREE LUNCH RATE, and used only the within-child variability of these three variables as instruments while using both the between and within variability of the other variables as instruments. As a result, despite the endogeneities detected in some variables, the GMM estimators provide robust and consistent estimates of all the regression coefficients in $b_{\text{GMMc}}$. This is an important correction to biased random effects estimators $b_{\text{RE}}$ and also a major improvement over fixed effects estimators $b_{\text{FEc}}$ that do not allow estimable coefficients at levels 2 and 3.

More broadly, this study showed that one can obtain multiple GMM estimators with the same set of instruments in multilevel models. The GMM estimators balance the robustness of fixed effects estimators with the efficiency of random effects estimators. When the difference between the fixed and random effects estimators is large, it is useful to have multiple GMM estimators that bridge the two extreme treatments by multiple small steps. In our example, we obtained two GMM estimators $b_{\text{GMMc}}$ and $b_{\text{GMMs}}$. The former was more robust and less efficient than the latter, and it was found that only $b_{\text{GMMc}}$ and not $b_{\text{GMMs}}$ was unbiased in the example. However, if both $b_{\text{GMMc}}$ and $b_{\text{GMMs}}$ had been robust, we would have preferred $b_{\text{GMMs}}$ because it is more efficient.

7 Extensions

In the multilevel data set-up, no instruments from outside the model are required. However, it is important to note that the theory developed in this paper does not restrict one from introducing instruments from outside the model variables; indeed, this is the usual case with instrumental variables estimation. Thus, one may readily use instruments of the form $H = (QX, PX_1, H_2)$, where $H_2$ represents variables not included in the model. Additional instruments are necessary for
handling so-called “level-1 dependencies,” where the disturbance term $\{\epsilon^{(1)}_{sc}\}$ is correlated with one or more model variables. In this case, the sweeping out procedures do not protect us from such dependencies and additional moment assumptions, through additional instruments $H_2$, are needed.

Moreover, analysts need not use only $Q$ matrices that sweep out entire levels. To illustrate, Verbeke, Spiessens and Lesaffre (2001) described observational clinical studies where $t = 1$ corresponds to baseline measurements and the explanatory variables are interpreted in terms of changes from the baseline. Because of the nature of the sampling scheme, it is intuitively plausible to think of unobserved characteristics affecting baseline measurements. Further, they argued that certain subject-specific changes, such as age or time, are strictly exogenous. Analogous to equation (8), this application suggests using a transform such as

$$Q_{1,sc}^{(1)} = \mathbf{I}_{sc} - P(\mathbf{Z}_{2,1,sc}),$$

(24)

where $\mathbf{Z}_{2,1,sc}$ is the first column of $\mathbf{Z}_{2,sc}$ corresponding to the intercept terms. This transform will sweep out intercepts but not slopes associated with $\mathbf{Z}_{2,sc}$. Because $\mathbf{Z}_{2,1,sc}$ is in the subspace generated by $P(\mathbf{Z}_{2,sc})$, it is easy to see from equation (24) that $Q_{1,sc}^{(1)}$ is nested within $Q_{sc}^{(1)}$. Thus, the estimation and hypothesis testing procedures developed in this paper can readily be used with this transform.

One can also extend the procedures in this paper to handle dynamic dependencies, or “feedback” into the system. For example, one can envision a situation where current values of a response, such as test scores, influence future values of covariates, such as grade level. In two-level random intercepts panel data models, Arellano and Bond (1991) showed how to use instrumental variable/GMM estimation techniques to handle these types of endogeneities. Essentially, they proposed using a $Q$ matrix to sweep out individual-specific intercepts and then to use current and past values of covariates as instruments. This seems a natural topic for consideration in our multilevel context, and we propose it as an area for future research.

Another potential extension involves incorporating Bayesian methods to model correlated effects. A variety of simulation techniques, such as Monte Carlo Markov Chain (known by the acronym MCMC), are widely available in several statistical software packages that have made Bayesian techniques easy to implement (for example, Gelman et al., 2004, provide examples of R and BUGs). Bayesian methods have the advantages that (1) they automatically incorporate parameter uncertainty into inference procedures and (2) handle non-Gaussian data equally as well as Gaussian data. Of course, this additional flexibility comes at a price. First, one must specify priors at each model level, and the choice of priors can influence inference results (Natarajan and Kass, 2000). Second, Bayesian techniques require that one explicitly model the correlated effects.
As noted in Section 1, there are several approaches for doing this including, for example, simultaneous equations methods. This approach presumes greater knowledge of the source of correlated effects than assumed in this paper but should also yield more precise inferences. Koop et al. (1997) provide an analysis of a two-level random intercepts model using Bayesian methods.

8 Summary and Conclusions

When explanatory variables are correlated with random components in multilevel models, standard estimation methods do not provide consistent estimates of model parameters. By employing GMM techniques to multilevel models, this paper introduces a new statistical technique that provides consistent estimators in the presence of correlated effects. The GMM estimators in multilevel models provide an overarching framework that includes the well-known fixed effects estimators and random effects estimators as special cases and also provides intermediate options between the two extremes. Also, the GMM estimators can be expressed in terms of familiar instrumental variables, which enhances the interpretability of the procedures. We also provided hypothesis tests to compare these different estimators.

An important advantage of the multilevel modeling GMM estimators is that in many circumstances the instruments are internal so that the researcher is not forced to scurry out and look for useful additional variables that were not involved in the original model formulation. For non-hierarchical data (without replications within clusters), one cannot obtain the GMM estimators without external instruments. Therefore, the ability to utilize internal instruments is a major benefit of multilevel GMM estimators.

Although the idea of using instrumental variable estimators as an intermediate approach between fixed effects estimators and random effects estimators is not new, the current study offers important new results. First, our theory extends the original work of Hausman and Taylor (1981) to the general multilevel framework. While Hausman-Taylor estimators are limited to two-level random-intercept models (i.e., referred to panel data models in economics), the presented GMM estimators are applicable for models with random intercepts and random slopes at multiple levels. Therefore, the GMM estimators are designed to handle a multiple hierarchical structure of data and consequently more complex covariance structures. Equally important, this paper provides a procedure for testing between \( b_{FE} \) and corresponding \( b_{GMM} \) directly as a tool to evaluate \( b_{GMM} \), provided there are enough degrees of freedom for the test (see Section 4), rather than testing the difference between \( b_{GMM} \) and \( b_{RE} \). Although it was commonly used in previous studies, a significant difference between \( b_{GMM} \) and \( b_{RE} \) does not necessarily warrant the consistency of \( b_{GMM} \).
because it is possible that $b_{GMM}$ is still biased, even if to a lesser extent than $b_{RE}$ (e.g., $b_{GMMs}$ in section 6).

The work presented here also extends previous research in Kim and Frees (2005, 2006) that documented systematic ways of handling omitted variables bias through “sweeping out” suspect parts of the model. Although that approach requires relatively weak assumptions, it also precludes researchers from estimating important aspects of the model. In this paper, we show how researchers can estimate all model regression coefficients in the presence of correlated effects by making additional assumptions regarding the exogeneity of certain model variables. Moreover, we provide procedures for checking these additional assumptions.

As a final note, there is a great potential for the GMM estimators for advancing the understanding and practice of multilevel models in the social sciences. In educational research, for example, despite immense research effort and investment, results from educational productivity studies remain inconsistent (Ehrenberg & Brewer, 1994; Goldhaber & Brewer, 1997; Ludwig & Bassi, 1999). There are mixed findings concerning what district, school, teacher, neighborhood, family, and student variables make significant differences on educational outcomes such as improving students’ academic achievement or reducing dropout rates, let alone the size of their effects. This inconsistency is largely due to inherent difficulties in isolating the effects of variables involved in the complex educational process among factors within and outside the school system. Yet, it is often infeasible to design comprehensive experimental studies in school settings due to apparent ethical and/or practical reasons.

Consequently, there is always a danger in education as well as other disciplines in the social sciences that decision making and policy implementation can be based on inaccurate information or in some cases misleading findings. The methodology presented in this paper allows researchers to examine whether and how much the results of a nonexperimental study are valid and trustworthy. Furthermore, this paper provides alternative robust estimators for unbiased and consistent solutions when the standard random effects estimators fail to provide reliable outcomes. In light of the strong interest and growing applications of multilevel models in the social sciences, the GMM estimators can help researchers in the direction of exploiting the rich hierarchical data and, at the same time, impeding the improper use of powerful multilevel models.
References


Rice, N., Jones, A., & Goldstein, H. (1998). Multilevel models where the random effects are correlated with the fixed predictors: a conditioned iterative generalised least squares estimator (CIGLS). York, University of York, Centre for Health Economics.


Appendix

A.1 GMM Estimator

The one-step GMM estimator in equation (5) can be written as $F_{b_{GMM}} = \hat{\Gamma}(H) \frac{1}{n} H^T W y$, where

$$\hat{\Gamma}(H) = \left( \hat{G}^T \hat{M}_{HH} \hat{G} \right)^{-1} \hat{G}^T \hat{M}_{HH}$$

and

$$\hat{G} = \frac{1}{n} H^T W x, \quad \hat{M}_{HH} = \frac{1}{n} H^T H.$$ 

Further, we define

$$\Gamma(H) = \left( G^T M_{HH}^{-1} G \right)^{-1} G^T M_{HH}^{-1}.$$ (25)

Some projections sweep out omitted effects and produce non-unique parameter estimates; thus, we use a matrix $F$ to isolate the unique portion of the estimates. For example, we may take $F$ to be equal to $(I : 0)$ for fixed effects estimators or $F = I$ for instrumental variable estimators.

**Proposition 1** The estimator $F_{b_{GMM}}$ has an asymptotic (as $n \to \infty$) normal distribution with mean $F \beta$ and asymptotic variance $F \Gamma(H) \Lambda \Gamma(H)' F$ where

$$\Lambda = \lim_{n \to \infty} \frac{1}{n} E H^T W V W H.$$ 

**Proof.** With the notational machinery that we have developed, we may now apply standard GMM estimation proofs. That is, with $F_{b_{GMM}} = \hat{\Gamma}(H) \frac{1}{n} H^T W y$, we see that $F_{b_{GMM}}$ is asymptotically equivalent to $F \Gamma(H) \frac{1}{n} H^T W y$, which is a linear combination of independent random vectors. Then, subject to standard moment conditions, we have that $F_{b_{GMM}}$ has an asymptotic normal distribution with asymptotic mean $F \beta$ and variance-covariance matrix $F \Gamma(H) \Lambda \Gamma(H)' F$.

**Proposition 2** Define the residuals $e_i = y_i - X_i b_{GMM}$ and

$$\hat{\Lambda} = \frac{1}{n} \sum_{i=1}^{n} H_i^T W_i e_i e_i^T W_i H_i.$$ (26)

Then, a robust estimator of the asymptotic variance is $F \hat{\Gamma}(H) \hat{\Lambda} \hat{\Gamma}(H)' F$.

Further, $\left( F \hat{\Gamma}(H) \hat{\Lambda} \hat{\Gamma}(H)' F \right)^{-1/2} F (b_{GMM} - \beta)$ has an asymptotic standard normal distribution.

**Proof (Sketch)** We have $b_{GMM} = \hat{\Gamma}(H) \frac{1}{n} H^T W (X \beta + \delta)$ and

$$e_i = y_i - X_i b_{GMM} = \delta_i - X_i (b_{GMM} - \beta) = \delta_i - X_i \left( F \hat{\Gamma}(H) \frac{1}{n} H^T W (X \beta + \delta) - \beta \right) = \delta_i - X_i \hat{\Gamma}(H) \frac{1}{n} H^T W \delta,$$

because $X \left( I - \hat{\Gamma}(H) \frac{1}{n} H^T W X \right) = 0$. Now,

$$\hat{\Lambda} \approx \frac{1}{n} \sum_{i=1}^{n} H_i^T W_i \delta_i \delta_i^T W_i H_i \approx \Lambda.$$ (27)

Thus, $F \hat{\Gamma}(H) \hat{\Lambda} \hat{\Gamma}(H)' F$ is an approximate unbiased estimator of $F \Gamma(H) \Lambda \Gamma(H)' F$, regardless of the heteroscedasticity and/or serial correlation is $V_i$. ■
A.2 Testing moment conditions with nested projections

We now consider two weighted instrumental variable estimators defined in equation (5) with a common weight \( W \). The estimators are thus determined by their instruments; we use the instrument \( H_j \) to determine the estimator \( b_{j,GMM} \), for \( j = 1, 2 \). Moreover, we require that there is a common matrix \( F \) so that \( Fb_{j,GMM} \) is uniquely defined, for \( j = 1, 2 \). To compare these estimators using different instruments, we assume that the projections are nested so that \( P(H_1)P(H_2) = P(H_2) \). Moreover, define \( P_{2 \star} = P(H_1) - P(H_2) \), a projection that is orthogonal to \( P(H_2) \).

For notation, we define

\[
\hat{G}_j = \frac{1}{n} H_j'WX, \quad G_j = \lim_{n \to \infty} \frac{1}{n} H_j'WX, \\
\hat{M}_{jj} = \frac{1}{n} H_j' H_j, \quad M_{jj} = \lim_{n \to \infty} \frac{1}{n} H_j' H_j,
\]

and

\[
\Lambda_{ij} = \lim_{n \to \infty} \frac{1}{n} H_i' W V W H_j, \quad i, j = 1, 2.
\]

**Proposition 3** The difference \( Fb_{2,GMM} - Fb_{1,GMM} \) has an asymptotic normal distribution with mean zero. Further, with \( \Gamma_j = \Gamma(H_j) \), \( j = 1, 2 \), the asymptotic variance is

\[
\Omega = F (\Gamma_2 A_{22} \Gamma_2' + \Gamma_1 A_{11} \Gamma_1' - \Gamma_2 A_{21} \Gamma_1' - \Gamma_1 A_{12} \Gamma_2') F'.
\]  

**Proof.** First note that

\[
F \left( \hat{\Gamma}(H_2) \frac{1}{n} H_2' W - \hat{\Gamma}(H_1) \frac{1}{n} H_1' W \right) X
= F \left( \left( G_2 M_{22} G_2 - \hat{G}_2 \hat{M}_{22} \hat{G}_2 - \left( \hat{G}_1 \hat{M}_{11} \hat{G}_1 \right) \right) \left( \hat{G}_1 \hat{M}_{11} \hat{G}_1 \right) \right) = 0.
\]

Thus, the difference of the estimators can be expressed as

\[
F (b_{2,GMM} - b_{1,GMM}) = F \left( \hat{\Gamma}(H_2) \frac{1}{n} H_2' W - \hat{\Gamma}(H_1) \frac{1}{n} H_1' W \right) \delta.
\]

This is asymptotically equivalent to \( \frac{1}{n} F \left( \Gamma_2 H'_2 - \Gamma_1 H'_1 \right) W \delta' W \) which is a linear combination of independent random vectors. Then, subject to standard moment conditions, we have that \( Fb_{2,GMM} - Fb_{1,GMM} \) has an asymptotic normal distribution. The asymptotic variance is

\[
\frac{1}{n} F \left( \Gamma_2 H'_2 - \Gamma_1 H'_1 \right) W \delta' W \left( H_2 \Gamma'_2 - H_1 \Gamma'_1 \right) F'
= \frac{1}{n} F \left( \Gamma_2 H'_2 W \delta' H_2 \Gamma'_2 + \Gamma_2 H'_2 W \delta' H_1 \Gamma'_1 - \Gamma_1 H'_1 \Gamma'_2 W \delta' H_2 \Gamma'_2 - H_2 \Gamma'_2 W \delta' H_1 \Gamma'_1 \right) F'.
\]

With the definition of \( \Lambda_{ij} \) and the law of large numbers, this is sufficient for the result. \( \blacksquare \)

**Corollary 4** Let \( W = V^{-1/2} \). Then, the asymptotic variance of \( Fb_{j,GMM} \) is \( \Gamma_j A_{jj} \Gamma_j' = \left( G_j M_{jj} G_j \right) \), \( j = 1, 2 \). Further, the asymptotic variance of \( F (b_{2,GMM} - b_{1,GMM}) \) is

\[
F \left( \Gamma_2 A_{22} \Gamma_2' - \Gamma_1 A_{11} \Gamma_1' \right) F'.
\]
Proof. With $W = V^{-1/2}$, we have that $A_{ij} = M_{ij}, j = 1, 2$. Further, we have

$$\Gamma_2 \Lambda_{22} \Gamma' = (G_2 M_{22} G_2)^{-1} G_2 M_{22} M_{22} G_2 (G_2 M_{22} G_2)^{-1} = (G_2 M_{22} G_2)^{-1}$$

and similarly for $j = 1$.

From the requirement that $P(H_1) P(H_2) = P(H_2)$, we have

$$\hat{G}'_1 \hat{M}_{12} \hat{M}_{12} \hat{M}_{22} G_2$$

Thus, by the law of large numbers, we have

$$\Gamma_1 \Lambda_{12} \Gamma' = (G_1 M_{11} G_1)^{-1} G_1 M_{11} M_{12} G_2 (G_2 M_{22} G_2)^{-1} = (G_1 M_{11} G_1)^{-1}.$$

This is sufficient for the proof. $\blacksquare$

**Proposition 5** The test statistic

$$TS = (b_{2,GMM} - b_{1,GMM})' F' \Omega^{-1} F (b_{2,GMM} - b_{1,GMM})$$

has an asymptotic chi-square distribution with degrees of freedom $df = \text{rank}(\Omega)$. Further, if $W = V^{-1/2}$, then the asymptotic degrees of freedom is $df = \text{rank}(X' V^{-1/2} P_{2*1})$.

Proof. By Proposition 3, we have that $TS$ has an asymptotic chi-square distribution, with degrees of freedom $df = \text{rank}(\Omega)$. See for example, Rao (1973, p. 188). With $W = V^{-1/2}$ and the corollary, we have

$$\text{rank}(\Omega)$$

$$= \text{rank} \left\{ F \left( G_2 M_{22} G_2 \right)^{-1} \left[ (G_1 M_{11} G_1)^{-1} \right] F' \right\}$$

$$= \text{rank} \left\{ F \left( G_2 M_{22} G_2 \right)^{-1} \left( G_1 M_{11} G_1 - G_2 M_{22} G_2 \right) (G_1 M_{11} G_1)^{-1} F' \right\}$$

$$= \text{rank} \left\{ G_1 M_{11} G_1 - G_2 M_{22} G_2 \right\}$$

$$= \lim_{n \to \infty} \text{rank} \left\{ \frac{1}{n} X' V^{-1/2} (P(H_1) - P(H_2)) V^{-1/2} X \right\}$$

$$= \lim_{n \to \infty} \text{rank} \left\{ \frac{1}{n} X' V^{-1/2} P_{2*1} V^{-1/2} X \right\} = \lim_{n \to \infty} \text{rank} \left\{ \frac{1}{n} X' V^{-1/2} P_{2*1} \right\},$$

thus establishing the result. $\blacksquare$

**Proposition 6** Define the residuals $e_i = y_i - X_i b_1$ and $\hat{V} = \text{blkdiag}(e_1 e_1', \ldots, e_n e_n')$, $\Lambda_{ij} = \frac{1}{n} H_i' W \tilde{V} W H_j$, $i, j = 1, 2$. Then, a robust estimator of the asymptotic variance of $F (b_{2,GMM} - b_{1,GMM})$ is

$$\hat{\Omega} = F \left( \hat{\Gamma}_2 \hat{\Lambda}_{22} \hat{\Gamma}' + \hat{\Gamma}_1 \hat{\Lambda}_{11} \Gamma' - \hat{\Gamma}_2 \hat{\Lambda}_{21} \hat{\Gamma}' - \hat{\Gamma}_1 \hat{\Lambda}_{12} \hat{\Gamma}' \right) F',$n

where $\hat{\Gamma}_j = \hat{\Gamma}(H_j)$, $j = 1, 2$. The test statistic

$$TS_{\text{Robust}} = (b_{2,GMM} - b_{1,GMM})' F' \hat{\Omega}^{-1} F (b_{2,GMM} - b_{1,GMM})$$

has the same asymptotic chi-square distribution as in Proposition 5.

Proof. Similar to the proof of Proposition 2, and is omitted. $\blacksquare$
Table 1: Properties of estimators and average bias over 1,000 replications

<table>
<thead>
<tr>
<th>Correlated Effects (Design)</th>
<th>( \eta )</th>
<th>( b_{\text{FEc}} )</th>
<th>( b_{\text{GMMCc}} )</th>
<th>( \beta_{13} )</th>
<th>( b_{\text{FEs}} )</th>
<th>( b_{\text{GMMs}} )</th>
<th>( b_{\text{RE}} )</th>
<th>( b_{\text{FEc}} )</th>
<th>( b_{\text{GMMCc}} )</th>
<th>( \beta_{23} )</th>
<th>( b_{\text{FEs}} )</th>
<th>( b_{\text{GMMs}} )</th>
<th>( b_{\text{RE}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( E(\mathbf{X}\delta) = 0 )</td>
<td>0</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>—</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
</tr>
<tr>
<td>2. ( E(\mathbf{X}<em>{\text{sct}}\epsilon</em>{\text{sc}}) \neq 0 )</td>
<td>&gt; 0</td>
<td>UB</td>
<td>UB</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>3. ( E(\mathbf{X}<em>{\text{sc}}\epsilon</em>{s}) \neq 0 )</td>
<td>&gt; 0</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>B</td>
<td>B</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>4. ( E(\mathbf{X}<em>{\text{sct}}\epsilon</em>{\text{sc}}) \neq 0 )</td>
<td>&gt; 0</td>
<td>UB</td>
<td>UB</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>UB</td>
<td>UB</td>
<td>UB</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
</tbody>
</table>

\( \mathbf{X} = (\mathbf{X}_{\text{sct}}| \mathbf{X}_{\text{sc}}| \mathbf{X}_{s}) \). \( \delta = \epsilon^{(1)}_{\text{sct}} + \epsilon^{(2)}_{\text{sc}} + \epsilon^{(3)}_{s} \). \( \eta \): size of correlated effects. UB: unbiased, B: biased.
Table 2: RMSE of the estimators

<table>
<thead>
<tr>
<th>Correlated Effects (Design)</th>
<th>N</th>
<th>nS, nC, nT</th>
<th>( \beta_{13} )</th>
<th>( \beta_{23} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( E(\mathbf{X}\mathbf{d})^* = 0 )</td>
<td>450</td>
<td>15, 10, 3</td>
<td>0.25 0.25 0.21 0.21</td>
<td>0.36 0.25 0.25 0.25</td>
</tr>
<tr>
<td></td>
<td>540</td>
<td>18, 10, 3</td>
<td>0.23 0.23 0.20 0.20</td>
<td>0.33 0.24 0.24 0.24</td>
</tr>
<tr>
<td></td>
<td>680</td>
<td>17, 10, 4</td>
<td>0.19 0.19 0.17 0.17</td>
<td>0.32 0.23 0.23 0.23</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>15, 15, 4</td>
<td>0.17 0.17 0.15 0.15</td>
<td>0.27 0.19 0.19 0.19</td>
</tr>
<tr>
<td></td>
<td>1080</td>
<td>18, 15, 4</td>
<td>0.15 0.15 0.14 0.14</td>
<td>0.25 0.18 0.18 0.18</td>
</tr>
<tr>
<td></td>
<td>1275</td>
<td>17, 15, 5</td>
<td>0.14 0.14 0.13 0.13</td>
<td>0.26 0.18 0.18 0.18</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>15, 20, 5</td>
<td>0.13 0.13 0.12 0.12</td>
<td>0.24 0.17 0.17 0.17</td>
</tr>
<tr>
<td></td>
<td>1700</td>
<td>17, 20, 5</td>
<td>0.12 0.12 0.11 0.11</td>
<td>0.23 0.15 0.15 0.15</td>
</tr>
<tr>
<td></td>
<td>2160</td>
<td>18, 20, 6</td>
<td>0.11 0.11 0.10 0.10</td>
<td>0.23 0.15 0.15 0.15</td>
</tr>
<tr>
<td></td>
<td>2400</td>
<td>20, 20, 6</td>
<td>0.10 0.10 0.10 0.10</td>
<td>0.21 0.14 0.14 0.14</td>
</tr>
<tr>
<td>2. ( E(X_{sc3}\epsilon_{sc}) \neq 0 )</td>
<td>450</td>
<td>15, 10, 3</td>
<td>0.28 0.29 1.38 1.38</td>
<td>0.36 0.79 0.79 0.79</td>
</tr>
<tr>
<td></td>
<td>540</td>
<td>18, 10, 3</td>
<td>0.25 0.26 1.39 1.39</td>
<td>0.33 0.78 0.78 0.78</td>
</tr>
<tr>
<td></td>
<td>680</td>
<td>17, 10, 4</td>
<td>0.22 0.23 1.33 1.33</td>
<td>0.33 0.69 0.69 0.69</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>15, 15, 4</td>
<td>0.19 0.20 1.34 1.34</td>
<td>0.28 0.69 0.69 0.69</td>
</tr>
<tr>
<td></td>
<td>1080</td>
<td>18, 15, 4</td>
<td>0.17 0.18 1.34 1.34</td>
<td>0.25 0.68 0.68 0.68</td>
</tr>
<tr>
<td></td>
<td>1275</td>
<td>17, 15, 5</td>
<td>0.16 0.16 1.29 1.29</td>
<td>0.26 0.61 0.61 0.61</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>15, 20, 5</td>
<td>0.14 0.15 1.30 1.30</td>
<td>0.24 0.61 0.61 0.61</td>
</tr>
<tr>
<td></td>
<td>1700</td>
<td>17, 20, 5</td>
<td>0.14 0.14 1.29 1.29</td>
<td>0.22 0.61 0.61 0.61</td>
</tr>
<tr>
<td></td>
<td>2160</td>
<td>18, 20, 6</td>
<td>0.12 0.12 1.24 1.24</td>
<td>0.22 0.55 0.55 0.55</td>
</tr>
<tr>
<td></td>
<td>2400</td>
<td>20, 20, 6</td>
<td>0.11 0.11 1.24 1.24</td>
<td>0.21 0.54 0.54 0.54</td>
</tr>
<tr>
<td>3. ( E(X_{sc3}\epsilon_{sc}) \neq 0 )</td>
<td>450</td>
<td>15, 10, 3</td>
<td>0.25 0.25 0.22 0.25</td>
<td>0.45 0.30 0.46 0.66</td>
</tr>
<tr>
<td></td>
<td>540</td>
<td>18, 10, 3</td>
<td>0.23 0.23 0.19 0.23</td>
<td>0.41 0.27 0.42 0.66</td>
</tr>
<tr>
<td></td>
<td>680</td>
<td>17, 10, 4</td>
<td>0.19 0.19 0.18 0.20</td>
<td>0.41 0.26 0.41 0.63</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>15, 15, 4</td>
<td>0.17 0.17 0.15 0.16</td>
<td>0.34 0.23 0.34 0.48</td>
</tr>
<tr>
<td></td>
<td>1080</td>
<td>18, 15, 4</td>
<td>0.16 0.16 0.14 0.15</td>
<td>0.32 0.21 0.31 0.47</td>
</tr>
<tr>
<td></td>
<td>1275</td>
<td>17, 15, 5</td>
<td>0.14 0.14 0.13 0.14</td>
<td>0.34 0.21 0.32 0.47</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>15, 20, 5</td>
<td>0.13 0.12 0.12 0.12</td>
<td>0.30 0.20 0.27 0.37</td>
</tr>
<tr>
<td></td>
<td>1700</td>
<td>17, 20, 5</td>
<td>0.12 0.12 0.11 0.11</td>
<td>0.29 0.18 0.26 0.37</td>
</tr>
<tr>
<td></td>
<td>2160</td>
<td>18, 20, 6</td>
<td>0.10 0.10 0.10 0.10</td>
<td>0.29 0.18 0.25 0.38</td>
</tr>
<tr>
<td></td>
<td>2400</td>
<td>20, 20, 6</td>
<td>0.10 0.10 0.09 0.10</td>
<td>0.27 0.17 0.24 0.38</td>
</tr>
<tr>
<td>4. ( E(X_{sc3}\epsilon_{sc}) \neq 0 )</td>
<td>450</td>
<td>15, 10, 3</td>
<td>0.28 0.30 1.39 1.43</td>
<td>0.46 0.89 1.05 1.21</td>
</tr>
<tr>
<td></td>
<td>540</td>
<td>18, 10, 3</td>
<td>0.26 0.28 1.40 1.43</td>
<td>0.42 0.89 1.03 1.22</td>
</tr>
<tr>
<td></td>
<td>680</td>
<td>17, 10, 4</td>
<td>0.22 0.23 1.33 1.35</td>
<td>0.41 0.78 0.92 1.09</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>15, 15, 4</td>
<td>0.19 0.20 1.35 1.36</td>
<td>0.34 0.78 0.87 0.98</td>
</tr>
<tr>
<td></td>
<td>1080</td>
<td>18, 15, 4</td>
<td>0.17 0.18 1.35 1.36</td>
<td>0.31 0.78 0.86 0.99</td>
</tr>
<tr>
<td></td>
<td>1275</td>
<td>17, 15, 5</td>
<td>0.16 0.16 1.29 1.30</td>
<td>0.32 0.69 0.76 0.87</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>15, 20, 5</td>
<td>0.14 0.15 1.30 1.31</td>
<td>0.29 0.69 0.74 0.82</td>
</tr>
<tr>
<td></td>
<td>1700</td>
<td>17, 20, 5</td>
<td>0.13 0.14 1.30 1.31</td>
<td>0.27 0.62 0.66 0.74</td>
</tr>
<tr>
<td></td>
<td>2160</td>
<td>18, 20, 6</td>
<td>0.12 0.12 1.25 1.25</td>
<td>0.27 0.62 0.66 0.74</td>
</tr>
<tr>
<td></td>
<td>2400</td>
<td>20, 20, 6</td>
<td>0.11 0.11 1.25 1.25</td>
<td>0.26 0.62 0.65 0.74</td>
</tr>
</tbody>
</table>
Table 3: Bias and RMSE of the variance of the random components

<table>
<thead>
<tr>
<th>Correlated Effects (Design)</th>
<th>( \eta )</th>
<th>bias ( \epsilon_{act}^{(1)} )</th>
<th>bias ( \epsilon_{sc}^{(2)} )</th>
<th>bias ( \epsilon_{s}^{(3)} )</th>
<th>RMSE ( \epsilon_{act}^{(1)} )</th>
<th>RMSE ( \epsilon_{sc}^{(2)} )</th>
<th>RMSE ( \epsilon_{s}^{(3)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( E(\mathbf{X}\delta)^* = 0 )</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
<td>0.41</td>
</tr>
<tr>
<td>2. ( E(\mathbf{X}<em>{act}\epsilon</em>{sc}) \neq 0 )</td>
<td>1</td>
<td>0.04</td>
<td>-0.33</td>
<td>0.00</td>
<td>0.07</td>
<td>0.35</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.11</td>
<td>-0.82</td>
<td>0.00</td>
<td>0.12</td>
<td>0.82</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.08</td>
<td>-0.93</td>
<td>0.00</td>
<td>0.10</td>
<td>0.93</td>
<td>0.38</td>
</tr>
<tr>
<td>3. ( E(\mathbf{X}<em>{sc}\epsilon</em>{s}) \neq 0 )</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.06</td>
<td>0.05</td>
<td>0.12</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.27</td>
<td>0.05</td>
<td>0.12</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.00</td>
<td>0.04</td>
<td>-0.60</td>
<td>0.05</td>
<td>0.13</td>
<td>0.70</td>
</tr>
<tr>
<td>4. ( E(\mathbf{X}<em>{act}\epsilon</em>{sc}) \neq 0 )</td>
<td>1</td>
<td>0.04</td>
<td>-0.34</td>
<td>-0.26</td>
<td>0.07</td>
<td>0.36</td>
<td>0.41</td>
</tr>
<tr>
<td>E(( \mathbf{X}<em>{sc}\epsilon</em>{s} ) \neq 0)</td>
<td>2</td>
<td>0.11</td>
<td>-0.82</td>
<td>-0.65</td>
<td>0.12</td>
<td>0.82</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.08</td>
<td>-0.93</td>
<td>-0.82</td>
<td>0.10</td>
<td>0.93</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 4: Rejection rates

<table>
<thead>
<tr>
<th>Correlated Effects (Design)</th>
<th>( \eta )</th>
<th>b_{GMMc} vs. b_{RE}</th>
<th>b_{GMMc} vs. b_{RE}</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E(\mathbf{X}\delta)^* = 0 )</td>
<td>1</td>
<td>( \chi^2 )</td>
<td>df</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>187.03</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>200.43</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>135.16</td>
<td>6.00</td>
</tr>
<tr>
<td>( E(\mathbf{X}<em>{act}\epsilon</em>{sc}) \neq 0 )</td>
<td>( \chi^2 )</td>
<td>df</td>
<td>( \chi^2 )</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>6.24</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.26</td>
<td>5.97</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9.40</td>
<td>5.54</td>
</tr>
<tr>
<td>( E(\mathbf{X}<em>{sc}\epsilon</em>{s}) \neq 0 )</td>
<td>( \chi^2 )</td>
<td>df</td>
<td>( \chi^2 )</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>185.84</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>187.73</td>
<td>5.98</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>112.90</td>
<td>5.68</td>
</tr>
</tbody>
</table>

\( df \): the average degrees of freedom over 1,000 replications, \( p \): the average rejection rate, \( SE \): standard errors
Table 5: Five estimators from most robust to most efficient. Empirical standard errors in parens.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$b_{FEc}$</th>
<th>$b_{GMMc}$</th>
<th>$b_{FEs}$</th>
<th>$b_{GMMs}$</th>
<th>$b_{RE}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE</td>
<td>–</td>
<td>3.387 (0.20)</td>
<td>3.368 (0.21)</td>
<td>3.358 (0.21)</td>
<td>3.375 (0.21)</td>
</tr>
<tr>
<td>RETAINED</td>
<td>14.554 (2.21)</td>
<td>14.554 (2.21)</td>
<td>9.338 (2.70)</td>
<td>9.333 (2.70)</td>
<td>9.205 (2.73)</td>
</tr>
<tr>
<td>SWITCH SCHOOLS</td>
<td>-1.544 (1.04)</td>
<td>-1.544 (1.04)</td>
<td>-0.301 (0.76)</td>
<td>-0.297 (0.76)</td>
<td>-0.365 (0.75)</td>
</tr>
<tr>
<td>SCHOOL FREE</td>
<td>1.994 (2.35)</td>
<td>1.994 (2.35)</td>
<td>0.583 (1.22)</td>
<td>0.333 (1.20)</td>
<td>-0.227 (1.22)</td>
</tr>
<tr>
<td>LUNCH RATE</td>
<td>–</td>
<td>-1.246 (0.78)</td>
<td>-1.200 (0.82)</td>
<td>-1.231 (0.78)</td>
<td>-1.234 (0.78)</td>
</tr>
<tr>
<td>FEMALE</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MALE</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AFRICAN AMERICAN</td>
<td>–</td>
<td>-4.705 (1.92)</td>
<td>-5.680 (1.80)</td>
<td>-4.786 (1.85)</td>
<td>-4.745 (1.84)</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>–</td>
<td>-3.971 (1.97)</td>
<td>-3.213 (1.93)</td>
<td>-3.677 (1.92)</td>
<td>-3.608 (1.91)</td>
</tr>
<tr>
<td>OTHER</td>
<td>–</td>
<td>6.704 (1.78)</td>
<td>4.783 (1.71)</td>
<td>6.472 (1.72)</td>
<td>6.526 (1.72)</td>
</tr>
<tr>
<td>CAUCASIAN</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>COHORT</td>
<td>–</td>
<td>1.455 (0.19)</td>
<td>1.503 (0.20)</td>
<td>1.483 (0.19)</td>
<td>1.497 (0.19)</td>
</tr>
<tr>
<td>P.MINORITY</td>
<td>–</td>
<td>-0.163 (0.09)</td>
<td>–</td>
<td>-0.128 (0.07)</td>
<td>-0.116 (0.07)</td>
</tr>
<tr>
<td>TEACHER EXPERIENCE</td>
<td>–</td>
<td>0.158 (0.17)</td>
<td>–</td>
<td>0.155 (0.17)</td>
<td>0.157 (0.17)</td>
</tr>
<tr>
<td>CLASS SIZE</td>
<td>–</td>
<td>0.088 (0.17)</td>
<td>–</td>
<td>0.071 (0.17)</td>
<td>0.069 (0.17)</td>
</tr>
</tbody>
</table>

Table 6: Comparing child-level IV estimator and random effects estimator

<table>
<thead>
<tr>
<th>Variable</th>
<th>$b_{GMMc}$</th>
<th>$b_{RE}$</th>
<th>Difference</th>
<th>Std Err</th>
<th>Individual coeff. test($\chi^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE</td>
<td>3.387</td>
<td>3.375</td>
<td>0.013</td>
<td>0.050</td>
<td>0.063, $p = 0.802$</td>
</tr>
<tr>
<td>RETAINED</td>
<td>14.554</td>
<td>9.205</td>
<td>5.349</td>
<td>2.480</td>
<td>4.654, $p = 0.031$</td>
</tr>
<tr>
<td>SWITCH SCHOOLS</td>
<td>-1.544</td>
<td>-0.365</td>
<td>-1.179</td>
<td>0.633</td>
<td>3.466, $p = 0.063$</td>
</tr>
<tr>
<td>SCHOOL FREE</td>
<td>1.994</td>
<td>-0.227</td>
<td>2.221</td>
<td>1.849</td>
<td>1.443, $p = 0.230$</td>
</tr>
<tr>
<td>LUNCH RATE</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-1.246</td>
<td>-1.234</td>
<td>-0.012</td>
<td>0.040</td>
<td>0.090, $p = 0.765$</td>
</tr>
<tr>
<td>MALE</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AFRICAN AMERICAN</td>
<td>-4.705</td>
<td>-4.745</td>
<td>0.040</td>
<td>0.183</td>
<td>0.047, $p = 0.827$</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>-3.971</td>
<td>-3.608</td>
<td>-0.363</td>
<td>0.274</td>
<td>1.759, $p = 0.185$</td>
</tr>
<tr>
<td>OTHER</td>
<td>6.704</td>
<td>6.526</td>
<td>0.177</td>
<td>0.256</td>
<td>0.478, $p = 0.489$</td>
</tr>
<tr>
<td>CAUCASIAN</td>
<td>1.455</td>
<td>1.497</td>
<td>-0.042</td>
<td>0.046</td>
<td>0.823, $p = 0.364$</td>
</tr>
<tr>
<td>COHORT</td>
<td>-0.163</td>
<td>-0.116</td>
<td>-0.047</td>
<td>0.040</td>
<td>1.394, $p = 0.238$</td>
</tr>
<tr>
<td>P.MINORITY</td>
<td>0.158</td>
<td>0.157</td>
<td>0.001</td>
<td>0.008</td>
<td>0.017, $p = 0.896$</td>
</tr>
<tr>
<td>CLASS SIZE</td>
<td>0.088</td>
<td>0.069</td>
<td>0.019</td>
<td>0.010</td>
<td>3.405, $p = 0.065$</td>
</tr>
</tbody>
</table>

$\chi^2 = 25.679$, $df = 13$, $p = 0.019$