Controlling for Student Heterogeneity in Longitudinal Achievement Models

J.R. LOCKWOOD AND DANIEL F. MCCAFFREY

WR-471-IES
February 2007
Controlling for Student Heterogeneity in Longitudinal Achievement Models

J.R. Lockwood and Daniel F. McCaffrey
February 14, 2007

Abstract

Research and policies concerning primary and secondary school education in the United States are increasingly focused on student achievement test scores, in part because of the rapidly growing availability of data tracking student scores over time. These longitudinal data are currently being used to measure school and teacher performance, as well as to study the impacts of teacher qualifications, teaching practices, school choice, school reform, charter schools and other educational interventions. Longitudinal data are highly valued because they offer analysts possible controls for unmeasured student heterogeneity in test scores that might otherwise bias results. Two approaches are widely used to control for this student heterogeneity: fixed effects models, which condition on the means of the individual students and use ordinary least squares to estimate model parameters; and random effects or mixed models, which treat student heterogeneity as part of the model error term and use generalized least squares for estimation. The usual criticism of the mixed model approach is that correlation between the unobserved student effects and other educational variables in the model can lead to biased and inconsistent parameter estimates, whereas under the same assumptions, the fixed effects approach does not suffer these shortcomings. This paper examines this criticism in the context of longitudinal student achievement data, where the complexities of standardized test scores may create conditions leading to bias in fixed effect estimators. We show that under a general model for student heterogeneity in observed test scores, the mixed model approach can have a certain “bias compression” property that can effectively safeguard against bias due to uncontrolled student heterogeneity, even in cases in which fixed effects models may lead to inconsistent estimates. We present several examples with simulated data to investigate the practical implications of our findings for educational research and practice.

This material is based on work supported by the Department of Education Institute of Education Sciences under Grant No. R305U040005 and by RAND. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations.
1 Introduction

The rapidly growing availability of data tracking student achievement over time has made longitudinal data analysis an increasingly prominent tool for education research. Longitudinal data analysis is now common practice in research on identifying effective teaching practices, measuring the impacts of teacher credentialing and training, and evaluating other educational interventions (Bifulco & Ladd, 2004; Gill, Zimmer, Christman, & Blanc, 2007; Goldhaber & Anthony, 2004; Hanushek, Kain, & Rivkin, 2002; Harris & Sass, 2006; Le, Stecher, Lockwood, Hamilton, Robyn, Williams, Ryan, Kerr, Martinez, & Klein, 2006; Schacter & Thum, 2004; Zimmer, Buddin, Chau, Daley, Gill, Guarino, Hamilton, Krop, McCaffrey, Sandler, & Brewer, 2003). Recent computational advances and empirical findings about the impacts of individual teachers have also intensified interest in “value-added” methods (VAM), where the longitudinal trajectories of students’ test scores are used to estimate the contributions of individual teachers or schools to student achievement (Ballou, Sanders, & Wright, 2004; Braun, 2005; Jacob & Lefgren, 2006; Kane, Rockoff, & Staiger, 2006; Lissitz, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders, Saxton, & Horn, 1997). With teacher and school accountability at the forefront of education policy, and with educators and researchers seeking more sophisticated ways of putting test score data to good use, longitudinal methods are likely to remain critical to education research.

One of the most important attributes of longitudinal data for education research is that they can lead to less biased and more precise estimates of the effects of educational variables (e.g. measurable teacher characteristics) on student achievement than is generally possible with purely cross-sectional observational data. Researchers have consistently found that observable background characteristics of students typically available in administrative databases, such as race/ethnicity, socio-economic status indicators, and limited English proficiency status, are correlated with student achievement, but that there remains substantial unexplained heterogeneity among students in achievement profiles after accounting for these observable characteristics (Harris & Sass, 2006). In observational studies with cross-sectional data, this unmeasured heterogeneity threatens to bias estimates of the effects of educational variables being studied because of non-random allocation of students to educational settings. However, the repeated measures on individual students inherent to longitudinal achievement data provide opportunities to control for this unmeasured heterogeneity, thereby improving the quality of the estimates.

Recently, there has been a growing divide in the educational research literature regarding the statistical methods used to control for unmeasured student heterogeneity in longitudinal analyses. While linear statistical models underlie the majority of empirical analyses, educational statisticians and economists have taken different approaches to implementing these models. The difference primarily involves whether model parameters quantifying student heterogeneity are treated as fixed parameters of the model mean structure (e.g. by using student fixed effects), or as part of the model error structure (e.g. by using student random effects or mixed models). Econometricians
primarily have used variations on the fixed effects approach (Gill et al., 2007; Goldhaber & Anthony, 2004; Harris & Sass, 2006; Koedel & Betts, 2005; Rivkin, Hanushek, & Kain, 2005; Todd & Wolpin, 2004; Zimmer et al., 2003), while educational statisticians more commonly use variations on the random effects or mixed model approaches by way of hierarchical linear models and related methods (Le et al., 2006; May, Supovitz, & Perda, 2004; Raudenbush & Bryk, 2002; Rowan, Correnti, & Miller, 2002; Zvoch & Stevens, 2003). These two distinct approaches to longitudinal modeling have evolved somewhat independently in the educational research literature, and there has not been much communication between users of the different approaches. For example, the divide is evident in the recent literature on value-added methods for estimating individual teacher effects, where econometric analyses have used student (and often teacher) fixed effects (Harris & Sass, 2006; Koedel & Betts, 2005; Rockoff, 2004) while the most prominent methods in the educational statistics literature treat student heterogeneity as part of the model error structure (Ballou et al., 2004; Lockwood, McCaffrey, Mariano, & Setodji, 2006; Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez, 2007; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Nye, Konstantopoulos, & Hedges, 2004; Raudenbush & Bryk, 2002; Sanders et al., 1997).

The usual criticism of the random effects or mixed model approach is that when treating student heterogeneity as part of the model error term, correlation between the unobserved student effects and other educational variables in the model can lead to biased and inconsistent estimates of the effects of those variables, while fixed effects approaches do not suffer these same shortcomings (Frees, 2004; Harris & Sass, 2006; Wooldridge, 2001). The goal of this paper is to examine this criticism in the context of longitudinal student achievement data, where the complexities of standardized test scores may introduce tradeoffs of the methods that require additional consideration. We begin by defining the standard model with time-invariant student-specific parameters in Section 2. We compare the fixed and random effects estimators under this model, review existing results, and develop new results about the conditions under which the two approaches will lead to similar estimates of the effects of educational variables. In Section 3 we then generalize the simple student effects model to a general mixed model and show how, under mild assumptions, the mixed model approaches have a certain “bias compression” property that can provide effective safeguards against bias due to uncontrolled student heterogeneity, even in cases where fixed effects models might be less appropriate. We use this framework to show that under some plausible models for student heterogeneity, the fixed effects approaches can be inconsistent, while the random effects approaches can be lead to approximately correct inferences, as the number of test scores available for each student increases. We provide empirical examples in Section 4, and conclude with a discussion of practical considerations and suggestions for future research in Section 5.
2 Fixed and Random Effects Estimators in the Standard Case

In this section we consider fixed and random effects estimators under the standard structural model of time-invariant student-specific effects. We present some known results about the relationship between the estimators and its dependence on the number of observed test scores for each student and the strength of their correlation within students, and provide conditions under which they will converge to a common value. We then provide an alternative view of the relationship between the estimators that facilitates comparison to the more complex cases considered later in the paper.

2.1 Model Specification

We assume that $T$ achievement measures are tracked for each of $n$ students. The students may represent multiple cohorts but for simplicity of the description we act if is the students belong to a single cohort. The model that we specify in this section is most applicable when these achievement measures are taken on a single subject (e.g. reading or mathematics) at different time points, and thus we refer to measurements $t = 1, \ldots, T$ in terms of time, but the results of the section do not require this. In the simplest case with one achievement measure per grade per student, $T$ is equal to the number of grades, but this is not required. The case of $T$ greater than the number of grades is practically relevant when, for example, students are tested on multiple assessments (e.g. separate district and state tests) during each school year, or when the data are such that historical achievement data available on students are available but teacher links or other covariate information are available for only more recent grades.

We let $Y_{it}$ be the achievement score of student $i$ at time $t$. We let $Y_i$ denote the vector of scores for student $i$ and posit the model

$$Y_i = Z_i \theta + 1 \delta_i + \epsilon_i \tag{1}$$

The design matrix $Z_i$ is $(T \times k)$ and has an associated $(k \times 1)$ parameter $\theta$ which is unknown and is the objective of inference (note that in general, $k$ may be a function of $T$ because of the addition of covariates (e.g. timepoint means) as $T$ grows). Each student has a specific effect $\delta_i$ that applies to all scores (via the $(T \times 1)$ vector $1$). The treatment of this effect as either fixed or random in the process of estimating $\theta$ is the primary consideration of this section. We make no assumption about the relationship of $\delta_i$ to the other covariates in the model, leaving open the possibility that $E(\delta_i|Z_i) \neq 0$. The residual error term $\epsilon_i$ is assumed to be $N(0, \sigma^2 I)$, is assumed to be independent

---

1We use “achievement score” generally, and we allow the possibility that the scores are actually annual gain scores which are commonly used directly as outcome variables in the education research literature. The algebra applied to the model to produce the estimators is equally valid if the $Y_{it}$ were to be annual gain scores rather than level scores, but the assumptions of the model might be more appropriate for level scores than for gain scores.
of \( \delta_i \), and is assumed to satisfy \( E(\epsilon_i|Z_i) = 0 \). (Throughout the paper we use \( I \) to denote identity matrices of conforming size.)

The design matrix \( Z_i \) is general and may include time marginal means, time-invariant characteristics of individual teachers, time-varying teacher-level or classroom-level predictors, time-varying or time-invariant (if students switch schools) school factors, or time-varying student characteristics. We assume that \( Z_i \) does not include time-invariant student characteristics. In that case, the parameters for those characteristics are identified only if the individual student effects are treated as random effects. If such coefficients are of primary interest, then one is forced to use random effects (Frees, 2004; Wooldridge, 2001) and the comparison of the two estimators is meaningless. We are more interested in considering cases where both approaches are possible.

In the context of value-added models where the parameters \( \theta \) are individual teacher effects, the level of generality we assume about \( Z_i \) covers all of the additive teacher effects models considered in the literature, including the Tennessee Value Added Assessment System (TVAAS) “layered” model of Sanders, Saxton and Horn (1997), the cross-classified model of Raudenbush and Bryk (2002) as well as the variable persistence model of McCaffrey et al. (2004) if the persistence parameters are treated as known. However, in these models the individual teacher effects are treated as random effects, an issue which we revisit in the Discussion.

The expanded version of the model in Equation 1 for the stacked vector of student outcomes \( Y \) (of length \( nT \)) across both students and time is

\[
Y = Z\theta + D\delta + \epsilon \tag{2}
\]

where \( Z \) is \((nT \times k)\) obtained by stacking the \( Z_i \), \( D \) is a \((nT \times n)\) matrix of indicators or dummy variables linking records to the student effects given by the \( n \)-vector \( \delta \), and \( \epsilon \) is a \( nT \)-vector with distribution \( N(0, \sigma^2 I) \), independent of \( \delta \), and satisfies \( E(\epsilon|Z) = 0 \).

### 2.2 Fixed Effects Estimator for \( \theta \)

The fixed effects estimator is obtained by ordinary least squares. To make clear the parallel between this estimator and the random effects estimator, it is convenient to express the estimator in its two-stage regression form. In the first stage, we regress the student fixed effects (\( D \)) out of both \( Y \) and each of the columns of \( Z \). \( \hat{\theta}_F \) is then obtained by regression on the resulting sets of residuals.\(^2\) Letting \( H_D = D(D'D)^{-1}D' \) be the “hat” or projection matrix from the first-stage regression on the student fixed effects, the residuals from the regression of \( Y \) on \( D \) are given by \((I - H_D)Y\), and similarly the residuals from the regression of \( Z \) on \( D \) are given by \((I - H_D)Z\).

\(^2\)We are implicitly assuming that the matrix \([Z|D]\) is full column rank, which would not be the case if for example \( Z \) included indicator columns for all teachers. In that case, columns for some teachers would need to be excluded from \( Z \) and the appropriate changes would need to be made to the random effects model to maintain comparability.
Using the fact that \((I - H_D)\) is symmetric and idempotent, the estimator \(\hat{\theta}_F\) from the second stage regression is
\[
\hat{\theta}_F = (Z'(I - H_D)'(I - H_D)Z)^{-1}Z'(I - H_D)'(I - H_D)Y
= (Z'(I - H_D)Z)^{-1}Z'(I - H_D)Y
\]

(3)

2.3 Random Effects Estimator for \(\theta\)

The random effects estimator treats the student effects \(\delta_i\) as random effects with variance \(\nu^2\). Conditional on known values of \(\nu^2\) and the residual variance \(\sigma^2\) (an important assumption that we revisit later), and letting \(\hat{R}\) be the covariance matrix of the terms \(D\delta + \epsilon\), the \(\theta\) is estimated by generalized least squares via
\[
\hat{\theta}_R = (Z'R^{-1}Z)^{-1}Z'R^{-1}Y
\]

(4)

Motivation for the generalized least squares estimator and the standard properties of the estimator are based on the assumption that the \(\delta_i\) are independent of the other variables in the model. However, the estimator can be used even if that assumption is violated and the resulting estimator can still have desirable properties.

Relating the random effects estimator to the fixed effects estimator for \(\theta\) requires evaluating \(R^{-1}\). Because the random student effects are assumed to be independent across students, \(R\) is block diagonal with each block being \(R_1 = \sigma^2 I + \nu^2 J\), where \(J\) is a \(T \times T\) matrix with every element equal to 1. Block diagonal matrices appear numerous times in this remainder of this paper. When the blocks are equal, as they are in \(R\), we use a bold-face English letter to denote the block diagonal matrix and the same letter with the subscript “1” to denote an individual block. When the blocks differ by student we use the letter with the subscript \(i\) to denote the individual blocks. It is convenient to re-express \(R_1\) using the student-level intra-class correlation of \(\rho = \nu^2 / (\nu^2 + \sigma^2)\), so that each block is
\[
(\nu^2 + \sigma^2)((1 - \rho)I + \rho J).
\]

The standard result on the inverse of a compound symmetric matrix (Searle, Casella, & McCulloch, 1992) implies that
\[
((\nu^2 + \sigma^2)((1 - \rho)I + \rho J))^{-1} = \frac{1}{\sigma^2} \left( I - \frac{\rho}{1 + \rho(T - 1)} J \right).
\]

The coefficient on \(J\) is important to the comparison of the estimators so we let
\[
\gamma = \frac{\rho}{1 + \rho(T - 1)}
\]

(5)
giving that $R^{-1}$ is block diagonal with block $R^{-1}_{ij} = \frac{1}{\sigma^2} (I - \gamma J)$. We can safely ignore the factor $\frac{1}{\sigma^2}$ because it cancels out of the random effects estimator. Noting that $DD' = I_n \otimes J_T$ gives that $R^{-1} \propto (I - \gamma DD')$. Returning to the matrix $H_D$ used in the fixed effects estimator and noting that $TH_D = TD(D'D)^{-1}D' = DD'$ gives that $R^{-1} \propto (I - \gamma TH_D)$. Thus the random effects estimator is

$$\hat{\theta}_R = (Z'(I - \gamma TH_D)Z)^{-1} Z'(I - \gamma TH_D)Y$$

(6)

### 2.4 Comparing the Estimators

Comparing the fixed effects estimator $\hat{\theta}_F$ from Equation 3 to the random effects estimator $\hat{\theta}_R$ in Equation 6 shows that the estimators have a highly similar structure and that they differ only insofar as $\gamma T$ differs from 1. Similar derivations of this result are presented by Wooldridge (2001) and Frees (2004). Note that for the fixed effects estimator, the transformation of $Z$ and $Y$ by $(I - H_D)$ subtracts from the elements of each vector the within-student averages of those elements. For example, the elements of $(I - H_D)Y$ are $(Y_i - \bar{Y}_i)$, where $\bar{Y}_i$ is the average of the scores for the $i$th student. Analogously, Wooldridge refers to the transformation by $(I - \gamma TH_D)$ as “quasi-time demeaning” because it is equivalent to subtracting a fraction $\gamma T$ of the within-student averages of the components.

Thus, if $\gamma T$ is close to one, the estimators tend to be very similar. There are two different ways in which $\gamma T$ can approach 1. Referring to the definition of $\gamma$ in Equation 5, as $\rho \to 1$ for fixed $T$, $\gamma T \to 1$, and as $T \to \infty$ for fixed $\rho > 0$, $\gamma T \to 1$. When $\rho \to 1$ for fixed $T$, the continuity of matrix operations implies that we can write $(\hat{\theta}_F - \hat{\theta}_R) \to QY$ where the elements of $Q$ are $o(1)$, so that $(\hat{\theta}_F - \hat{\theta}_R) \overset{p}{\to} 0$. The analogous result for fixed $\rho$ as $T \to \infty$ requires more assumptions because the dimensions of various quantities are growing with $T$. In Appendix A we show that under weak assumptions about design matrices $Z$ whose number of columns is not growing in $T$, $(\hat{\theta}_F - \hat{\theta}_R) \overset{p}{\to} 0$ as $T \to \infty$ for fixed $\rho > 0$. In practical terms, these results imply that if either many scores are available for each student, or if $\rho$ is large, $\gamma T \approx 1$ and the estimators should be similar across a broad range of $Z$. For example, when $T = 5$ and $\rho$ in the range of 0.7 to 0.8, values typical of actual longitudinal achievement data series, $\gamma T$ is in the range of 0.92 to 0.95.

The main feature of the fixed effects estimator that makes it attractive under the structural model in Equation 1 is that it removes selection bias on the basis of $\delta$, which we operationalize by $E(\delta|Z) \neq 0$. Under the structural model in Equation 1, the expected value of the fixed effects estimator from Equation 3 conditional on $Z$ is

$$E(\hat{\theta}_F|Z) = \theta + (Z'(I - H_D)Z)^{-1} Z'(I - H_D)DE(\delta|Z)$$

(7)

The fact that $(I - H_D)D \equiv 0$ implies that the second term on the RHS is zero, even if $E(\delta|Z) \neq 0$,
and thus \( E(\hat{\theta}_F|Z) = \theta \). A more intuitive way to view this property is that because the fixed effects estimator relies on within-individual differences to identify parameters, and within-individual differences do not depend on \( \delta_i \) under the model, the parameters are identified with information that does not depend \( \delta_i \).

Analogous calculations with the random effects estimator in Equation 6 (assuming \( \gamma \) is known) give

\[
E(\hat{\theta}_R|Z) = \theta + (Z'(I - \gamma TH_D)Z)^{-1}Z'(I - \gamma TH_D)DE(\delta|Z)
= \theta + (1 - \gamma T)Z'(I - \gamma TH_D)Z^{-1}Z'DE(\delta|Z)
\]

(8)

where the second line follows from that the fact that \( H_D D = D \). Unlike the fixed effects estimator, the second term on the right hand side is not in general zero when \( E(\delta|Z) \neq 0 \), and thus the random effects estimator does not guarantee the elimination of the selection bias. Even as \( n \to \infty \) the second term on the RHS does not in general tend to zero and thus the random effects estimator is not consistent, the standard result leading many practitioners to prefer fixed effects.

However, the leading coefficient of \( (1 - \gamma T) \) on the bias term in Equation 8 indicates that the random effects estimator “compresses” the selection bias toward zero, with the degree of compression increasing as \( \gamma T \to 1 \). This provides an alternative view to quasi-time demeaning of the mechanism by which random effects estimators can approximately remove bias. Importantly, this bias compression is a feature of the random effects estimator that holds under more general structural models considered next, where the fixed effects estimator no longer is guaranteed to completely remove bias.

3 Extensions to More Complex Models

In this section we consider a generalization of the structural model in Equation 1 that might be particularly relevant when modeling standardized test score data. We consider the fixed effects and mixed model estimators and provide a theorem about how the bias compression of the random effects estimator can generalize to the mixed model. We also present some illustrative examples.

3.1 Model Specification

The model in Equation 1 assumes that for each student, the student-specific effect \( \delta_i \) is related to each test score in exactly the same way. This assumption is the key to the ability of the fixed effects estimator to remove bias due to student heterogeneity, because it implies that within-student differences of test scores do not depend on \( \delta_i \). While this assumption may provide adequate approximations in many circumstances, it is unlikely to be exactly met with standardized achievement
assessments given the complexities of creating and scaling the tests. Issues such as multidimensionality, content shift over time, and vertical equating may make the constant additive effect of Equation 1 too rigid to adequately capture the relationships among multiple scores taken from the same student over time, even if those tests are from a single test developer and are intended to provide measurements on a common scale (Doran & Cohen, 2005; Hamilton, McCaffrey, & Koretz, 2006; Lockwood et al., 2007; Martineau, 2006; Schmidt, Houang, & McKnight, 2005). Even circumstances as relatively simple as differential reliability of assessments over time may be sufficient to invalidate the structural model in Equation 1. And with criterion-referenced tests that are not vertically scaled becoming more common in response to the requirements of No Child Left Behind, it is likely that many longitudinal data series cannot be assumed to provide measures of a single, consistent construct over time.

We thus consider the following generalization to the structural model for the vector of scores for one student:

$$Y_i = Z_i \theta + A_1 \delta_i + \epsilon_i$$

(9)

The assumptions about the design matrix $Z_i$ are unchanged from the standard case in Equation 1. But we generalize the model for student heterogeneity by replacing the scalar $\delta_i$ specific to student $i$ with a $d$-dimensional vector of factors $\delta_i$, and we allow those factors to be represented in the tests as arbitrary linear combinations that can vary over time. We assume that the factors are mean zero, normally distributed with $V(\delta_i) = S_i$, a $(d \times d)$ positive definite matrix. The arbitrary covariance structure for the factors allows them to cover such cases as problem solving and computation abilities that might be positively correlated.

By allowing the $\delta_i$ to have multiple components, the model in Equation 9 can account for multi-dimensionality of tests and changing weights on these constructs over measurements. For example, suppose a test measures $d$ constructs so that scores depend on $d$ factors. We let the vector $\delta_i$ denote the time-invariant values on these factors for student $i$. Row $t$ of $A_1$ contains the weights for these factors for the $t$th measurement. As the measures change the values in the rows of $A_1$ change to allow differential weighting of the factors. Examples 1, 2, and 4 below provide specific examples of this type of scenario for test scores. Random polynomial growth models, considered in Example 3, are also a special case of Equation 9. In this case, $\delta$ contains the random coefficients of the polynomial growth model and the columns of $A_1$ are the polynomials of time.

The factors are assumed to be independent across students, but as before we allow the possibility that $E(\delta_i|Z) \neq 0$. We further assume that $\text{rank}(A_1) = d$ for all $T$, which essentially means that we have chosen a parameterization of the factors $\delta_i$ with minimal dimension. If $\text{rank}(A_1) = r < d$, it can be shown that it is possible to reduce the factor to one of dimension $r$ and recover the same marginal covariance structure of the student heterogeneity terms, so without loss of generality we assume that we are dealing with this maximally parsimonious representation
at the outset.

The residual error term $\epsilon_i$ is assumed to be mean zero, to be independent of $\delta_i$ for each $i$, to be independent across students, and to satisfy $E(\epsilon_i | Z) = 0$ for each $i$. We let $V(\epsilon_i) = \Psi_1$, a positive definite matrix with diagonal elements that are bounded both away from zero and away from $\infty$. In many practical cases it may be reasonable to assume that $\Psi_1$ is diagonal, but this restriction is not required. For example, we allow the possibility that $\epsilon_i$ has an autoregressive structure.

These assumptions imply that $V(A_1\delta_i + \epsilon_i) = A_1S_1A_1' + \Psi_1$, the usual form considered in factor analysis when $S_1 = I$ (Anderson, 1984; Morrison, 1990). This structural model recovers the standard model considered in Section 2 by taking $d = 1$, $A_1 = 1$ and $\Psi_1 = \sigma^2 I$. However, by allowing the possibility of multiple student-specific factors that link differentially to the test scores, and by allowing different residual variances across time points, this model is capable of expressing arbitrary covariance structures of the student-specific portion of the model.

The expanded version of the model for all student outcomes

$$Y = Z\theta + A\delta + \epsilon$$

(10)
is analogous to the one presented in Equation 2, where again $Z$ is $(nT \times k)$ obtained by stacking the $Z_i$, $A$ (the “expanded” $A_1$) = $I_n \otimes A_1$ is $(nT \times nd)$, $\delta$ is length $nd$ obtained by stacking the student factor vectors $\delta_i$, and $\epsilon$ is the vector of $nT$ residual errors with covariance matrix $I_n \otimes \Psi_1$.

The covariance matrix of the student portion of the model, $(A\delta + \epsilon)$ is thus $R = I_n \otimes R_1 = I_n \otimes (A_1S_1A_1' + \Psi_1)$.

### 3.2 Comparing the Estimators

The fixed effects estimator under this model is identical to that in Equation 3, and the mixed model estimator, conditional on the value of $R$, is identical to the random effects estimator of Equation 4, except that $R$ will have more complex structure in the mixed model than it did in the simple random effects model. The mixed model estimator is also the generalized least squares estimator assuming the residual covariance matrix $R$. In some circumstances $R$ may be parameterized and estimated via random effects or random coefficients (e.g. in HLM models) but this is not necessary. For example, in value-added modeling of individual teacher effects, both the TVAAS layered model (Sanders et al., 1997) and the variable persistence model of McCaffrey et al. (2004) allow $R$ to be unstructured. Later we discuss practical considerations about when it is feasible to impose no structure on $R$ and when parameter-reducing assumptions might be required.

Under the structural model in Equation 10, the expected value of the fixed effects estimator conditional on $Z$ is

$$E(\hat{\theta}_F | Z) = \theta + (Z'(I - H_D)Z)^{-1} Z'(I - H_D)A E(\delta | Z)$$

(11)
and analogous calculations with the mixed model estimator (assuming known \( \mathbf{R} \)) give

\[
E(\hat{\theta}_R | \mathbf{Z}) = \theta + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{R}^{-1} \mathbf{A} E(\delta | \mathbf{Z})
\] (12)

The relative abilities of the estimators to remove bias when \( E(\delta | \mathbf{Z}) \neq 0 \) are driven in large part by the matrix \((\mathbf{I} - \mathbf{H}_D)\mathbf{A}\) for the fixed effects estimator compared to the matrix \( \mathbf{R}^{-1}\mathbf{A} \) for the mixed model estimator. In general, neither matrix is zero and so in general, neither estimator is unbiased or consistent as \( n \to \infty \) when \( E(\delta | \mathbf{Z}) \neq 0 \). It is possible for the structure of \( \mathbf{A} \) to imply \((\mathbf{I} - \mathbf{H}_D)\mathbf{A} = 0\) (for example, when the columns of \( \mathbf{A} \) have constant elements), but assuming the diagonals of \( \Psi \) are bounded away from zero, \( \mathbf{R}^{-1} \) is positive definite which implies that \( \mathbf{R}^{-1}\mathbf{A} = 0 \) only in the degenerate case of \( \mathbf{A} \equiv 0 \).

However, the elements of \( \mathbf{R}^{-1}\mathbf{A} \) can be very close to 0 in many cases when those of \((\mathbf{I} - \mathbf{H}_D)\mathbf{A}\) will not be. In Appendix B we prove the following theorem, which implies that in many practically relevant cases, the elements of \( \mathbf{R}^{-1}\mathbf{A} \) approach zero as \( T \to \infty \):

**Theorem 1:** Let \( \mathbf{A}_1 \) and \( \Psi_1 \) be defined as above. Then sufficient conditions for the elements of \( \mathbf{R}^{-1}\mathbf{A} \) to go to zero uniformly as \( T \to \infty \) are:

1. The smallest eigenvalue of \( \mathbf{A}_1'\Psi_1^{-1}\mathbf{A}_1 \) goes to infinity as \( T \to \infty \); and
2. There exists a number \( C \) independent of \( T \) such that the elements \( a_{it} \) of \( \Psi_1^{-1/2} \), the symmetric square root of \( \Psi_1^{-1} \), satisfy \( \sum_{t=1}^{T} |a_{it}| < C \) for all \( i \)

These conditions, while abstract, are met in many practical cases. For example, the standard model in Equation 1 has \( d = 1, \mathbf{A}_1 = \nu \mathbf{1} \) with \( \nu > 0 \), and \( \Psi = \sigma^2 \mathbf{I} \). Thus \( \mathbf{A}_1'\Psi_1^{-1}\mathbf{A}_1 \) is the scalar \((\nu^2/\sigma^2)T \to \infty \) as \( T \to \infty \), and the rowsums of \( \Psi_1^{-1/2} \) are identically equal to \( 1/\sigma \). Thus the conditions of the theorem are met, reiterating the results presented in Section 2. We provide a few additional examples in which the conditions are met in the simplified case where \( \Psi = \sigma^2 \mathbf{I} \) with \( \sigma > 0 \), which is representative of more general cases of \( \Psi = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2) \) with \( 0 < \sigma_{\text{lower}}^2 \leq \sigma_i^2 \leq \sigma_{\text{upper}}^2 < \infty \) (both of which satisfy condition 2 of the Theorem). Cases with more general \( \Psi_1 \) are generally analytically intractable.

- **Example 1:** Suppose we generalize the standard student effects model in Equation 1 to allow the student effect to be weighted differently for each measurement. In this case, \( d = 1 \) and \( \mathbf{A}_1 \) is the vector \((a_1, \ldots, a_T)' \). Then \( \mathbf{A}_1'\Psi_1^{-1}\mathbf{A}_1 = (1/\sigma^2) \sum_{t=1}^{T} a_t^2 \), and so condition 1 of the theorem is met as long as the series \( \sum_{t=1}^{T} a_t^2 \) is divergent. A sufficient condition for this divergence is \( a_t \) remain bounded away from zero as \( T \to \infty \); that is, as long as each test provides an amount of information about \( \delta \) that is not diminishing to zero.

- **Example 2:** Suppose \( d = 2 \) and row \( t + 1 \) of \( \mathbf{A}_1 \) is \( \frac{1}{T-t}(T-1-t,t) \) for \( t = 0, \ldots, T-1 \). That is, the sequence of tests gradually changes from all weight on the first factor to all
weight on the second factor. Then \( A'_1 \Psi^{-1}_1 A_1 = \frac{1}{\sigma^2(T-1)^2} C \) where \( c_{11} = c_{22} = \frac{1}{3} T^3 + o(T^3) \) and \( c_{21} = c_{12} = \frac{1}{6} T^3 + o(T^3) \). The eigenvalues of \( C \) can be shown to be \( c_{11} \pm c_{21} \), with the smaller eigenvalue thus being \( \frac{1}{6} T^3 + o(T^3) \). Multiplying by \( \frac{1}{\sigma^2(T-1)^2} \) gives that the smallest eigenvalue of \( A'_1 \Psi^{-1}_1 A_1 \) behaves like \( \frac{1}{6\sigma^2 T} \to \infty \) as \( T \to \infty \).

- **Example 3:** (Random linear growth model) Suppose \( d = 2 \) and suppose that the columns of \( A_1 \) express a random linear growth model parameterized such that the first column is \( (1, 1, \ldots, 1)' \), and the second is \( (1, 2, \ldots, T)' \). Then \( A'_1 \Psi^{-1}_1 A_1 \) is \( (T/\sigma^2) \) times a matrix for which the smallest eigenvalue is bounded away from zero as \( T \to \infty \), and thus \( A'_1 \Psi^{-1}_1 A_1 \) satisfies the conditions of the theorem.

- **Example 4:** Suppose any \( d \geq 1 \) and suppose that for each measurement the rows of \( A \) are a random draw from Dirichlet distribution with parameter \( \omega \), a \( d \times 1 \) vector such that for \( j = 1 \ldots d \omega_j > 0 \) and \( \omega_0 = \sum_{j=1}^d \omega_j \). Then \( A'_1 \Psi^{-1}_1 A_1 \to \frac{1}{\sigma^2} (\Omega + c\omega \omega') \), where \( \Omega \) is a diagonal matrix with elements \( \omega_j \omega_j \) and \( c = \frac{\omega_0 (\omega_0 + 1)^{-1}}{\omega_0^2 (\omega_0 + 1)^{-1}} > 0 \). Theorem 3 on p. 116 of Bellman (1960) states that if \( B_1 \) and \( B_2 \) are symmetric with \( B_2 \) is positive semi-definite, then the smallest eigenvalue of \( B_1 + B_2 \) is greater than or equal to the smallest eigenvalue of \( B_1 \). Letting \( B_1 = \Omega \) and \( B_2 = c\omega \omega' \) gives that the smallest eigenvalue of \( \frac{1}{\sigma^2} (\Omega + c\omega \omega') \) is greater than or equal to the smallest diagonal element of \( \frac{1}{\sigma^2} \Omega \) which is greater than zero. Hence, the eigenvalues of \( A'_1 \Psi^{-1}_1 A_1 \) will diverge.

In practical terms, the theorem suggests that when many test scores are available for individual students, and when the covariance structure of these test scores deviates from that implied by the simple time-invariant constant offset model of Equation 1, the mixed model estimator may be equally effective, and potentially more effective, at mitigating bias in parameter estimates than the fixed effects estimator across a broad range of design matrices \( Z \). However, generally characterizing the circumstances under which the elements of \( R^{-1} A \) going to zero implies that the mixed model estimator will compress bias as \( T \to \infty \) is complex because it depends on the structure of \( Z \). A particular challenge is that in most circumstances, growing \( T \) implies a growing number of predictors in the model (i.e. columns of \( Z \)), and in many cases those added predictors apply to only a single time point (e.g. time point means or interactions of other variables with time).

One sufficient condition to ensure bias compression as \( T \to \infty \) is that the sums of the absolute values of the rows of \( (Z'R^{-1}Z)^{-1} Z' \) are uniformly bounded by a constant (not depending on \( T \)) as \( T \to \infty \). This is not a particularly intuitive condition, and the asymptotic result might not be well approximated for realistic values of \( T \). Hence, we now consider several empirical examples building on the examples above to understand when the condition seems likely to hold and when it does not. We compare the behavior of the mixed model estimator to the fixed effect and simple OLS estimators to demonstrate the importance of Theorem 1 for reducing bias in estimated effects in situations similar to some applied settings in terms of strength of correlation and \( T \).
In the examples, we include the OLS estimator to calibrate the strength of the selection bias and demonstrate the power of mixed models to reduce, if not completely remove, that bias.

4 Empirical Examples

In this section we consider a series of four examples, building on the examples presented in Section 3, to understand the implication of Theorem 1 in real applications. The residual errors of the data generating model of each example $(A\delta + \epsilon)$, satisfy the conditions of the theorem. For each example, we consider alternative values for the “treatment” variables $Z$ to understand applications where the theorem will be sufficient for mixed models to remove bias from non-random assignment and where it will not. We also consider different non-random assignment mechanisms that relate components of $\delta$ to $Z$. For reference, Table 1 summarizes the empirical examples that we consider in this section in terms of the model for student heterogeneity, the treatment assignment mechanism, and the configuration of the treatment variables.

4.1 Example 1, continued

The model for residual errors in this example is the same as it was in Example 1 of Section 3. Specifically, data for 1000 students are generated from a model where each student has a single student effect that is weighted differentially by each measure. The model used to generate the data is:

$$Y_{it} = z_{it}'\beta + a_t\delta_i + \epsilon_{it}$$

where $Y_{it}$ is student $i$’s score in time period $t$ for $t = 1 \ldots T$, $z_{it}$ is a vector of independent variables that can depend on the time period and varies with different scenarios we consider (described below), $\delta_i$ is a random normal variable with mean zero and variance one, $a_t$ is a weight that varies across time, and the $\epsilon_{it}$ are independent random normal variables with mean zero and variance $1 - a_t^2$. Thus the variance of the residuals after accounting for treatment effects is equal to 1 for all time points. The $a_t$ vary from 0.7 to 0.9 as $t$ increases from 1 to $T$, so that the correlations of the residuals within students over time are bounded between about 0.5 and 0.8, values consistent with real longitudinal achievement data. If the $a_t$ were constant then the model would be the standard student effect model, so this example is a slight deviation from that case.

To explore how $Z$ affects the implications of Theorem 1, we consider four scenarios for the $z_{it}$. All four involve simple dichotomous treatment indicators where the log odds of treatment assignment equal $\delta_i$ - that is, student assignment to treatment is non-random. In scenario 1, a student’s treatment assignment remains the same for all values of $t$ and there is a single treatment effect that is constant for all values of $t$. In scenario 2, a student’s treatment assignment can vary across time periods, but there is again only a single treatment effect that is constant for all values
<table>
<thead>
<tr>
<th>Student Heterogeneity Model</th>
<th>Treatment Assignment</th>
<th>Treatment Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single factor with varying weights over time</td>
<td>Depends on factor</td>
<td>Four scenarios:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1: Single Tx effect, students remain in same Tx condition at all time points</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2: Single Tx effect, students’ Tx conditions vary across time points</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3: Time-varying Tx effects, students remain in same Tx condition at all time points</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4: Time-varying Tx effects, students’ Tx conditions vary across time points</td>
</tr>
<tr>
<td><strong>Example 2</strong></td>
<td></td>
<td>Treatment only in final year</td>
</tr>
<tr>
<td>Two correlated factors with weights gradually changing from one to the other over time</td>
<td>Three scenarios:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1: Tx assignment depends equally on both factors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2: Tx assignment depends on only first factor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3: Tx assignment depends on only second factor</td>
<td></td>
</tr>
<tr>
<td><strong>Example 3</strong></td>
<td></td>
<td>Four scenarios:</td>
</tr>
<tr>
<td>Linear growth model with random slopes and intercepts; consider two cases: independent intercepts and slopes</td>
<td>Two scenarios:</td>
<td>1: Single Tx effect, students remain in same Tx condition at all time points</td>
</tr>
<tr>
<td></td>
<td>1: Assignment depends on the intercepts alone;</td>
<td>2: Single Tx effect, students’ Tx conditions vary across time points</td>
</tr>
<tr>
<td></td>
<td>2: Assignment depends on the slopes alone.</td>
<td>3: Time-varying Tx effects, students remain in same Tx condition at all time points</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4: Time-varying Tx effects, students’ Tx conditions vary across time points</td>
</tr>
<tr>
<td><strong>Example 3 (Teacher Effects)</strong></td>
<td>Assignment to teachers depends on both intercepts and slopes</td>
<td>Consider one to four tested subjects per year and four different models for teacher effect persistence</td>
</tr>
</tbody>
</table>

**Table 1:** Summaries of student heterogeneity model, treatment assignment mechanism, and treatment exposure scenarios for the empirical examples considered in this section.
of $t$. In scenario 3, a student’s treatment assignment remains the same for all values of $t$, but the mean and the treatment effect vary with $t$. In scenario 4, a student’s treatment assignment can vary across time periods and the mean and the treatment effect vary with $t$. For each scenario, we consider estimators for $T = 3, 5, 10, 15,$ and $20$ with 100 Monte Carlo iterations run at each value of $T$.

![Graphs showing standardized bias for different scenarios and estimators.](example_1_graphs)

Figure 1: (Example 1) Average absolute standardized bias as a function of $T$. Different scenarios are shown in different frames. Each frame contains the average absolute standardized bias for OLS (plus signs), mixed model (open circles) and fixed effects (triangles). When treatment assignment is constant in scenario 1 (upper left), the fixed effects estimates are unavailable. When treatment assignment is constant and treatment effect vary over time in scenario 3 (upper right), fixed effects estimates differences in the treatment effects, not the treatment effects themselves.

Figure 1 presents a standardized measure of absolute bias for each estimator (OLS, fixed effects, and mixed model) as a function of the scenario and the number of scores. For the mixed model estimator, we use the known value of $R$; we consider the behavior of the estimator when $R$ also must be estimated in Example 2. The standardized bias measure for a given estimator, scenario...
and number of scores is calculated as follows. Each Monte Carlo simulation provides estimates of the treatment effects (of which there are $T$ for scenarios 3 and 4), the absolute values of which are then divided by the marginal standard deviation in the scores. This makes the metric interpretable in terms of standardized effect sizes. For scenarios 3 and 4, these standardized quantities are then averaged over the multiple treatment effects of these scenarios. Finally the resulting quantities are averaged across Monte Carlo simulations.

The figure reveals several important facts about each estimator. First OLS is biased for every scenario and every value of $T$. The size of the bias depends on the strength of the selection. The fixed effects estimator cannot estimate the treatment effect in scenario 1 because treatment assignment is constant and perfectly collinear with the student indicator variables used to estimate the student effects. In the scenario 2, the treatment effect can be estimated because students’ treatment assignments change across time. In scenario 3 treatment assignment does not change but treatment effects change, and this allows the fixed effects estimator to recover differences in treatment effects but it cannot recover all $T$ individual treatment effects. In the scenario 4, the fixed effects estimator can recover all the treatment effects since the student change assignment over time.

In general under the scenarios of this example, the fixed effects estimator will be biased and inconsistent because the $a_t$ are not constant and consequently the residual error in scores less the student’s mean scores are not independent of treatment assignment. However, the bias is very small relative to the bias in OLS because the residual errors in the demeaned scores are more
weakly correlated with assignment than the raw scores, particularly given the values of \( a_t \) used here. The more variable are the \( a_t \) across years, the larger the bias can be. For scenario 2, the bias is particularly small because of the approximate balance of treatment assignment by year, leading to cancellation of the positive and negative biases that exist for the individual treatment effect estimates in scenario 4. Even in scenario 4 the average absolute bias across all estimated treatment effects is small, but the bias can be relatively large for treatment effects from very early or very late in the data series. Figure 2 shows the bias by individual treatment effect in scenario 4 for \( T = 20 \). The fixed effects estimators are negatively biased for treatment effects early in the data series, positively biased for those later in the data series, and approximately unbiased for those in the middle of the series. In contrast, the mixed model estimator is positively biased, with approximately constant magnitude, for all treatment effects.

The mixed model estimator greatly reduces the bias relative to OLS in scenarios 2 and 4 in which treatment assignment varies over time. The bias decreases as \( T \) increases in accordance with Theorem 1. The results holds in scenario 4 even though the number of treatment effects is growing with \( T \). However, when treatment assignment is constant (scenarios 1 and 3; top row of Figure 1), the mixed model estimators are biased and the bias does not decrease with \( T \). In these scenarios, the mixed model estimator is similar to the OLS estimator and performs about the same. The row sums of the absolute values of the elements of the matrix \( (Z' R^{-1} Z)^{-1} Z' \) appear to grow with \( T \) when the treatment assignment is constant but not when it is not. The reason for the difference between the estimators is that when treatment assignment is constant each student’s contribution to \( Z' R^{-1} Z \) is not growing in \( T \) because of cancellation between the negative and positive elements of \( R^{-1} \). When treatment assignment varies, cancellation does not occur and the elements of \( Z' R^{-1} Z \) grow with \( T \). This behavior alternatively can be viewed in terms of available information to estimate \( \delta_i \) - only when treatment assignment varies can the data provide the information about \( \delta_i \) necessary to correct for selection. We speculate that as long as both the \( a_t \) and the overall fraction of non-treatment scores remain bounded away from zero as \( T \to \infty \), the bounded rowsums condition will be met and the estimator will be consistent.

### 4.2 Example 2, continued

This example uses the model for residual errors first introduced in Example 2 of Section 3 in which \( \delta \) has two components and the contributions of these components to scores gradually changes across tests from one factor to the other. Such a situation might occur if math tests measure two constructs such as computation and problem solving, but the weight given to the two constructs varies systematically across tests. For example, tests for elementary students might focus more on computations whereas tests for middle school students might focus more on problem solving.

The \( Z \) used in this model was motivated by a case that is common in applications where \( T - 1 \) test scores are collected on students prior to treatment and then a fraction of the students receive
treatment with treatment assignment depending on unobserved characteristics of the student. It is clear that this $Z$ will meet the sufficient condition on the $(Z' R Z)^{-1} Z'$ and we showed in Section 3 that the assumptions of Theorem 1 are met. Hence, for large $T$, the treatment effect estimated from the mixed models estimator should be approximately unbiased. The motivation for this example is then to explore the properties of the estimator when $T$ is small and to contrast the performance of the mixed model estimator to the fixed effects estimator. The model clearly violates the assumptions required for the fixed effects estimator to be consistent. We also use this model to explore the impact of the treatment assignment mechanism on bias and to explore the impact of estimating $R$ on the performance of the mixed model estimator.

The score $Y_{it}$ for student $i$ on test $t = 1, \ldots, T$ is generated according to the model

$$Y_{it} = \mu_t + z_{it} \beta + a_{1t} \delta_{i1} + a_{2t} \delta_{i2} + \epsilon_{it}.$$  

The $\mu_t$ are the mean scores for all students by year, and $\beta$ is the effect of treatment and is the parameter of interest. The variable $z_{it}$ is a treatment assignment indicator, and because all but the final test are pre-treatment it is identically equal to zero for all $t < T$. For the roughly one half of the students who receive treatment $z_{iT} = 1$ and for all other students it is zero. The weights $a_{1t}$ are an evenly spaced sequence of values from .1 to .9 and the weights $a_{2t}$ are an evenly spaced sequence of values from .9 to .1. The $\delta_i$ are bivariate normal random variables, with mean zero, variance one and correlation .5. The $\epsilon_{it}$ are i.i.d. $N(0, .2)$ variables. The correlations among observations from the same student vary from about 0.5 to 0.8 with mean around 0.75.

Using this model, for a sequence of $T$s ranging from 2 to 20, we generated 100 Monte Carlo samples of 1,000 students independently under three scenarios for treatment assignment. In the first scenario, treatment assignment depends equally on $\delta_{i1}$ and $\delta_{i2}$ with the log of the odds of treatment equal to $4 \delta_{i1} + 4 \delta_{i2}$. In the second scenario, treatment assignment depends only on $\delta_{i1}$ with the log of the odds of treatment equal to $4 \delta_{i1}$ and in the third scenario, treatment assignment depends only on $\delta_{i2}$ with the log of the odds of treatment equal to $4 \delta_{i2}$. For each scenario, we consider the bias in the estimated the treatment effect using OLS, fixed effects, and the mixed model estimator using both the known $R$ as well as $R$ estimated from the data using restricted maximum likelihood implemented in SAS PROC MIXED. We also repeated the estimation using 5,000 observations. The results are extremely similar to the cases with 1,000 and are not presented in the figure.

Figure 3 gives the absolute bias in the estimated treatment effect standardized by the marginal standard deviation in scores for the last period for the all four estimators. For all three scenarios, students assigned to treatment have significantly higher $\delta$ values and the OLS estimator has substantial bias, especially in scenarios 1 and 2. For scenario 1, in which treatment assignment depends equally on both elements of $\delta$, the fixed effects estimator is consistent and consequently the bias is negligible for all values of $T$. For this scenario the random effects estimator also has
small bias for $T$ greater than 10.

In scenario 2, treatment assignment only depends on $\delta_1$ and fixed effects are inconsistent. The bias declines somewhat as $T$ increase from 3 to 4 but then it asymptotes to approximately 10 percent of the marginal standard deviation within treatment group. The bias in the mixed model estimator is smaller than the bias in the fixed effects estimator for $T > 4$ and it approaches zero for $T = 20$.

Treatment assignment in scenario 3 depends only on $\delta_2$ which has greater weight in the early scores than the later scores. Because assignment does not depend on both $\delta_1$ and $\delta_2$, again the fixed effects estimator is not consistent and its performance is very similar its performance in scenario 2. However, the OLS estimator using only the test score from the final period in which $\delta_2$ is severely downweighted to estimate treatment effects and consequently the bias in OLS is considerably smaller than in the other two scenarios. Similarly the small weights on $\delta_2$ in final period result in very low bias for the random effects estimator for all values of $T$.

In all three scenarios and at all values of $T$, the performance of the mixed model estimator with known $\mathbf{R}$ is very similar to the performance of the estimator with $\mathbf{R}$ estimated from the data. In scenarios 1 and 2, there is a trend for performance of the estimator with known $\mathbf{R}$ to improve relative to the estimator with an estimated $\mathbf{R}$ at larger values of $T$ but the differences between the estimators are always less than an percentage point or two. The difference for larger values of $T$ might be due to imprecision in estimating such a large covariance matrix but the similarity at $T = 20$ is very encouraging because the covariance matrix has 210 parameters but even with as few as 1000 students it estimated with sufficient precision to effectively eliminate the bias.
This example extends Example 3 of Section 3 considering a model where growth trajectories in students’ test scores follow student-specific linear models in time. This is an example of a random slopes and intercepts model (Raudenbush & Bryk, 2002). Again, Section 3 demonstrated that this model meets the assumptions of Theorem 1 when the slopes and intercepts are independent. Our goal is to demonstrate the effect of $Z$ on the compression by $R^{-1}A$, and to examine the relative performance of the mixed models and fixed effects estimators and how this relative performance might depend on correlation between the random intercepts and slopes.

The score $Y_{it}$ for student $i$ in year $t$ (here indexed from 0 to $T - 1$ for convenience) is generated according to the model

$$Y_{it} = \delta_i + \lambda_i t + z_{it}' \beta + \epsilon_{it}$$

We assume the following independent distributions for the parameters, where the parameters are also independent across students:

$$(\delta_i, \lambda_i) \sim N(0, \Sigma) \quad \text{with} \quad \Sigma_{11} = \sigma_\delta^2, \Sigma_{22} = \sigma_\lambda^2, \Sigma_{21} = \Sigma_{12} = r\sigma_\delta\sigma_\lambda$$

$$\epsilon_{ist} \sim iid \quad N(0, \sigma_\epsilon^2)$$

This allows the intercept and slope parameters for the same student to have correlation $r$.

The vectors $z_{it}$ and $\beta$ describe the marginal linear model in time and treatment effects which differ among scenarios. As with Example 1, we consider four scenarios for the treatment indicators. In scenario 1, there is a constant treatment effect that does not vary by period and a treatment effect on growth. In addition, each student’s treatment assignments do not vary so that a student is either assigned to treatment every period or control every period. The treatment effects in scenario 2 also include a constant treatment effect that does not vary by period and a treatment effect on growth. However, in this case a student’s treatment assignments can vary across time periods. In scenario 3, the model again includes a marginal slope and intercept but the treatment effect now changes with each time period. Each student’s treatment assignments do not vary across time. Scenario 4 again has separate treatment effects for each time period but unlike scenario 3, each student’s treatment assignments can vary with time.

For all four scenarios, we consider two cases for treatment assignment. In the first case, treatment assignment depends on the random intercepts and in the second case it depends on the random slopes. We also consider cases where the random slopes and intercepts are independent and where they are positively correlated. We explored cases with negative correlation and the results were qualitatively the same.

Figure 4 presents results for 100 Monte Carlo samples from scenario 4 (because the results of the other scenarios provide similar insights into estimation as Example 1, they are not presented in detail but summarized later). This scenario has a separate treatment effect for each time period.
Figure 4: (Example 3) Average absolute standardized bias as a function of $T$ for data generated with random growth curves and treatment assignment changing over time. Each frame contains the average absolute standardized bias for OLS (plus signs), mixed model (open circles) and fixed effects (triangles).
This scenario is analogous to using growth modeling to estimate the effects of a teacher attribute, which can change with each measure, such as National Board Certification, on student achievement and allowing the effect to depend on the student’s grade level. For each estimator the frames of the figure plot by $T$ the average of the absolute standardized bias in estimated treatment effects, calculated similarly as described in Example 1.

As shown in the figure, the mixed model estimator greatly reduces the bias relative to OLS and for $T > 10$ the bias is very small for all the cases considered in this scenario. When selection depends only on the intercept and the slope and intercept are uncorrelated, the fixed effect estimator again works well because the demeaned data is independent of treatment assignment. However, when assignment depends on the slope removing the mean does not make the scores independent of treatment assignment. Moreover, even if assignment only depends explicitly on the intercept, correlation between the intercept and the slope can make the fixed effects estimator have substantial bias.

The results for scenarios 1 and 3 (not shown) are similar those for scenarios in Example 1 where treatment assignment is constant for students. The mixed model estimator is biased for all values of $T$ and absolute values of the bias is not decreasing with increasing values of $T$. In scenario 1, where there is a constant treatment effect fixed effects cannot identify the treatment effect estimate. In scenario 3, where treatment effects vary with time period but assignment is constant, fixed effects provides unbiased estimates of differences in treatment effects when assignment depends on the random intercept and slopes and intercepts are independent. If assignment depends on the slopes explicitly or implicitly through correlation between the slopes and intercepts, fixed effect estimates are biased because the residual errors in the demeaned test scores are not independent of treatment assignment.

In scenario 2 (not shown), treatment assignment varies and just as in Example 1, the bias in mixed model decrease as $T$ increase regardless of treatment assignment mechanism and regardless of whether slopes and intercepts are independent or correlated. Fixed effects can estimate the treatment effect and the estimates are essentially unbiased when assignment does not depend on the slope but has bias that does not vanish with increasing $T$ otherwise.

### 4.4 Example 3, continued: Teacher Effect Estimation

One of the primary motivating examples for this paper is how best to model student heterogeneity when estimating individual teacher effects. The estimation of teacher effects using multivariate longitudinal models is extremely difficult to treat analytically, primarily because of the complexity of $Z$ arising from the crossing of students with teachers over time, the assumed model for the persistence of the teacher effects into future test administrations, and the fact that the number of columns of $Z$ grows as more grades and/or subjects are considered. Thus we examine how the mixed model approach behaves through a sequence of simulations. We base the simulations
on random slopes and intercepts model discussed above but unlike the previous example, $Z$ is determined by the teacher effects specification rather than treatment effect specifications.

We consider $n = 1000$ students followed for a period of $G$ consecutive grades, with achievement measured once per grade on $S$ different academic subjects (e.g. reading, mathematics, science, etc). Thus students have a total of $T = GS$ scores. The students are assigned to teachers in each grade, and those teachers are linked to all of the subject scores in each grade. We assume that each class has 25 students, so in this case there are $n/25 = 40$ teachers per grade and thus $40G$ teachers in total across all grades. A separate effect is estimated for each teacher for each subject, so that the total number of teacher effects being estimated is $40GS$.

To simplify the evaluation of the simulation, we generate the data in such a way that there are truly no teacher effects. The score $Y_{isg}$ for student $i$ on subject $s$ in grade $g$ (indexed from 0 to $(G - 1)$) is generated according to the model

$$Y_{isg} = \delta_i + \delta_{is} + (\lambda_i + \lambda_{is})g + \epsilon_{isg}.$$ 

This is similar to the random growth model used in the previous Example 3, except it applies to multiple subjects and the trajectories from different subjects are correlated through the common intercept parameter $\delta_i$ and shared slope parameter $\lambda_i$. We assume the following independent distributions for the parameters, where all parameters are also independent across students:

$$
(\delta_i, \lambda_i) \sim N(0, \Sigma) \quad \text{with} \quad \Sigma_{11} = \sigma_\delta^2, \Sigma_{22} = \sigma_\lambda^2, \Sigma_{21} = \Sigma_{12} = r\sigma_\delta\sigma_\lambda
$$

$$
\delta_{i1} \ldots \delta_{iS} \sim \text{iid } N(0, \nu_\delta^2)
$$

$$
\lambda_{i1} \ldots \lambda_{iS} \sim \text{iid } N(0, \nu_\lambda^2)
$$

$$
\epsilon_{isg} \sim \text{iid } N(0, \sigma_\epsilon^2)
$$

(13)

Like Example 3 above, the only parameters that are allowed to be correlated are the common intercept and slope parameters, which have correlation $r$. Under this model, the marginal variance of the scores is the same for all subjects in a given grade, and is $(\sigma_\delta^2 + \nu_\delta^2) + g^2(\sigma_\lambda^2 + \nu_\lambda^2) + 2gr\sigma_\delta\sigma_\lambda + \sigma_\epsilon^2$ for $g = 0, \ldots, (G - 1)$. In general scores are correlated across both subjects and grades, with covariances determined by the parameters and the time lags.

Students are regrouped into classes each grade. We introduce spurious teacher effects by making these assignments non-random, and in particular making assignments dependent on the parameters $\delta_i$ and $\lambda_i$. For each student in each grade, we calculate the quantity

$$\eta_{ig} = 0.3(\delta_i/\sigma_\delta) + 0.3(\lambda_i/\sigma_\lambda) + 0.4\xi_{ig}$$

where $\xi_{ig}$ are independent standard normal variables. For each grade, we assign the smallest 25 $\eta_{ig}$
to class 1, the next smallest 25 $\eta_{ig}$ to class 2, etc, all the way to the largest 25 $\eta_{ig}$ to class 40. This results in selection into classrooms that is moderately strong on both student intercepts and student slopes.

In order to estimate teacher effects from these data (the true values of which are identically zero), we need to make assumptions about the form of the design matrix $Z$ that links teacher effect indicators given by the $40GS$ columns to sequences of test scores for students. We assume in all cases that the teacher effect for a given subject affects only the outcomes (potentially current and future) for that same subject. For each subject, we assume that the effect of a teacher experienced in grade $g_1$ persists into grade $g_2 \geq g_1$ by the amount $\alpha^{g_2-g_1}$ for some $0 \leq \alpha \leq 1$. The parameter $\alpha$ is taken to be the same for all subjects. The case $\alpha = 0$ corresponds to no persistence of past teacher effects, the case $\alpha = 1$ corresponds to complete persistence of past teacher effects (i.e. the assumption made by the TVAAS layered model (Sanders et al., 1997)), and $0 < \alpha < 1$ corresponds to decaying persistence depending on the lag. The design matrices $Z_i$ are fully determined given the sequence of class assignments and $\alpha$.

Our simulation uses $G = 5$, and then varies the number of subjects per grade from 1 to 4 and also considers values of $\alpha$ of 0, 0.3, 0.7 and 1, for a total of 16 design points. We use a common set of variance components and selection model parameters across all design points. In particular we set $\sigma_\delta^2 = 0.5$, $\sigma_\lambda^2 = 0.125$, $r = 0.3$, $\nu_\beta^2 = 0.2$, $\nu_\delta^2 = 0.05$, $\sigma_\epsilon^2 = 0.8$, which leads to $R$ with marginal variances of 1.500, 1.825, 2.500, 3.525, and 4.900 by grade for each subject, and correlations (both within and across subjects) ranging from about 0.3 to 0.8 with an average of 0.5.

For each design point, we independently generate 100 datasets, and for each dataset we consider three different estimators for the teachers effects: completely unadjusted classroom means, the OLS estimator, and the mixed model estimator using the known value of $R$. The unadjusted classroom means are provided as a strawman to calibrate the strength of selection of students to teachers, like the OLS estimators in the other Examples. We did not consider the fixed effects estimator in this example. Given that selection to teachers depends on student growth parameters, student fixed effects fit to the level scores will be inconsistent. Student fixed effects on grade-to-grade gain scores, a common empirical strategy when there are concerns about selection bias due to differential grade-to-grade growth in achievement (Harris & Sass, 2006; Zimmer et al., 2003), will be consistent for all teacher effects except those in the first grade.

For each estimator, rather than summarize the bias in each individual teacher effect, we report estimated variance components of teachers by grade, expressed as a percentage of the corresponding marginal variance in that grade. This standardized measure of estimated teacher variability is commonly used in the literature to summarize the heterogeneity of teacher effects on the scale of student outcomes (Aaronson, Barrow, & Sander, 2003; Koedel & Betts, 2005; McCaffrey et al., 2004; Nye et al., 2004; Rockoff, 2004). Because our data generating model contains no true teacher effects, the correct value is zero and percentages close to zero indicate that an estimator is behaving effectively. Because of the simplified balanced design of our simulations, the behavior of the
estimators is exchangeable across subjects for a given total number of subjects and thus we average the estimated variance components across subjects within grade.

The results are summarized in Figure 5. Each frame of each plot corresponds to a different value of $\alpha$ and presents the estimated teacher variance components by grade and for numbers of subjects of 1, 2, 3 and 4. The lines connect the estimates for a given number of subjects, with the dotted lines corresponding to unadjusted means, dotdash lines corresponding to OLS, and solid lines corresponding to the mixed model estimator. The bias in the unadjusted means and OLS estimators are (up to Monte Carlo error) invariant to the number of subjects, while the bias in the
mixed model estimator changes with the number of subjects. In all cases the bias decreases as the number of subjects increase. That is, the lowest mixed model trajectory in each frame corresponds to 4 subjects, and the highest corresponds to 1 subject.

The unadjusted means indicate that the spurious variation among teachers increases across grades, which makes sense because students are selected into classrooms partially on the basis of growth. When $\alpha = 0$, OLS is equivalent to the unadjusted means but diverges from it as $\alpha$ increases. The decreasing bias in OLS as $\alpha$ increases for grades after the first grade is because the OLS estimator approaches a first-differenced (gain score) estimator as $\alpha$ goes to one. This would be sufficient for OLS to remove all bias for teachers beyond the first grade if students were selected into classes on the basis of intercepts alone; however the selection on growth ensures that OLS remains biased. Alternatively, the mixed model estimator is generally effective at removing bias, particularly when $\alpha$ is small, when the bias is uniformly small and effectively zero when multiple subjects are available. As $\alpha$ grows the bias for the mixed model estimator remains for first grade teachers, and when $\alpha = 1$, the mixed model estimator is inconsistent. This is analogous to the cases considered in previous examples where treatment assignment was constant for all time points, because when $\alpha = 1$ the first grade teacher effect is similarly constant for all time points. Thus in value added studies using mixed model estimators, it is customary not to report estimated teacher effects from the first grade of available teacher links (Ballou et al., 2004; Lockwood et al., 2006)

5 Discussion

As the use of longitudinal data analysis becomes more common for studying what educational factors impact student outcomes, understanding the costs and benefits of different statistical model specifications is imperative. This paper critically examines the tradeoffs between the fixed-effects and the random effects or mixed modeling approaches to analysis of longitudinal achievement data. The standard criticism of random effects or mixed models is that they will provide inconsistent estimates when unobserved student variables are correlated with the variables of interest. Our primary finding suggests that this criticism requires deeper consideration. The results of the econometrics literature provide conditions when fixed effects and random effects can control for bias created by unobserved student heterogeneity and demonstrate that the two estimators will provide similar results when student heterogeneity accounts for much of the unexplained variation in students outcomes and when many measurements are available for each student. Moreover, consideration of the nature of achievement tests suggests that longitudinal test score data are likely to be incompatible with the conditions for fixed effects to yield consistent results. However, our results demonstrate that provided there are sufficiently many test scores, mixed model estimators can have minimal bias even in circumstances when fixed effects estimators would fail to remove
bias. This result, in conjunction with the well-established results about the circumstances under which mixed model estimators can be more efficient than fixed effects estimators (Aitken’s Theorem; see, e.g., Theil (1971)), suggests that the mixed model approach may be beneficial in many applications.

Intuitively, the mixed model estimator can effectively remove bias because pre-multiplication of scores by \( R^{-1} \) serves as a type of regression adjustment to the raw scores. By the results for inverting partitioned matrices (Searle et al., 1992), pre-multiplication of the outcome vector \( y \) by \( R^{-1} \) results in values in which each score \( y_{it} \) is replaced by a scaled version of the residual \( y_{it} - \hat{y}_{it} \) where \( \hat{y}_{it} \) is based on the regression of \( y_{it} \) on all of the other available scores for each student. This adjustment is in contrast to the adjustment made by the fixed effects estimator, which replaces each score by its deviation from the average score for each student. As demonstrated in Section 2, under the standard unobserved effects model, differencing is optimal because it ensures that the adjusted scores no longer depend on the unobserved student effects. However, the regression adjustment made by the mixed model estimator can have equal or better properties when the true model for student heterogeneity and student selection is more complex and demeaning the data is not sufficient to remove student heterogeneity. As long as student heterogeneity can be adequately represented by a low-dimensional factor, and the signal about that heterogeneity is not swamped by the noise in the tests, then the residuals from regression adjustment can be approximately unrelated to unobserved student effects, and estimates of educational variables can be approximately unbiased. The degree of bias compression improves as both the number of available scores, and the signal to noise ratio of those scores, increases. Our results suggest that with the correlations typically observed in longitudinal achievement data series (on the order of 0.7), if only two or three tests are available and it is believed that these tests are sufficiently similar, the fixed effects approach is probably preferable in terms of bias reduction. In cases with a larger number of diverse tests that may differ in content and quality, then the mixed model approach may provide more robust inference.

Importantly, our empirical examples demonstrate that the bias compression can hold simultaneously over a suite of parameters whose dimension may be growing as the number of test scores grows, such as with the treatment-by-time interactions considered in Examples 1 and 3 and the individual teacher effects considered in the second part of Example 3. Allowing for this kind of generality in the models may be important in the very circumstances in which the mixed model approach may be preferable - when there is concern that the sequence tests are not measuring the same construct in the same way over time. It also makes the mixed model approach particularly relevant to jointly modeling outcomes on different tested subjects. Nearly all longitudinal achievement data series contain test outcomes from multiple subjects in each year, commonly including both math and reading, and often science and social studies as well. It is common in fixed-effects analyses of educational variables to model each subject separately and to report estimates of the variables of interest separately for each subject (Goldhaber & Anthony, 2004; Rockoff, 2004; Zimmer et al., 2004).
2003). Our analytical and simulation results suggest that jointly modeling scores from multiple subjects increases the information available for controlling for student heterogeneity, which can lead to more effective bias compression for all estimated treatment-by-subject effects simultaneously. The benefits of exploiting the redundancy in repeated measures of multiple outcomes is also noted by Thum (1997). Jointly modeling multiple subjects also helps to make the larger values of \( T \) considered in our empirical examples more tenable. While following students for \( T = 20 \) years is unrealistic, using repeated measures on a vector of annual outcomes from different subjects or different tests of the same subject makes achieving large numbers of test scores feasible, particularly with the growing availability of rich longitudinal data sets.

The ability of the mixed model approach to perform well simultaneously for a number of parameters that grows as the number of test scores grows is particularly relevant for estimating teacher effects. For example, the TVAAS model (Sanders et al., 1997) as applied in Tennessee and elsewhere simultaneously models up to 25 scores for individual students (five subjects for five years), and estimates separate teacher-by-subject effects, analogous to Example 4. William Sanders (personal communication) has claimed that jointly modeling 25 scores for students, along with other features of the TVAAS approach, is extremely effective at purging student heterogeneity bias from estimated teacher effects. The analytical and simulation results presented here largely support that claim.

It is however important to note that models such as TVAAS, and the cross-classified model of Raudenbush and Bryk (2002), and the variable persistence model of McCaffrey et al. (2004) differ in one notable way from the models considered in this article: the effects of individual teachers are modeled as random effects rather than unknown parameters of indicator variables in a regression model. That is, these models can be viewed as “random-random” estimators because both teacher and student effects are treated as part of the model error term. In the TVAAS model and variable persistence models, student heterogeneity is captured by allowing the residual errors to have an unstructured covariance matrix that is estimated from the data, and in the cross-classified model, student heterogeneity is model via student random intercepts and slopes. In all these models the data provide an estimate of the covariance matrix \( \mathbf{R} \) of the residuals \( \mathbf{r} \) after controlling for any other covariates in the model, and the teacher effects are estimated by their best linear unbiased predictors (BLUPs) given by the equation

\[
\hat{\theta} = \left( \mathbf{Z}' \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \right)^{-1} \mathbf{Z}' \mathbf{R}^{-1} \mathbf{r}
\]

Here \( \mathbf{Z} \) here is analogous to those considered in Example 4, providing linkages of individual teacher effects to present and potentially future scores of individual students. \( \mathbf{G} \) is the covariance matrix of the random teacher effects, generally taken as diagonal with different blocks of teacher effects (e.g. those from the same year and subject) having a common variance that is estimated from the data. Comparing this equation to the mixed model estimator in Equation 4 shows that the
estimators differ only by the addition of $G^{-1}$ to $Z' R^{-1} Z$. This results in the so-called “shrinkage” of the estimated teacher effects toward their means (typically zero), which results in estimates that are biased but have lower mean squared error relative to treating teachers as fixed regressors. While we did not consider this class of estimators in our Example 4, simulation studies not reported here and the structural similarity of it to the mixed model estimator that treats teacher effects as fixed suggest that our findings about the bias compression in that case are likely to carry over to the random-random estimator.

The primary advantage of the random-random estimator is that shrinking the estimators can reduce the average mean squared error in estimated coefficients or teacher effects. However, the random-random estimator can be biased by controlling for student or classroom level covariates when unobserved teacher effects are correlated with the observed values, for example, if better teachers are more likely to teach in schools serving students from higher income families (Ballou et al., 2004). With the mixed models considered in this paper, treating teacher effects as fixed allow for controlling for additional students and classroom variables without introducing bias. Post-hoc shrinkage can then be performed to reduce the variability of the resulting estimators and improve the mean squared error (Kane et al., 2006).

One critical caveat of our findings is that there are circumstances in which the mixed model approach can fail. Our empirical examples showed cases in which the mixed model approach is unable to remove bias, even with increasing amounts of test score data on individual students. All of these circumstances occurred when student treatment status did not vary over time. This is not a particular indictment of the mixed model approach relative to fixed effects; the fixed effects estimator is similarly ineffective because the treatment variable is confounded with the average of the unobserved student effects for treated students. Intuitively it makes sense that using repeated measures on students to control for student heterogeneity generally is not going to be effective under these circumstances because there is no within-student information upon which to gauge an individual student’s performance in the absence of treatment.

A related circumstance when random effects and mixed models cannot control for bias due to unobserved student heterogeneity is when the population is stratified as described in McCaffrey et al. (2004). A stratified population is one in which there are disjoint groups of students such that students within a group share teachers but students in different groups never share any teachers. For examples, students from different school districts where there is no interdistrict transfer are a stratified population. As discussed in McCaffrey et al. (2004), differences in strata means are not removed by pre-multiplication by $R^{-1}$. This is because the teacher effects from different strata form “treatments” that are constant for students across the entire time period of the data and, as with constant treatment assignments, random effects cannot yield unbiased estimates. It is important to note that fixed effects estimators are unidentified for stratified populations and require additional constraints that make interpretation of resulting estimates difficult.

Another potential limitation of our results is that our analytical results, and all of our examples
other than Example 2, treated \( R \) as known, whereas \( R \) must be estimated from the data in nearly all practical settings. When \( T \) scores are being modeled, \( R \) has \( T(T+1)/2 \) unknown parameters if no additional parametric structure is imposed, and unless the number of students is large relative to \( T \) it is likely that \( R \) may be estimated with substantial error. Moreover there are cases when \( R \) may not even be estimated consistently, such as when students do not switch treatment status over time as in Examples 1 and 3. Our results from Example 2 where we compared the bias compression from the mixed model estimator using both the known \( R \) and \( R \) estimated from the data warrant cautious optimism, at least about the effect of estimation error in \( R \). In that case the relative abilities of these two estimators to compress bias were almost identical, even for \( T = 20 \), where our estimated \( R \) had 210 parameters and only 1000 students were used in the simulation. We conjecture that for the \( R \)s that are likely to exist in longitudinal student achievement data series - where a large portion of the residual variance is dominated by a low-dimensional student-specific factor - that estimation error in \( R \) is not likely to substantially degrade the bias compression property of the mixed model estimator because the redundant information available in the vector of scores is likely to lead to estimates of \( \hat{y}_{it} \) that are largely insensitive to error in the estimated coefficients of the regression adjustment performed by \( R^{-1} \). Further exploration of this issue in settings more complex than that considered in Example 2 is an important area for future research.

Finally, all of the cases considered here had balanced, completely observed score data for all students. Actual data sets invariably contain a substantial amount of missing test score information, and when multiple cohorts are being modeled, it is common for different cohorts to have different configurations of available test scores. To explore the effects of missing data, we expanded Example 2 by adding cases where 50 percent of the data are missing at random. The bias continues converge to zero with increasing \( T \) but the decay is reduced so that the bias with \( T = 20 \) and 50 percent of observations missing is similar to the bias with \( T = 15 \) and all the data are observed.

This suggests that our general findings about the bias compression of the mixed models approach are not invalidated by the complexities of missing data, but it is likely that incompleteness in the test score data will in general degrade the bias compression to some extent. On the other hand, the mixed models approach makes use of all of the information available for each student in estimating the unknown parameters - in essence estimating \( \hat{y}_{it} \) from the regression on the available scores for each student - and so might lead to particular efficiency gains relative to fixed effects estimation when missing data are substantial. A specific case where this may be true is when using fixed effects on student gain scores rather than student level scores (Harris & Sass, 2006; Zimmer et al., 2003). This approach has strict data requirements because multiple year-to-year gain measures must be available on a student in order for that student to contribute to the estimation of the model parameters, a restriction that can result in substantial reduction in the amount of usable information in the data when data are missing. The mixed model estimator does not impose similar restrictions. Understanding how the mixed model estimator behaves in terms of both bias and variance relative to fixed effects methods when data are incomplete warrants further consideration.
6 References


### 7 Appendix A

**Theorem:** Consider the structural model in Equation 2 and let the fixed effects estimator for the $k$-dimensional parameter $\theta$ based on data from $T$ years from $n$ students be as specified in Equation 3 and the random effects estimator for known values of the variance components (and thus $\rho$) be given by Equation 6. Assume that $\Omega = \frac{1}{T} Z'(I - H_D)Z$ is nonsingular for all $T$ and converges as $T \to \infty$ to a nonsingular matrix. Further assume that for each $T$, the absolute values of the
elements of the \((nT \times k)\) design matrix \(Z\) are bounded by a number \(a < \infty\) independent of \(T\). Then for any \(\rho > 0\), \((\hat{\theta}_F - \hat{\theta}_R) \overset{p}{\to} 0\) as \(T \to \infty\).

**Proof:** All matrices depend on \(T\) so we suppress that notation. Rewrite the random effects estimator in Equation 6 as

\[
\hat{\theta}_R = (Z'(I - H_D)Z + (1 - \gamma T)Z'H_DZ)^{-1}(Z'(I - H_D)Y + (1 - \gamma T)Z'H_DY)
\]

\[
= \left(\Omega + \frac{(1 - \gamma T)}{T^2}Z'DD'Z\right)^{-1}\left(\frac{1}{T}Z'(I - H_D)Y + \frac{(1 - \gamma T)}{T^2}Z'DD'Y\right)
\]

where the second line follows from the fact that \(H_D = \frac{1}{T}DD'\). By the Schur complement formula (Searle et al., 1992), the inverse of the matrix on the left is \(\Omega^{-1} - \frac{(1 - \gamma T)}{T^2}\Gamma\) where

\[
\Gamma = \frac{1}{T^2}\Omega^{-1}Z'D\left(I + \frac{(1 - \gamma T)}{T^2}D'\Omega^{-1}Z'D\right)^{-1}D'\Omega^{-1}
\]

Then

\[
\hat{\theta}_R = (\Omega^{-1} - (1 - \gamma T)\Gamma)\left(\frac{1}{T}Z'(I - H_D)Y + \frac{(1 - \gamma T)}{T^2}Z'DD'Y\right)
\]

\[
= \hat{\theta}_F + (1 - \gamma T)\left(\frac{1}{T^2}\Omega^{-1}Z'DD' - \frac{1}{T}\Gamma Z'(I - H_D) - \frac{(1 - \gamma T)}{T^2}\Gamma Z'DD'\right)Y
\]

\[
= \hat{\theta}_F + (1 - \gamma T)CY
\]

By the boundedness assumption on the elements of \(Z\), the elements of \(\frac{1}{T^2}D'\Omega^{-1}Z'D\) are \(O(1)\), and thus the elements of \(\Gamma\) are also \(O(1)\). Analogous considerations for the remaining components of the matrix \(C\) give that \(CY\) is \(O_p(1)\). For any \(\rho > 0\), \((1 - \gamma T) = o(1)\) as \(T \to \infty\). Thus \((\hat{\theta}_F - \hat{\theta}_R) = o(1)O_p(1) = o_p(1)\) as \(T \to \infty\).

### 8 Appendix B

**Theorem:** Let \(A_1\) and \(\Psi_1\) be defined as in Section 3. Then sufficient conditions for the elements of \(R^{-1}A\) to go to zero uniformly as \(T \to \infty\) are:

1. The smallest eigenvalue of \(A_1'\Psi_1^{-1}A_1\) goes to infinity as \(T \to \infty\); and
2. There exists a number \(C\) independent of \(T\) such that the elements \(a_{it}\) of \(\Psi_1^{-1/2}\), the symmetric square root of \(\Psi_1^{-1}\), satisfy \(\sum_{t=1}^{T} |a_{it}| < C\) for all \(i\)

**Proof:** Throughout, all matrices except \(S_1\) and its root are assumed to depend on \(T\) so we suppress that notation. Because \(R^{-1}A = (I_n \otimes R_1^{-1})(I_n \otimes A_1) = I_n \otimes R_1^{-1}A_1\), it is sufficient to consider
only the elements of $R_1^{-1}A_1$. Because all matrices would thus be subscripted by “1”, we suppress that notation as well and use, for example, $R$ for $R_1$, $A$ for $A_1$, $Ψ$ for $Ψ_1$, etc. We also assume that all matrix roots are symmetric roots.

Recall that $R = ASA' + Ψ$. By the Schur complement formula (Searle et al., 1992)

$$R^{-1} = Ψ^{-1} [I - AS^{1/2}(I + S^{1/2}A'Ψ^{-1}AS^{1/2})^{-1}S^{1/2}A'Ψ^{-1}]$$

so that

$$R^{-1}A = Ψ^{-1}A [I - S^{1/2}(I + S^{1/2}A'Ψ^{-1}AS^{1/2})^{-1}S^{1/2}A'Ψ^{-1}A]$$

$$= Ψ^{-1}AS^{1/2} [I - (I + S^{1/2}A'Ψ^{-1}AS^{1/2})^{-1}S^{1/2}A'Ψ^{-1}AS^{1/2}] S^{-1/2}$$

Let $X = Ψ^{-1/2}AS^{1/2}$ of dimension $(T \times d)$. Then

$$R^{-1}A = Ψ^{-1/2}X [I - (I + X'X)^{-1}X'X] S^{-1/2}$$

Singular value decompose $X$ as $UΛ^{1/2}V'$ where $U$ is $(T \times d)$ with orthonormal columns, $Λ^{1/2} = diag(\sqrt{λ_1}, \ldots, \sqrt{λ_d})$, and $V$ is $(d \times d)$ and orthogonal. Because $A$ is assumed to have full column rank, $X$ has full column rank and so $\sqrt{λ_m} > 0$ for $m = 1, \ldots, d$. Note that $X'X = S^{1/2}A'Ψ^{-1}AS^{1/2} = VΛV'$ where $λ_1, \ldots, λ_d$ are the eigenvalues of $X'X$.

Now consider the matrix

$$[I - (I + X'X)^{-1}X'X] = [I - (I + VΛV')^{-1}VΛV']$$

$$= [VV' - V(I + Λ)^{-1}V'VΛV']$$

$$= [V(I - (I + Λ)^{-1}Λ)V']$$

$$= [Vdiag(\frac{1}{1 + λ_1}, \ldots, \frac{1}{1 + λ_d})V']$$

Thus

$$R^{-1}A = Ψ^{-1/2}UΛ^{1/2}V' [Vdiag(\frac{1}{1 + λ_1}, \ldots, \frac{1}{1 + λ_d})V'] S^{-1/2}$$

$$= Ψ^{-1/2}UΛ^{*}V'S^{-1/2}$$

where $Λ^*$ is $diag(\sqrt{λ_1}/(1 + λ_1), \ldots, \sqrt{λ_d}/(1 + λ_d))$.

An arbitrary element $r_{ij}$ of $R^{-1}A$ is the inner product of a row $a_i'$ of $Ψ^{-1/2}$ and a column $b_j$ of $UΛ^*V'S^{-1/2}$. Let $s$ be the absolute value of the largest element of $S^{-1/2}$. By condition 1 of the theorem and Lemma 1, all the eigenvalues of $X'X$ are getting arbitrarily large for sufficiently large values of $T$, so that for any $ε > 0$ there exists a $T_ε$ such that $\sqrt{λ_m}/(1 + λ_m) < ε/Csd^2$ for
all \( m = 1, \ldots, d \) and for all \( T > T_\epsilon \). Because \( V \) is an orthogonal matrix and the columns of \( U \) are orthonormal, their elements cannot exceed 1 in absolute value, and so the absolute value of the largest element of \( U \Lambda^* V S^{-1/2} \) is bounded by \( \epsilon/C \) for all \( T > T_\epsilon \). Then, for any \( i \) and \( j \) and for all \( T > T_\epsilon \),

\[
|r_{ij}| = |a'_i b_j| = \left| \sum_{t=1}^{T} a_{i,t} b_{j,t} \right| \leq \sum_{t=1}^{T} |a_{i,t}||b_{j,t}| \leq (\epsilon/C) \sum_{t=1}^{T} |a_{i,t}| < \epsilon
\]

where the last inequality follows by condition 2 of the theorem.

**Lemma 1.** Let \( B \) be a positive definite \( d \times d \) matrix and let \( B^{1/2} \) be a symmetric root of \( B \). Let \( M_T \) be a matrix sequence of \( d \times d \) matrices. Let \( \lambda_T \) denote the smallest eigenvalue of \( M_T \) and \( \omega_T \) denote the smallest eigenvalue of \( Q_T = B^{1/2} M_T B^{1/2} \). Then \( \lambda_T \to \infty \) as \( T \to \infty \) if and only if \( \omega_T \to \infty \) as \( T \to \infty \).

**Proof:** Suppose \( \lambda_T \to \infty \) as \( T \to \infty \). Let \( \psi_{\text{min}} > 0 \) denote the smallest eigenvalue of \( B \). Then for every \( d \)-vector \( x \) such that \( x' x = 1 \)

\[
x' Q_T x = (x'Bx) \frac{x'Q_T x}{x'Bx} \\
\geq \psi_{\text{min}} \frac{x'Q_T x}{x'Bx} \\
= \psi_{\text{min}} \frac{u'M_T u}{u'u}, \text{ where } u = B^{1/2} x \\
\geq \psi_{\text{min}} \lambda_T.
\]

Because \( B \) is positive definite, division by \( x'Bx \) is well defined. By assumption \( \psi_{\text{min}} \lambda_T \to \infty \) with \( T \) so that \( \omega_T \), the minimum of \( x'Q_T x \), must converge to infinity as well.

Now suppose that \( \omega_T \to \infty \) as \( T \to \infty \). Let \( \psi_{\text{max}} > 0 \) denote the largest eigenvalue of \( B \). Then for any \( d \)-vector \( a, a'Ba \leq \psi_{\text{max}} a'a \). Now, let \( u \) be a vector such that \( \lambda_T = (u'M_T u)/u'u \) and let \( x = B^{-1/2} u \) so that \( B^{1/2} x = u \). Then

\[
\lambda_T = \frac{x'B^{1/2} M_T B^{1/2} x}{x'Bx} \\
\geq \frac{x'B^{1/2} M_T B^{1/2} x}{\psi_{\text{max}} x'x} \\
\geq \frac{\omega_T}{\psi_{\text{max}}} \to \infty \text{ as } T \to \infty.
\]