A COMPARISON AMONG MAJOR VALUE-ADDED MODELS: A GENERAL MODEL APPROACH

by

YUAN HONG

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

written under the direction of

Jimmy de la Torre

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2010

ABSTRACT OF THE DISSERTATION

A Comparison among Major Value-Added Models:

A General Model Approach

By YUAN HONG

Dissertation Director:

Jimmy de la Torre

Value-added models (VAMs) are becoming increasingly popular within accountability-based educational policies as they purport to separate out the effects of teacher and schools from student background variables. Given the fact that evaluations based on the inappropriate use of VAMs would significantly impact students, teachers and schools in a high-stake environment, the literature has advocated empirical evaluations of VAM measures before they become formal components of accountability systems. The VAM label is attached to a number of models, which range from simple to highly sophisticated models. However, in practice, educators and policymakers are often being misled into believing that these approaches give nearly identical results, and making decisions without understanding the strengths and limitations of these models. In addition, the empirical evaluations to date have shown that the VAM measures of teacher effects are sensitive to the form of the statistical model and to whether and how student background variables are controlled.

This study proposes a multivariate joint general VAM to investigate the issues

raised by the applications of all the currently prominent VAMs, which can be seen as restricted cases of this general model. The general model provides a framework for comparing the restricted models and for evaluating the sensitivity of VAM measures (e.g., teacher and school effects) to the model choice. Markov chain Monte Carlo algorithm is used in a Bayesian context to implement both the general and the restricted models.

A simulation study was conducted to investigate the feasibility and robustness of the general model when the data were generated under varying assumptions. For each condition, three consecutive years of testing scores were generated for 400 students grouped into 16 classes. Real data consisting of three years of longitudinally linked student-level data from a large statewide achievement testing program were also analyzed. The results show that the proposed general model is more robust than other models to different assumptions and the inclusion of the background variable has significant impact on some models when the school/class has an unbalanced mix of advantaged and disadvantaged students.

Acknowledgement

First, I thank my advisor, my dissertation Chair, Prof. Jimmy de la Torre, for his continuous support in the Ph.D. program. He has always been there to listen and to give advice. His firm belief that persistence is needed for any work has influenced my approach to thesis writing, as well as to my daily life.

I would also like to extend my sincere gratitude to my committee members: Prof. Gregory Camilli, who always asked good questions and gave advices mixed with a sense of humor about life, Dr. Lihua Yao, for her willingness to discuss this topic at its initial stage, her friendship and encouragement, and Prof. Bruce Baker, who gave insightful comments and reviewed my work on a very short notice.

A special thank to CTB/McGraw-Hill company, who recognized the value of my work, provided with monetary support and allowed me to access to the real testing data. I would especially like to thank the scientists at CTB, Dr. Richard Patz, Prof. Wim van der Linden and Dr. Daniel Lewis for providing perspective comments on this work.

Many thanks to my friends, Lei, Peijia, Lu, Haihui and Kui, for always being there when I needed you.

Last, but not least, I thank my family: my mom, Pingli, for her unconditional love and meticulous care, my grandparents, for educating me with aspects from both arts and sciences, and my husband, Zhaohua, for listening to my complaints and frustrations, and for believing in me.

Table of Contents

Lists of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Background

Written with the intention and spirit of ensuring all children are reached by America's public schools, No Child Left Behind (NCLB) is the federal government's mandate that all students are considered proficient by 2014 in reading/language arts and mathematics. The law outlines and requires a scientific and systematic approach to achieving reform and improvement in all areas of school life. To garner compliance, each school receiving Title 1 funding must develop and adopt assessments and procedures to evaluate the annual performance of schools at the state wide level on a variety of indicators, the most important of which is academic (Henderson-Montero, 2003). Harsh sanctions are imposed for failure to make steady, demonstrable progress toward improving student achievement (Wanker, 2005).

At the heart of NCLB is the development, at the state level, of content standards linked to assessments for reading/language arts and mathematics. There are 40 key requirements of No Child Left Behind, the most highly publicized and debated is Adequate Yearly Progress (AYP), which required states to start testing students in grades 3 through 8 in mathematics and reading/language arts by the 2005-2006 school year, and reach 100% proficiency by 2013-2014 (Wanker, 2005).

But though NCLB has been promoted, in part, as a way of equalizing better-off, predominately white schools with low-income, largely minority ones, it provides no methods for improving schools that are lagging. Nor does it identify the teachers who are most effective-who deserve recognition and whose skills should be emulated. A major criticism of NCLB is that it mandates student achievement without offering

methods for obtaining it (Carey, 2004). Due for reauthorization, many experts are recommending the addition of value-added assessment (VAA) to the new legislation as a way to track growth of each student since one of the greatest criticisms of AYP is that it aims for a goal with no commitment to growth (Barton, 2004). The proponents of value-added modeling call its results fairer and more accurate than those produced by AYP, which is currently based entirely on standardized test scores.

Value-Added Modeling

Value-added modeling (VAM), also known as VAA, is a method of measuring student academic progress over time even after the proficient level has been reached. "Value-added assessment system" does not refer to one particular test format. Rather, value-added refers to any one of several models that are used to interpret test scores in a way that evaluates the growth or progress in a student's academic achievement over time, usually over several academic years (Rubin, Stuart, & Zanutto, 2004).

History of VAM

Developed by Tennessee statistician Dr. William Sanders, value-added was first used in the field of agriculture genetics. Learning of the controversies in public education in the early 1980's, Sanders and his group felt they could actually apply their knowledge to education and showed that growth modeling was a great improvement over a single cut-score on a standardized test. Appealing directly to the governor of Tennessee, Sanders was awarded rights to assessment results of students in Knox County Schools and was able to simultaneously measure teaching and student effects using previous test results.

Relying on pilot studies that Sanders and his colleagues conducted on the value-added model during the 1980s, the Tennessee legislature embraced the model as its

methodology of choice for measuring the performance of students, teachers, schools, and school systems. The legislation defines the Tennessee Value-Added Assessment System (TVAAS) as a "statistical system for educational outcome assessment which uses measures of student learning to enable the estimation of teacher, school, and school district statistical distributions." (Kupermintz, 2003) TVAAS becomes the centerpiece of an ambitious educational reform effort implemented by the Tennessee Education Improvement Act of 1992. Since then TVAAS has been credited with leading to the implementation of value-added assessment systems in states, districts, and nationwide (Carey, 2004; Hershberg, Simon, & Lea-Kruger, 2004; Kupermintz, 2003).

Recently, more than a dozen states, including Colorado, California, Florida, Ohio, New York, Pennsylvania and Michigan are studying, and in some cases, applying value-added modeling. The U.S. Department of Education has accepted applications from up to ten states to meet their part of their AYP with value-added modeling. Beginning in the fall of 2006, the AYP in those states' schools is calculated by using both the new progress method and the usual standardized tests.

VAM versus Simple Growth Scores

The concept of an assessment that measures a student's achievement growth over several years, commonly known as longitudinal assessment, has long existed in education (Goldschmidt, Choi, & Martinez, 2003). However, value-added assessment represents an approach to evaluating student achievement growth that is distinct from traditional growth models. Currently, schools that miss AYP, those missing by a small amount along with those missing by a large amount, are both labeled as "failing" regardless of any demonstrable progress. Growth models help move beyond the current "blame game" and instead highlight areas in need of improvement regardless of whether

or not the school is already strong (Gooden & Nowlin, 2006). VAM can be considered as a special growth model because it measures the individual progress of schools and students. A growth score is typically calculated as the difference between a student's scores for the current year and the previous year. VAM is more statistically complex because it is intended to separate out the non-educational factors, such as student's demographics and socio-economic status (SES). Once these factors are isolated, their impact is removed from the measure of the student's achievement growth. Then, the student's true achievement growth can be attributed to the educational practice of the district, school and teacher (Drury & Doran, 2004; Hershberg et al., 2004; McCaffrey et al., 2003). Therefore, from the VAM's perspective of view, schools would give credit for increasing student achievement, even when their AYP is missed.

More Applications of VAM

VAM holds great promise because it claims to separate out the school effect, the teacher effect, and the student's own effect that together contribute to student progress. It is the highly interactive relations between these effects that make VAM so attractive. Researchers also believe growth information can be instructional in improving practice. Value-added measures can provide valuable information about the effects of curriculum, instructional techniques and other instructional practices. Using the data, teachers and administrators can determine areas of success and improvement and work to best meet the needs of their students. In addition, administrators can analyze the data and target professional development for staff, or use it as the basis for school improvement plans (Hershberg, Simon, & Lea-Kruger, 2004). The same data from VAM can also provide principals with valuable information for assigning students to specific teachers. From the data results, teachers, grade levels, groups of students (learning disabled or gifted)

can be identified and then a precise match between the teachers' individual strengths and the students' needs can be achieved (Hershberg et al., 2004). Certain value-added measures can also be used to evaluate teacher preparation programs in public universities. Berry and Fuller (2006) have researched a Value-Added Teacher Preparation Program Assessment Model that has the capacity to connect growth in student learning to public university teacher preparation programs.

Statement of the Problem

The "value-added modeling" label is attached to a number of models, which range from being roughly simplistic to very sophisticated. Differences among these models stem from the efforts by statisticians to resolve various technical problems. None of the models can solve all the technical problems and no single approach has been proved superior to any other. At this point, even the experts on these models have no agreement on the appropriateness of each model. However, in practice, educators and policy makers are often being misled into believing that these approaches give nearly identical results and making decisions without understanding their strengths and limitations. Evaluations based on the inappropriate use of VAM would significantly impact students, teachers and schools in a high-stake environment. Thus, the literature has advocated empirical evaluations of VAM measures before they become formal components of accountability systems or are used to inform high stakes decisions about teachers and students (Braun, 2005; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). The empirical evaluations to date have considered the sensitivity of VAM measures of teacher effects to the form of the statistical model (McCaffrey, Lockwood, Mariano, & Setodji, 2005; Rowan, Correnti, & Miller, 2002) and to whether and how student background variables are controlled (Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). Lockwood

et al. (2007) considered the sensitivity of estimated VAM teacher measures to two different subscales of a single mathematics achievement assessment.

Although the studies above have yielded results that are meaningful in practice, the different approaches provide partial and fragmented answers only to the problem of interest, and are inevitably affected by the limitations of the methods involved. Moreover, the test data used in each study are different, making it difficult to attribute the results to either the methods or the data examined. Therefore, it might be inappropriate to consolidate their findings to be a systematic whole. This dissertation aims to provide more systematic and thorough investigation on the sensitivity of the VAM results to different methods through a general model approach.

McCaffrey et al. were concerned with creating a system for classifying VAM models as a means of specifying the conditions under which one or another would be a valid methodology. They were concerned with creating a system for classifying VAM models as a means of specifying the conditions under which one or another would be a valid methodology. Their approach was also to specify a general model, and then show how different models suggested by themselves or others would be special cases of this general model.

As the most complex special case of the McCaffrey et al. general model, the variable persistence model (Lockwood et al., 2004) poses computational challenges that render likelihood methods practically infeasible for all but small data sets. To address this problem, Lockwood et al. propose a Bayesian formulation of the variable persistence model that scales well to the extremely large and complex data sets that challenge alternative approaches to parameter estimation. Another contribution of their study is that their formulation includes an extension to jointly modeling outcomes from

multiple tested academic subjects (e.g., mathematics and reading) in each year, which has been proven to provide higher quality parameter estimates.

Inspired by both McCaffrey et al.'s study and Lockwood et al.'s study, this work proposes a multivariate joint general VAM model to investigate the issues raised by the application of all the currently prominent VAM models, which can be seen as restricted cases of this general model. The general model provides a framework for comparing the restricted models and for evaluating the sensitivity of VAM measures (e.g. teacher effect, school effect) to the model choice. The general model is estimated under the Bayesian framework. Although the less complex restricted models other than the variable persistence model could have been estimated using maximum likelihood method, all the restricted models are estimated using MCMC algorithm under Bayesian framework for comparability purposes. That is, by using same estimation method we can attribute the differences we observed on the differences in model specifications rather than the difference in the estimation methods.

CHAPTER 2
LITERATURE REVIEW

This chapter is organized as follows: Section I provides an overview of the theoretical ground of VAM. Section II identifies the most important problems existing in the VAM application. Section III presents a thorough review of several principle existing VAM models, comparing their underlying assumptions, model specifications, strengths, weaknesses and potential problems in their use. Finally, section IV explores the possibility of more general VAM approaches to evaluate the school/teacher effect by summarizing the major factors that influence the features of different VAM methods and several empirical studies in this direction are discussed.

An Overview of the Theoretical Ground of VAM

The question of how to evaluate school and teacher effectiveness is fundamental to educational policy and practice. Our common practice is to compare schools or teachers by comparing unadjusted mean levels of achievements or the percent of students in a school or class who are classified as proficient. As Ballou, Sanders, and Wright (2004) note, it is unfair to hold schools accountable for mean achievement levels when students enter those schools with large variances in achievement. Moreover, changes in mean achievement at the school level may have little relation to instructional effectiveness if the mobility of students across schools is remarkable.

There is a common agreement in the VAM literature is that the contributions of school and teacher to student learning be estimated. The literature advocates that we should compare schools or teachers by comparing their “value added” to student learning gains rather than by comparing the mean level achievement. The value-added philosophy is to hold schools and teachers accountable for the learning gains of students they serve. The philosophy seems simple, but the underlying technique details are

numerous. The foremost question that should be clarified is: "What are VAMs trying to estimate?" Raudenbush (2004) endeavored to answer this question from a potential outcomes view.

The student's potential outcomes would be a function of pre-assigned student characteristics, S, random error, e, and two aspects of schools: school context, C, which contains the social environment of the school and the social composition of the school, and school practice, P. What is the interested by the policymakers and district officials is the component P, over which the school leaders and teachers have direct influence. Although school administration and teacher instruction have no or little direct influence on C, C and P are highly correlated factors.

According to Raudenbush and Willms (1995), the first or Type A effect is the difference between a child's potential outcome in school j and that child's potential outcomes in school j'. Type A effect is expected to be estimated from the experiment in which the students having a common S are randomly assigned to school j or j'. With the assumption of randomization, the expected estimate of the difference between two schools would depend only on C, P, C', P'. In contrast, the Type B effect is the difference between student i's potential outcome in school j when school practice P_j is in operation and when school practice P_j is in operation. As Raudenbush and Willms point out, it does not seem to be possible to separate teacher and school effects using currently available accountability data. Therefore, "VAM are best aimed at assessing the Type A effect defined as combined effects of context and practice at the classroom and school levels. A useful way to do so is to view each student as possessing a smooth trajectory that would describe that students' growth if that student encountered

average teachers and schools. The Type A effect in any year is then defined as a deflection from this expected curve" (Raudenbush & Willms, 1995, p. 124).

The figure shown in Raudenbush (2004) displayed the fundamental idea of VAM from a potential outcomes view. There is a hypothetical student's expected trajectory from time point t to $t+2$ given "average" schools and teachers. If instead, this student has an above (or below) average observed score $Y_{t+1}^{(j)}$. The difference between the observed and expected score $Y_{t+1}^{(j)} - Y_{t+1}^{(0)}$ is associated with the attendance in school j. Therefore, the student's gains over years can be partitioned into two parts: the part attributed to student's own expected gain and the part attributed to the school or teacher effect.

Major Issues Arising from the Use of VAM

As mentioned in the previous section, there are a number of different models in use in different accountability systems. Differences in the models stem from efforts by researchers to resolve the various technical problems that have arisen in this field. Before reviewing and assessing some of these models, this section will first identify the most important problems raised by the use of VAM to estimate school or teacher effect.

The first, and perhaps most significant problem, is that the students are not randomly assigned to schools or teachers. The characteristics of students and communities are correlated with classrooms and schools. That is, for example, the most effective teachers tend to be able to select their assignments (Goldhaber, 2004; Hibpshman, 2004b), and as a result are more likely to have highly motivated students. It is difficult to determine whether the higher average levels of achievement of their students is due to teacher's instruction, or to the highly motivated students they teach.

The second problem has to do with the uncertainty about which variables are

important to the models. Educators, researchers and policymakers have long been recognized that schooling is only one of many factors that affect student learning. The numerous family background characteristics and social environment factors are also strong predictors of student achievement. Shkolnick et al. (2002) showed that background characteristics predict gains for some population. However, their relationship with growth or gains measured by VAM has not been explicitly defined. McCaffrey et al. (RAND; 2003) stated that these characteristics variables could be confounded with the teacher effects and therefore bias estimates of teacher effects, and the question of whether it is necessary to include these variables in the model has been an important point of debate in the VAM literature.

The third problem is that of the complexity of the models. Researchers are interested in determining how much difference there is in the effect scores produced by the simpler and more complex models. Some researchers note that considerations that may have importance in theory may make little difference in a practical sense, and if a simpler model produces results comparable to more complex models, it may be preferable because of its intuitive appeal. On the other hand, some researchers believe the benefits of using a complex model because they believe the important issues, which can affect the evaluation of school or teacher effectiveness, need to be addressed by more advanced models. For example, the key feature of longitudinal achievement data for modeling teacher contributions to student achievement is the sequential regrouping of students into different classrooms with different teachers. The results in data where students who are nested under a common teacher for one measurement are not nested together for another measurement. Moreover, scores for students who share a common teacher at one point in time might continue to be positively correlated at subsequent test administrations. The

resulting model structures necessary to accommodate these complexities are known as "multiple-membership" models (Browne, Draper, Goldstein, & Rasbash, 2002; Rasbash & Browne, 2002) because individual scores depend on the effects from multiple "members" of the grouping units (e.g., past and current teachers).

The various models will be reviewed in the following section are designed in response to one or more of these problems. However, none of the models solves all of these problems.

Review of the Principal Existing VAM Models

Several major existing models are reviewed in this section. They are gain score (GS), covariant adjustment (CA), layered (LA), cross-classified (CC), and persistence (PS) model. The GS and CA model are considered more generic and more widely used than the other models. The GS and CA model have been referred to as "single wave" or "univariate" models, as they only use two points in time. The CC, LA and PS models all utilize more than two points in time and they have been referred to as "multiple wave" or "multivariate" approaches.

CA Model

In the CA model (Rowan, Correnti, & Miller, 2002; Diggle, Liang, & Zeger, 1996; Meyer, 1997), prior scores are used as the covariant in the model, with the current score as a function of prior year score and is linked to only the current teacher. Student i's score at the gth grade is modeled as follows:

$$y_{ig} = \mu_g + \beta_g x_i + \gamma_{ig}^{'} z_{ig} + \gamma y_{ig-1} + \theta_g + \varepsilon_{ig} \qquad (1)$$

where the θ_g denotes the grade g teacher effect of the current teacher, which is measured by the deviation in class-level mean from the overall system mean. The

x_i and z_{ig} are time invariant and time varying covariates for student i. The time invariant variables include student-level covariates such as gender and ethnicity. The time varying variables may include family income and testing circumstances. The teacher effect θ_g is considered either fixed or random normal with mean zero and variance, $\sigma^2_{\theta_g}$. β_g and γ'_{ig} are vectors that contain the coefficients associated with the student's background variables. The ε_{ig} are *i.i.d.* $N(0, \sigma^2_{\varepsilon g})$ residual error terms. The residuals across years are assumed to be independent of each other. That is, $Corr(\varepsilon_{ig}, \varepsilon'_{ig}) = 0$ for $g \neq g'$. This assumption avoids the biased estimates of fixed effects by the standard mixed model estimation due to the correlation between the covariates and the residual error term.

Assuming the current teacher effect is the random effect, the expectation and variance conditional on the observed covariates and previous year's score, y_{ig-1}, are $E(y_{ig}) = \mu_g + \beta_g x_i + \gamma'_{ig} z_{ig} + \gamma y_{ig-1}$ and $Var(y_{ig}) = \sigma^2_{\theta g} + \sigma^2_{\varepsilon g}$.

The advantages of this approach are that 1) it is simple to understand and easy to use; 2) it can model the effects of previous year's experiences; and 3) previous year's teacher effects are estimated, not assumed. The disadvantages of this approach are that 1) it ignores student performance information from prior years; 2) students transfer or retain in grade are excluded; 3) there is no statistical adjustment for student ability; and 4) it does not take measurement error into account.

GS Model

The GS model is a special case of the CA model, which specifies gain score between two adjacent years as a function of the covariates and the current year teacher/school effect (Rowan et al., 2002; Shkolnik, Hikawa, Suttorp, Lockwood,

Stecher, & Bohrnstedt, 2002). Let $d_{ig} = y_{ig} - y_{ig-1}$, the model for grade g gains is

$$d_{ig} = \delta_g + \beta_g x_i + \gamma_{ig}^{'} z_{ig} + \theta_g + \varepsilon_{ig} \qquad (2)$$

The coefficient δ_g denotes the mean gain in grade g. The random teacher effect and residual error terms follow the same assumptions as the CA model. Setting $\gamma = 1$ and moving y_{ig-1} from the right side to the left side of Equation (1), the CA model becomes the GS model. The GS model has all the disadvantages of the CA model, but it is simple and easy to understand.

CC Model

Raudenbush and Bryk(2002) develop the CC model that explicitly specifies the cross-grade correlations and the effects of the multiple years of teachers on student scores. Moreover, they consider random linear growth trajectories for students. The CC model for student i score in grades g is

$$y_{ig} = \mu + g\gamma + \mu_i + g\gamma_i + \sum_0^g \psi_{ig}\theta_g + \varepsilon_{ig} \qquad (3)$$

For example, the student i's scores in grades 0 to 3 are

$$y_{i0} = \mu + \mu_i + \psi_{i0}\theta_0 + \varepsilon_{i0}$$

$$y_{i1} = \mu + \gamma + \mu_i + \gamma_i + \psi_{i0}\theta_0 + \psi_{i1}\theta_1 + \varepsilon_{i1}$$

$$y_{i2} = \mu + 2\gamma + \mu_i + 2\gamma_i + \psi_{i0}\theta_0 + \psi_{i1}\theta_1 + \psi_{i2}\theta_2 + \varepsilon_{i2}$$

$$y_{i3} = \mu + 3\gamma + \mu_i + 3\gamma_i + \psi_{i0}\theta_0 + \psi_{i1}\theta_1 + \psi_{i2}\theta_2 + \psi_{i3}\theta_3 + \varepsilon_{i3}$$

The trend of the overall mean for each grade is denoted by $\mu + g\gamma$. The εs are assumed to be *i.i.d.* normally distributed random variables with mean zero and variance

σ_ε^2. The teacher effects θs are assumed to be independently, normally distributed with a constant variance across years. The ψ_{ij} measures the proportion of grade 0 education provided to student i by teacher j. Each student's growth over grades is modeled with a linear trend $y_{ig} = \mu + g\gamma + \mu_i + g\gamma_i$ and the random intercepts and slopes are assumed normally distributed with mean zero and variance τ_{00}^2 and τ_{11}^2, and covariance τ_{01}.

The CC model fitted by Rowan et al. (2002) included time-varying covariates for participation in educational programs, (e.g., special education) and age. Their model also included time-invariant covariates for student ethnicity, family structure and socioeconomic status. The random effect included in their model was school effect rather than teacher effect. In this model, the score for the ith student in school j at time t, y_{ijt}, is given by $y_{ijt} = \alpha + \beta t + \delta t^2 + \alpha_j + \beta_i t + \alpha_{ij} + \beta_{ij} t + \gamma x_{ijt} + \theta_{1(ij)} + ... + \theta_{t(ij)} + \varepsilon_{ijt}$

where α_i and β_i are the random intercept and slope for the school; α_{ij} and β_{ij} are the random intercept and slope for the student; x_{ijt} denotes a vector of student characteristics, some of which might vary over time; $\theta_{1(ij)}$ and $\theta_{t(ij)}$ are the effects for the student's teachers at testing times 1 to t; and ε_{ijt} is a residual error term. Thus, scores are models by a common quadratic function of time $\alpha + \beta t + \delta t^2$ plus school-specific and student-specific random linear time trends. The model assumes no variability in the nonlinear component of the model, implicitly, any variation in δ is captured in the residual error term. A teacher effect, $\theta_{t(ij)}$, is added for each year and these effects remain in the model undiminished at the future tests, which is why the

model for the score at time t includes terms for all previous teachers. Rowan et al. acknowledge that the $\theta_{t(ij)}$ are residual classroom effects, although they are referred to as teacher effects.

LA Model

LA model, also called TVAAS, was developed by Sanders, Saxton and Horn (1997) to account for the complicated linkage of students to teachers or schools over time, and the correlation of future scores for students who shared a common past teacher or in a same school, which was referred to as cross-classified or multiple member ship. It is called the layered model because the model for later years adds layers to the model for earlier years. The model for student i score in grade g is

$$y_{ig} = \mu_g + \sum_{0}^{g} \psi_{ig}\theta_g + \varepsilon_{ig} \tag{4}$$

Therefore, the student i scores in grade 0 to 3 is

$$y_{i0} = \mu_0 + \psi_{i0}\theta_0 + \varepsilon_{i0}$$

$$y_{i1} = \mu_1 + \psi_{i0}\theta_0 + \psi_{i1}\theta_1 + \varepsilon_{i1}$$

$$y_{i2} = \mu_2 + \psi_{i0}\theta_0 + \psi_{i1}\theta_1 + \psi_{i2}\theta_2 + \varepsilon_{i2}$$

$$y_{i3} = \mu_3 + \psi_{i0}\theta_0 + \psi_{i1}\theta_1 + \psi_{i2}\theta_2 + \psi_{i3}\theta_3 + \varepsilon_{i3}$$

The ε_{ig} s are assumed normally distributed and independent across students. Within a student the variance-covariance matrix of the ε s is unrestricted allowing for different variance at each time point and possibly nonzero and nonconstant correlation of scores from different grades or years. The variance-covariance parameters are assumed constant across all students. The LA model allows the variance of school or teacher

effects to vary across grades and the correlation between scores from the same student across subjects (and grades). It also assumes that schools or teachers have separate and independent effects for each subject and these effects persist undiminished into all future test outcomes.

The CC model also accommodates the complex "multiple-membership" between students and teachers or schools, but differs from the LA model in several ways. For one, the CC model uses random growth curves to model to correlation among scores within a student, whereas the layered model accounts for this correlation with an unspecified covariance matrix. These two models share the common assumption that the random teacher or school effect persists undiminished for students' future performance.

PS Model

The PS model (McCaffrey et al., 2004) defines the persistence of the past teacher or school effects on the current achievement. In this regard, the LA and CC models can also be considered as special cases of the PS model in that the persistence is assumed to be a fixed value. This kind of PS model is referred to as "complete persistence model". They explicitly parameterize and estimate the strength of past teacher or school effects on the current scores rather than assuming them to be known. This kind of PS model is called "variable persistence model". This specification makes the PS model more complex and computationally challenging. Moreover, Lockwood et al. generalize the McCaffrey et al.'s PS model to multiple subjects per year and provides a multivariate formulation for it. A special case of this model that includes teacher effects but not school effects is the following:

Let $\mathbf{y}_i$ denote the vector of test scores for student i. $\mathbf{y}_i$ is of length ST, the number of subjects (S) times the number of years (T). $\mathbf{X}_i$ denotes the ($ST \times p$) design matrix of both time-invariant and time-varying student background variables for the p-dimensional vector of regression coefficients $\boldsymbol{\mu} = (\mu_{11},...,\mu_{ST},\beta_{11}^{'},...,\beta_{ST}^{'})$. The teacher effects are organized by subject and year as $(\boldsymbol{\theta}_{11}^{'},...,\boldsymbol{\theta}_{1T}^{'},\boldsymbol{\theta}_{21}^{'},...,\boldsymbol{\theta}_{2T}^{'},...\boldsymbol{\theta}_{S1}^{'},...,\boldsymbol{\theta}_{ST}^{'})$ of length n_θ, where $\boldsymbol{\theta}_{ST}^{'}$ provides the teacher effects for subject s in year t. The matrix $\boldsymbol{\Phi}_i$ specifies the linkage of students to teachers by subject. $\boldsymbol{\Phi}_i$ is ($ST \times n_\theta$) with only 0 or 1 entries and row sums equal to 1, with the nonzero element in each row corresponding to student i's teacher for a given year and subject. Hence, the contribution of teacher effects to the outcomes for student i is then given by $\mathbf{A}\,\boldsymbol{\Phi}_i\,\boldsymbol{\theta}_{ST}^{'}$, where $\mathbf{A}$ is a ($ST \times ST$) block diagonal matrix consisting of S distinct ($T \times T$) lower triangular blocks corresponding to subjects. The (t,t^*) element of the block for subject s is α_{s,tt^*} for $t > t^*$ and 0 otherwise, where α_{s,tt^*} denotes the teacher effect persistence parameters for subject s. When all α_{s,tt^*} are equal to 1, the model becomes the complete PS model.

Therefore, the distribution for a single student's score vector, $\mathbf{y}_i$, conditional on the model parameters, teacher effects, and all covariates and linkage information is

$$\mathbf{y}_i \left| \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma} \right. \sim N_{ST}(\mathbf{x_i}\boldsymbol{\mu} + \mathbf{A}\boldsymbol{\Phi}_\mathbf{i}\boldsymbol{\theta}, \boldsymbol{\Sigma}) \qquad (5)$$

where N_{ST} denotes the ST dimensional multivariate normal distribution and $\boldsymbol{\Sigma}$ is a ($ST \times ST$) unstructured positive definite covariance matrix. Outcomes for different

students are assumed to be conditionally independent given all of these parameters. The components of $\boldsymbol{\theta}'_{ST}$ are *i.i.d.* $N(0, \tau^2_{\theta,st})$. Finally, the teacher effects are assumed to be independent across subjects and years.

The McCaffrey et al. and Lockwood et al. exploit the availability of tests in multiple subjects to improve the precision of estimation of teacher effects on any specific subject. As Ballou et al. indicated out, this multivariate outcome approach not only reduces confounding of teacher assignment with student background, also increases the robustness of results to non-ignorable missing values. However, the complexity of the models poses computational challenges that render likelihood methods practically infeasible for all but small data sets. Lockwood et al. propose a Bayesian formulation of the variable PS model that scales well to the extremely large and complex data sets that challenge alternative approaches to parameter estimation.

Relationships among the Existing Principal VAM

McCaffrey et al. summarize five features of these models: parameterization of the overall time trend, inclusion of covariates, the distribution of residual error terms, the persistence of teacher effects on future outcomes, and translations between modeling scores and gains. According to these features, relationship among models is summarized as: the GS and CC model without covariates are special cases of the LA model with restrictions to the overall time trend and/or the distribution of residual errors. The LA model is a special case of the PS model with restrictions on the α s and without covariates. The CA and GS model with covariates are special cases of the PS model with restrictions on the distribution of residual errors and the α s.

Findings on the Major Issues from the Recent Studies

Modeling the School or Teacher Effects as Fixed or Random

VAM can specify school or teacher effects as either fixed or random effects. If the effects are treated as fixed, then the observed schools or teachers are assumed to be the only units of interest. Random effects assume that the units is a sample from a larger population. In VAM application, to model school or teacher effects as fixed or random is the primary design choice. Tekwe et al. (2004) addressed this issue by comparing models with different specifications using simulated data. The first model is fixed-effects models (FEM), where school effects (i.e., the improvement in student achievement due to teacher or school efforts) are taken to be fixed rather than random. This is the simplest of all models, requiring little computational complexity and not much mathematical knowledge. This model thus has intuitive appeal to policymakers, since the interpretation of the results is much easier to comprehend. An extension of this model, the simple fixed effects model, or SFEM, is an intuitively simple model that incorporates no student background factors, does not consider the complex linkage between students, teachers and schools, and by the nature of the statistics used, does not produce shrunken estimates. This model estimates school effects by comparing school effects only to the effect sizes of the districts to which they belong. Another type of VAM is hierarchical linear models (HLM), which assume that school effects are random. These models produce shrunken effects towards the mean, and there are two of them. First is the simple unadjusted change score HLM (UHLM) with random intercept. This model does not account for compositional or student-level covariates. Second is the demographic and intake score adjusted HLM (AHLM), where outcome is defined by a change score, and contains student and school-level covariates.

The results of the simulation study showed very strong correlations (higher than 0.9) between results provided by SFEM, LMEM, and UHLM, but much more modest correlation between the results of AHLM and all other models. Tekwe et al. concluded on the basis of these results that the SFEM performed about as well as the other two models that did not incorporate compositional or student-level covariates, and could be expected to produce similar results at a much lower computational cost. It was noted that these results were based on only two years of student achievement data and that the incorporation of more years of data might affect the relationships among effects generated by the three models. The difference between AHLM and all other models was notable, and indicate that when compositional and student-level covariates are included in the analysis, the estimates change. Although AHLM takes into more factors that do indeed affect student learning, it is arguable that the AHLM produces more precise estimates than do other models. Tekwe et al. finally noted that considerations that may have importance in theory may make little difference in a practical sense, and if a simpler model produces results comparable to more complex models, it may be preferable because of its intuitive appeal.

Ballou et al. criticized Tekwe et al.' s finding by stating that Tekwe et al.' s interest is confined to estimating school effects with large samples of students and data with two time points. As the sample size becomes large, the fixed effects and random effects estimates converge. When teacher effects are of interest, large sample size is not realistic. And fixed effects models are suboptimal when multiple time points and multiple cohorts are available. Therefore, Ballou et al. recommend random effects as a general approach, although fixed effect estimates have good properties in some circumstances.

McCaffrey and colleagues (RAND; 2003) were also interested in discussing the advantages and disadvantages of specifying school or teacher effects as fixed or random. They stated that one advantage of estimating teacher effects with a random effects model is that shrinking reduces the variance of an estimate of an individual teacher effect compared to the fixed effect estimate. The downside of a random-effects model is that the shrinking effect forces the estimated teacher effects to deviate from the true effects if the teacher's class is small. Although the fixed effects do not shrink estimates toward the mean, they will not necessarily move teachers toward the middle of the distribution if the class is small. Thus, the fixed effect estimates for teachers with small classes will be more likely to be in the extremes of the distribution. In sum, specifying school or teacher effects as fixed or random provide similar conclusions about the variability of teachers but yield different estimates of individual teacher effects.

Inclusion of the Covariates

McCaffrey et al. (RAND; 2003) stated that the importance of modeling background variables depends in a relatively complicated fashion on the interaction of several factors. These factors are the distribution across classes and schools of students with different characteristics, the relationship between the characteristics and outcomes, the relationship between the characteristics and true teacher effects, and the type of model used. Therefore, the importance of modeling student background characteristics when using VAM to estimate teacher effects remains an empirical question that must be addressed by each analyst in the context of these specific factors (See McCaffrey et al. RAND 2003 for details). In this section, several empirical studies focusing on the impact of inclusion of covariates will be reviewed.

Ballou et al. (2004) evaluated the TVAAS model and noted that studies of the

inclusion of contextual factors in HLM models almost always show that the results are sensitive to such effects They also noted that the TVAAS can include context factors if desired. Inclusion of these factors tends to bias measures of school and teacher effects towards zero.

Using data from the vast database accumulated by TVAAS, Ballou et al. conducted a simulation study to determine how much teacher effect sizes reported by the TVAAS would change if student and school compositional effects were entered into the model. The simulation study used student eligibility for free and reduced price lunch, race other than white, gender, the two-way interactions between these, and percent free and reduced price lunch by classroom as covariates. Thus, there were three student-level covariates and one school composition variable used in the study. The conclusion of this study was that student-level covariates showed only a moderate influence on teacher effectiveness scores. The scores produced by the two models were 2.7 times more likely to agree than to disagree in reading, 3.5 times more likely in language arts, and 8.5 times more likely in mathematics.

With respect to the school composition variable (free and reduced lunch), Ballou et al. found that there was a significant effect on the magnitude of teacher effectiveness scores, but they noted that the direction and magnitude of the regression coefficients showed that the relationship between the percent free and reduced variable and teacher effectiveness was unstable, and therefore not much confidence could be placed in the results. Having concluded that student-level covariates had little effect on teacher performance cores, Ballou et al. offered four possible explanations for the result of the simulation study. First, if the great majority of teachers have roughly the same mix of poor and non-poor students, white and non-white, then adjusting for demographics will

not change estimated teacher effects. However, the reality is that the mix of poor and nonwhite children does indeed vary widely from one classroom to another. Second, the impact of student variables is not large enough to make an appreciable difference to estimated teacher effects. But if we compare the result from the TVAAS with that from the fixed effect model, this explanation might be doubted. Because the results from the models are significantly different. Third, the high correlation between adjusted and unadjusted effects is caused by shrinkage. Finally, student factors add little information beyond that contained in the covariance of test scores. That is, other test scores contain much of the same information.

McCaffrey et al. also systematically investigated the influences of covariates on the GS, CA, CC and LA models. As presented in the previous section, the GS and CA model could include student and compositional variables, although the models produce biased estimates when the covariate and residual error terms are correlated. The CC and LA model usually include no or only limited information on student characteristics because some analysts have suggested that the inclusion of intra-student correlation essentially removes the effects of omitted covariates. However, McCaffrey et al. found that the impact of omitted covariates on estimated teacher effects depends on both the distribution of the omitted covariates and the assignment of students to teachers. McCaffrey et al. noted that omitted variables that are randomly distributed should have little effect on the results of any of the models. But when omitted variables cluster by class, or when they differ by stratum, none of these models is capable of disentangling teacher effects from the effects of student-level covariates. The CC and LA model are most sensitive to the effect of omitted variables.

Teacher Effects Are Cumulative and Long Lasting

The papers reviewed in this section focusing on investigating the persistence of those teacher effects on students' future achievement. Sanders and Rivers (1996) used data from two school systems in Tennessee to study the cumulative effects of third, fourth, and fifth grade teachers on fifth grade math achievement. Rivers (1999) replicated this study using slightly different methods to measure the cumulative effects of fifth, sixth, seventh, and eighth grade teachers on ninth grade achievement. Mendro et. al (1998) replicated the Sanders and Rivers's study using data from Dallas public schools, and Kain (1998) provided a separate independent reanalysis of the Dallas data.

Sanders and Rivers (1996) purported to show that teacher effects accumulate and persist over time. They reported that for math tests, students taught by the least effective teachers for three consecutive years would score 52 to 54 percentile points below similar students taught by the most effective teachers for three consecutive years. In the paper, Sanders and Rivers use a two-stage approach. First, they estimate teacher effectiveness using the CA model. Separate models are fit to math scores for the 3rd, 4th, 5th grade students. The correlation among the residual errors from the same student is ignored. These models provide shrinkage estimates of the teacher effect θs for each grade. Then, teachers within each grade are ranked and assigned to scale 1 to 5 based on the estimated teacher effect quintiles. In the second stage, student scores from grade 5 are modeled as an additive linear function of teacher effectiveness (where the quintile assignments are treated as categorical variables) for grades 3 through 5, the second grade score, and residual error. This model is the ANCOVA model

$$Y_{i5} = \mu_5 + q_{3i} + q_{4i} + q_{5i} + \beta Y_{i2} + \varepsilon_{i5} \qquad (6)$$

Where q_{gi} is a five-level categorical variable representing the quintile of the grade g teacher for student i. The estimated differences between outcome Y s are compared to indicate the teacher effectiveness. However, the authors' ad hoc method has been criticized for using the same students in both stages of the analysis (Kupermintz, 2002).

Rivers (1999) replicated the Sanders and Rivers' design with several important modifications to address some of the criticisms of SR and still found persistent teacher effects. Rivers used the teacher effect estimates from the LA model rather than using Sander and Rivers' simple CA model. As discussed in the previous section, the LA model can simultaneously model scores from several subjects and several years. The model allows for correlation among scores from the same student, although it includes no student or school level covariates. Another feature one should note is that the LA model used by Rivers includes a separate parameter for the mean of every school system or district; so estimated teacher effects are relative to the other teachers in the district. The second difference between Rivers and Sanders and Rivers is that Rivers used two cohorts of students rather than one to estimate the persistence of teacher effects. The first cohort provided estimates of teacher effectiveness from the LA model. The second cohort was used to conduct Sanders and Rivers' second stage analysis. Rivers modeled ninth grade test scores as a function of fourth grade test scores and the students' fourth to eighth grade teachers' stage 1 effectiveness ratings based on the prior cohort. Thus, estimates of teacher effectiveness and the impact of varying effectiveness were estimated from two distinct cohorts of students. The final major difference between Rivers and Sanders and Rivers is that Rivers models outcomes on a different test than the test used for estimating effectiveness, and the outcome is measured at the end of ninth grade while teacher effects are measured for sixth, seventh, and eighth grades.

Rivers found that teacher effects from all four grades are statistically significantly related to scores in the fall of ninth grade. The effect of fifth and sixth grade teachers decreases when the students' fourth grade scores increase. That is, fifth and sixth grade teachers were estimated to matter more for students with lower baseline scores. The impact of fifth grade teachers on ninth grade tests is about two times greater for students at the mean of the lowest quartile of fourth grade scores than the impact for students at the mean of the highest quartile. The impact of sixth grade teachers is about 2.5 times greater for students in the lowest quartile compared with the highest quartile on the fourth grade test. Rivers also found that for students scoring low at fourth grade, fifth and sixth grade teachers have the strongest relationship with ninth grade scores, while for other students, eighth grade teacher effects have the strongest relationship with ninth grade scores. However, River's study also has its limitations. It excluded the students who transfer across schools or retain in grade. Thus, Rivers' result that teacher effect persists into the future tests can only apply to students who remain in the same school systems for six years.

The Sanders and Rivers study was also replicated by Mendro, Jordan, Gomez, Anderson, and Bembry (1998) using data from students in the Dallas Independent School District. Mendro et al. consistently found large persistent teacher effects across multiple cohorts and on both reading and math scores. They corroborated the results of Sanders and Rivers and Rivers, even with a very different approach. Mendro fit models analogous to those of Sanders and Rivers. The models include previous year's scores as a covariate, and teacher ratings as a categorical variable. Teacher ratings are from the estimated teacher effects using the Dallas Value Added Accountability System (DVAAS). DVAAS uses a three-stage approach to estimate teacher effects. In stage 1,

the fairness variables (e.g., gender and ethnicity) are removed from the from current-year and past-year scores. Stage 2 of the DVAAS estimation procedure models the first stage residual for the current-year score as a function of first-stage residuals for prior-year scores, prior-year attendance, and school-level variables. Stage 3 estimates teacher effects as the classroom averages of the stage 2 residuals. The procedure produces separate estimates for teacher effects on the math and reading scores. Details on the Dallas teacher effects are presented in Webster and Mendro (1997). The authors also found that students' loss due to an ineffective teacher in one year cannot be compensated by the additional years of schooling. Teacher in one year do not make up for this loss even after additional years of schooling. They demonstrated this effect by showing outcomes for pairs of groups of students who have similar average outcomes on the pretest, but one group in each pair had ineffective teachers in the first year. The results showed that the group with an ineffective teacher in the first year always scores lower on the final test, regardless of the effectiveness of the teachers in the ensuing year. In sum, the empirical studies conducted by Sanders and Rivers, Rivers, Mendro and other researchers provided consist findings that the persistence of the teacher effect exists, although the size of the effects vary across studies.

CHAPTER 3

GENERAL VAM

This chapter will first propose the multivariate general VAM in a matrix formulation. The parameter estimation of the general VAM will be implemented under the Bayesian framework. The second section will specify the Bayesian method and MCMC procedure conducted to estimate all the model parameters. The third section will show how each reduced model will be derived from the general model.

The Matrix Formulation of the General VAM

The matrix formation of the proposed general VAM is

$$\mathbf{y}_i = \boldsymbol{\Psi} \bullet \boldsymbol{\Pi}_{i1}\boldsymbol{\Lambda} + \boldsymbol{\Phi} \bullet \boldsymbol{\pi}_i \bullet \boldsymbol{\Pi}_{i2}\boldsymbol{\Theta} + \gamma_i \bullet \boldsymbol{\Pi}_{i3}\varepsilon_i + \boldsymbol{\Delta} \qquad (7)$$

where the operation $\bullet$ is entrywise production. Such, given two vectors $\vec{\mathbf{a}} = (a_1, \cdots, a_n)$, and $\vec{\mathbf{b}} = (b_1, \cdots, b_n)$,

$$\vec{\mathbf{a}} \bullet \vec{\mathbf{b}} = (a_1 b_1, \cdots, a_n b_n) \qquad (8)$$

Each variable in this formation is defined as follows.

- $\mathbf{y}_i$ contains *i*th student's scores. It is a vector of length ST, the number of subjects (S) times the number of years (T). The elements in vector Y_i are first arranged by subjects, then by time.

- $\boldsymbol{\Lambda}$ contains all fixed effects, which need to be estimated. The fixed effect parameters $\boldsymbol{\Lambda}_{ST\times F}$ can be decomposed into two parts, one part $ST \times \boldsymbol{\Lambda}_{F_1}$ is for time variant factors (e.g., mean score), and the other part $\boldsymbol{\Lambda}_{ST\times F_2}$ is for time invariant factors

(e.g., gender, ethnicity), and $F = F_1 + F_2$. Suppose mean score is used as time variant fixed effect, and gender is used as time invariant fixed effect, we have

$$\mathbf{\Lambda} = (m_{11}, m_{12}, \cdots, m_{1T}, G_1, \cdots, m_{S1}, m_{S2}, \cdots, m_{ST}, G_S)' \qquad (9)$$

- $\mathbf{\Psi}$ is a $ST \times F$ matrix that contains the coefficients that represent the relationship between the current and previous years' scores. They need to be estimated according to $\mathbf{\Lambda}$.

- $\mathbf{\Pi}_{i1}$ is an incidence matrix designed according to $\mathbf{\Lambda}$ for the *i*th student.

- $\mathbf{\Theta}$ is a random effect matrix contains ST vectors of length R. $R = N_1 T + N_2 T$, where N_1 is the number of schools; N_2 is the number of teachers. The first N_1 elements in each vector represent N_1 school effects, which are followed by N_2 teacher effects. The ST vectors are organized first by subjects, then by time. If the model only involves the N_1 school effects in three contiguous years for two subjects, the notation will be

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\theta}_{11}^{1} \\ \boldsymbol{\theta}_{11}^{2} \\ \vdots \\ \boldsymbol{\theta}_{11}^{N_1} \\ \vdots \\ \boldsymbol{\theta}_{13}^{1} \\ \boldsymbol{\theta}_{13}^{2} \\ \vdots \\ \boldsymbol{\theta}_{13}^{N_1} \\ \boldsymbol{\theta}_{21}^{1} \\ \boldsymbol{\theta}_{21}^{2} \\ \vdots \\ \boldsymbol{\theta}_{21}^{N_1} \\ \vdots \\ \boldsymbol{\theta}_{23}^{1} \\ \boldsymbol{\theta}_{23}^{2} \\ \vdots \\ \boldsymbol{\theta}_{23}^{N_1} \end{pmatrix}$$

where $\boldsymbol{\Theta}_{st}$ is the school effect vector of the *t*th year for the *s*th subject, which contains N_1 schools effects. $\boldsymbol{\Theta}_{st} \sim N(\boldsymbol{\mu}_{st}, \tau_{st}^2 \mathbf{I}_{N_1})$. In general, we assume $\boldsymbol{\mu}_{st} = \mathbf{0}$ and the vector that contains all τ_{st}'s needs to be estimated. The similar distribution is assumed for the teacher effect.

- $\boldsymbol{\Pi}_{i2}$ is an incidence matrix of dimension $ST \times R$ that indicates the linkage between student *i* and schools/teachers.

- $\boldsymbol{\pi}_i$ is also an incidence matrix of dimension $ST \times R$ that is previously assigned according to $\boldsymbol{\Pi}_{i2}$ for student *i*. It presents how long the student has been associated with a specific teacher or school.

• $\mathbf{\Phi}$ is a $ST \times ST$ block diagonal matrix contains the persistence parameters measuring how much the previous school or teacher effect contribute to the current year score. $\mathbf{\Phi}$ consists of S distinct ($T \times T$) lower triangular blocks corresponding to subjects. The (t, t^*) element of the block for subject s is ϕ_{s,tt^*}, for $t^* \leq t$ and 0 otherwise, where ϕ_{s,tt^*} denotes the school or teacher effect persistence parameters for subject s.

• $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2})$ presents random effects for the ith student. For example, we can assume $\varepsilon_{i1} \sim N(m_1, v_1)$, $\varepsilon_{i2} \sim N(m_2, v_2)$, and they present mean and slope of student growth curve, respectively. ε_i's are independent across students.

• $\mathbf{\Pi}_{i3}$ is also an incidence matrix.

• γ_i contains the coefficients that need to be estimated according to c.

• $\mathbf{\Delta}$ is the random error, which follows multivariate norm distribution $MVN(0, \Sigma)$ and independent of teacher or school effects. $\mathbf{\Sigma}$ is a $ST \times ST$ unstructured positive definite covariance matrix. If the residuals across years and subjects are independent, only diagonal elements need to be estimated. Normally, residuals across years and subjects are not independent. Student-specific effects on scores and the relationship between the scores across years and subject within one student can be reflected in the covariance matrix.

Bayesian Method for the General Value-Added Model

The conditional distribution for student i's score vector is

$$\mathbf{y}_i \mid \mathbf{\Lambda}, \mathbf{\Theta}, \mathbf{\Psi}, \boldsymbol{\tau}, \boldsymbol{\gamma}_i, \boldsymbol{\varepsilon}_i, m_1, m_2, v_1, v_2, \mathbf{\Sigma}) \sim N(\mathbf{\Psi} \bullet \mathbf{\Pi}_{i1}\mathbf{\Lambda} + \mathbf{\Phi} \bullet \pi_i \bullet \mathbf{\Pi}_{i2}\mathbf{\Theta} + \gamma_i \bullet \mathbf{\Pi}_{i3}\boldsymbol{\varepsilon}_i, \mathbf{\Sigma}) \quad (10)$$

where the likelihood function is given by the production of equation (10) across all examinees. The prior distributions used in this study are

$$\mathbf{\Lambda} \sim N_F(\boldsymbol{\mu}_\Lambda, \mathbf{V}_\Lambda) \tag{11}$$

$$\boldsymbol{\Theta}_{st} \sim N_{N_1}(0, \tau_{st}^2 \mathbf{I}_{N_1}) \tag{12}$$

$$\tau_{st} \sim U(a,b) \tag{13}$$

$$\boldsymbol{\varphi} \sim N_{\frac{ST(T-1)}{2}}(\boldsymbol{\mu}_{\boldsymbol{\varphi}}, \mathbf{V}_{\boldsymbol{\varphi}}) \tag{14}$$

$$\mathbf{\Sigma}^{-1} \sim W(\mathbf{d}_\Sigma, \mathbf{D}_\Sigma) \tag{15}$$

Here W denotes the Wishart distribution, and $U(a,b)$ denotes the uniform distribution on the interval (a, b).

The joint posterior distribution for all the parameters given the fixed hyperparameters is

$$P(\mathbf{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\tau}, \boldsymbol{\gamma}_i, \boldsymbol{\varepsilon}_i, m_1, m_2, v_1, v_2, \mathbf{\Sigma} \mid \mathbf{y}) \propto P(m_1)P(m_2)P(v_1)P(v_2)P(\mathbf{\Sigma})$$

$$P(\mathbf{\Lambda})P(\boldsymbol{\Theta} \mid \boldsymbol{\tau})P(\boldsymbol{\Psi})P(\boldsymbol{\tau})P(\boldsymbol{\gamma}_i)P(\boldsymbol{\varepsilon}_i \mid m_1, m_2, v_1, v_2)$$

$$L(\mathbf{y} \mid \mathbf{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\tau}, \boldsymbol{\gamma}_i, \boldsymbol{\varepsilon}_i, m_1, m_2, v_1, v_2, \mathbf{\Sigma}) \tag{16}$$

MCMC simulation can be used to draw samples iteratively from the full conditional distributions of the parameters given the data and the rest of the parameters. Below is an outline of how parameters can be sampled from their full conditional distributions.

1. Updating $\mathbf{\Sigma}$: First we obtain the residual $\mathbf{e}_i$ for each student. Then the full conditional distribution for $\mathbf{\Sigma}$ is Wishart distribution with degree of freedom $\mathbf{d}_\Xi + N$ and parameter $\mathbf{D}_\Xi + \sum_{i=1}^{N} e_i e_i'$.

2. Updating τ_{st} : the full conditional distribution of τ_{st} depends only on the current value of $\mathbf{\Theta}_{st}$, the length of the vector N_1 and the hyperparameters *v* and δ of the uniform distribution. A new value for the parameter will be sampled using the Metropolis-Hastings within Gibbs algorithm (Casella & George, 1995) because the full conditional distribution is not available in closed form. We update the transformed value

$$\eta_{st} = f(\tau_{st}) = \log\left(\frac{\tau_{st} - v}{\delta - \tau_{st}}\right) \tag{17}$$

So

$$\tau_{st} = f^{-1}(\eta_{st}) = (\delta e^{\eta_{st}} + v)/(1 + e^{\eta_{st}}) \tag{18}$$

and the prior distribution for η_{st} is

$$P(\eta_{st} \mid v, \delta) = e^{\eta_{st}}/(1 + e^{\eta_{st}}) \tag{19}$$

The Metropolis-Hastings algorithm is implemented as follows

i) Draw initial value $\eta_{st}^{(0)}$ from prior distribution.

ii) At iteration *m*, draw candidate $\eta_{st}^{(*)}$ from the proposal normal distribution with mean $\eta_{st}^{(m-1)}$ and known variance.

Accept each $\eta_{st}^{(m)} = \eta_{st}^{(*)}$ with probability

$$P(\eta_{st}^{(m-1)}, \eta_{st}^{(*)}) = \min\left(1, \frac{P_{N-1}(\theta_{st}^{(m)} \mid 0, (f^{-1}(\eta_{st}^{(*)})^2 \times \mathbf{I}_{N1}) P(\eta_{st}^{(*)}, \mid v, \delta)}{P_{N-1}(\theta_{st}^{(m)} \mid 0, (f^{-1}(\eta_{st}^{(m-1)}))^2 \times \mathbf{I}_{N1}) P(\eta_{st}^{(m-1)}, \mid v, \delta))}\right) \tag{20}$$

Otherwise $\eta_{st}^{(m)} = \eta_{st}^{(m-1)}$.

3. Updating $\mathbf{\Lambda}$: First we obtain the partial residual $\mathbf{e}_i^*$, which is the difference between the observed score and all the random effects part for each student. Then regress the vector $\mathbf{e}^*$ of these residuals on $\mathbf{\Psi} \times \mathbf{\Pi}_{i1}$ for all students, with known error covariance $\mathbf{V} = (\mathbf{I}_N \otimes \mathbf{\Sigma})$. Therefore, we draw Λ from the full conditional distribution for $MVN(\mathbf{Bb}, \mathbf{B})$, where

$$\mathbf{B}^{-1} = (\mathbf{\Psi} \times \mathbf{\Pi}_{i1})^{\mathbf{T}} \mathbf{V}^{-1} (\mathbf{\Psi} \times \mathbf{\Pi}_{i1}) + \mathbf{V}^{-1} \tag{21}$$

and

$$\mathbf{b} = (\mathbf{\Psi} \times \mathbf{\Pi}_{i1})^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{E}^* + \mathbf{V}_{\mathbf{\Lambda\mu\Lambda}} \tag{22}$$

4. Updating $\mathbf{\Theta}$: We update $\mathbf{\Theta}$ one element at a time. We obtain partial residual $\mathbf{e}_i^*$ for each student linked to the teacher of interest by subtracting the fixed effects structure and the part of the teacher structure that does not depend on the teacher effect being updated. Then regress these $\mathbf{e}_i^*$ on the single teacher effect $\mathbf{\theta}_{st}$, where the design matrix consisting of zeros, ones, and the appropriate components of $\mathbf{\Phi}$ (here we use $\mathbf{D}_\Phi$ to denote the current design matrix), and where the error covariance matrix is *V*. The full conditional distribution for $\mathbf{\theta}_{st}$ is $N(\mathbf{B}_\Theta \mathbf{b}_\Theta, \mathbf{B}_\Theta)$, where

$$\mathbf{B}_\Theta^{-1} = \mathbf{D}_\Phi^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{D}_\Phi + \mathbf{V}^{-1} \tag{23}$$

$$\mathbf{b}_\Theta = \mathbf{D}_\Phi^{\mathbf{T}} \mathbf{V}^{-1} \mathbf{E}^* \tag{24}$$

5. Updating $\mathbf{\Phi}$: Analogous to the previous step, here the $\mathbf{\theta}_{st}$ s serve as regressors with parameter $\mathbf{\Phi}$. We obtain partial residual $\mathbf{e}_i^*$ for each student by subtracting the fixed effects structure and the effects of all current year teachers for each score. Now the

design matrix consists of zeros and appropriately placed values of $\boldsymbol{\theta}_{st}$. The error covariance matrix is also $\boldsymbol{V}$. So the full conditional distribution for α is $MVN(\mathbf{B}_{\Theta}\mathbf{b}_{\Theta}, \mathbf{B}_{\Theta})$. Where

$$\mathbf{B}_{\Phi}^{-1} = \mathbf{D}_{\theta}^{T}\mathbf{V}^{-1}\mathbf{D}_{\theta} + \mathbf{V}^{-1} \tag{25}$$

$$\mathbf{b}_{\Phi} = \mathbf{D}_{\theta}^{T}\mathbf{V}^{-1}\mathbf{E}^{*} + \mathbf{V}_{\Phi}\boldsymbol{\mu}_{\Phi} \tag{26}$$

6. Updating the student's own random effects is analogous to the steps that update the school/teacher random effects.

Existing Major Value-Added Models

Now, suppose, 1) we only have scores for one subject S = 1; 2) there is only one teacher in each year N_t = 1, then the teacher effect $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_T)^T$; 3) and the mean score for T years and gender are the fixed effects, then we will derive the CA, GS, PS and LA and CC model.

CA Model

$$\boldsymbol{\Lambda} = (m_1, m_1, \cdots, m_T)^T \tag{27}$$

$$\boldsymbol{\Pi}_{i1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}_{T\times T} \tag{28}$$

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \psi & 1 & 0 & \cdots & 0 \\ \psi^2 & \psi & 1 & \cdots & 0 \\ \vdots & & & & \\ \psi^T & \psi^{T-1} & 1 & \cdots & 1 \end{pmatrix}_{T\times T} \tag{29}$$

$$\boldsymbol{\Pi}_{i2} = \boldsymbol{\Pi}_{i1} \tag{30}$$

$$\mathbf{\Phi} = \mathbf{\Psi} \tag{31}$$

$\mathbf{\Pi}_{i3}$ = 0, and $\mathbf{\Sigma}$ is $T \times T$ matrix that only diagonally element to be estimated, correlations are all 0. The covariant adjustment model for the t years can be derived as:

$$y_1 = m_1 + \theta_1 + e_1 \tag{32}$$

$$y_t = m_t + \psi y_{t-1} + \theta_t + e_t \tag{33}$$

GS Model

The GS model can be obtained if *b* in the CA models is specified to be 1. Therefore, the the GS model can be viewed as a special case of the CA model.

PS Model

$$\mathbf{\Lambda} = (m_1, m_1, \cdots, m_T, G)^T \tag{34}$$

$$\mathbf{\Pi}_{i1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 1 \\ \vdots & & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{T \times T+1} \tag{35}$$

$$\mathbf{\Psi} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & \psi_1 \\ 1 & 1 & 0 & \cdots & 0 & \psi_2 \\ \vdots & & & & & \\ 1 & 1 & 1 & 1 & 1 & \psi_T \end{pmatrix}_{T \times T+1} \tag{36}$$

Π_{i2} is the same as Equation (35). Suppose

$$\mathbf{\Phi} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \phi_{21} & 1 & \cdots & 0 \\ \vdots & & & \\ \phi_{T1} & \phi_{T2} & \cdots & 1 \end{pmatrix}_{T \times T} \tag{37}$$

LA Model

For the PS model above, if we further assume all the $\phi_{ij} = 1$, for $i > j$ in Equation (37), then we will obtain LA model. Therefore, the LA model can be seen as a special type of the PS model.

CC Model

The CC model is the only model that explicitly models individual growth curves. In this model, student's growth are student-specific, and of random effects. The teacher's effects are the same as the LA model. For student-specific random effects, we have

$$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}) \tag{38}$$

$$\boldsymbol{\Pi}_{i3} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \\ 1 & 1 \end{pmatrix}_{T\times 2} \tag{39}$$

$$\boldsymbol{\gamma}_i = \begin{pmatrix} 1 & r_1 \\ 1 & r_2 \\ \vdots & \\ 1 & r_T \end{pmatrix}_{T\times 2} \tag{40}$$

In Raudenbush and Bryk, they used fixed coefficient for γ_i , and $r_t = t, t = 1, \cdots, T$.

Table 3.1 summarizes the similarity and differences among the general and reduced VAM.

Table 3.1. Comparison among the General and Reduced VAM

	General VAM	Gain Score	Covariate Adjustment	Cross-Classified	Layered	Persistence
Covariates	Yes	Yes	Yes	No	No	Yes
Student random growth	Yes	No	No	Yes (linear trend)	No	No
Prior year teacher effect	Yes	Yes	No	Yes	Yes	Yes
Teacher effect persistence	Diminished	Undiminished	No	No	Undiminished	Diminished
Residual error	No restriction	Independent across years within a student	Independent across years within a student	Constant correlation within student across years	Allows correlation across years within student; Constant across all students	No restriction

CHAPTER 4

SIMULATION STUDY

Design

A simulation study is conducted to investigate the feasibility and robustness of the general model when the data are generated under varying assumptions. Six data sets are generated using the general model and the five reduced models (the GS, CA, CC, LA and PS model). Data (test scores) generated using the general model are fitted using the general model and all the reduced models. This allows comparison of the different fitted models when the data do not follow the assumptions of the reduced models. Data generated using a reduced model are fitted using both the same reduced model and the general model. This allows comparison of the model fit of the general model and specific reduced model when the reduced model assumptions hold. Table 4.1 summarizes this design and shows that 16 conditions result from the combination of the data generation and estimation methods.

For each model, three consecutive years of testing scores are generated for 1200 students grouped into 48 classes of 25 students each. The classes are grouped into 3 schools each with 16 classes. Systematic heterogeneity is introduced into the school means through the students. School A contains 80% students who are eligible for free and reduced lunch (FRL), school B contains 50% students who are eligible for FRL, and school C contains 20% FRL students. School C is considered a balanced mix. Within each school, the students are randomly assigned to two classes. The students without FRL are assumed to be advantaged students and those with FRL are disadvantaged students. The advantaged students have higher mean scores at the starting year and higher mean gain scores each year than the disadvantaged students.

Within each school, teachers with different effectiveness are assigned into 16 classes across three years. In this simulation study, the teachers who contribute positively to students' growth will be considered as the effective teachers and the teachers who contribute negatively are considered as the non-effective teachers. Various combinations of effective and non-effective teachers across three years yield four types of classes: Class NNN, Class NNE, Class EEN and Class EEE. For example, for Class NNE, non-effective teachers are assigned for the first two years and an effective teacher is assigned for the third year. For each type of combination, there are 4 classes. Table 4.2 summarizes the teacher arrangement for each class across three years.

Models for Generating Scores

According to formula (10), given the values of all the required parameters, student i's score vector can be assumed to have a multivariate normal distribution. The covariate variable considered in the simulation study is the students' SES; the random effect considered are teacher effect for all the models and students' own random effect for the general and CC model. Although schools are heterogeneous, the school effect is not examined here. The heterogeneity among schools is introduced only for the purpose of investigating the relationship between school composition and the inclusion of the covariates. The parameter values used to generate student scores are listed as follows.

1) The first year's mean scores are 220 and 200 for the advantaged and disadvantaged students, respectively; the average gain score for each year is 20 and 10 for the advantaged and disadvantaged students, respectively.

2) The marginal variance (the total variance of teacher effect, student random effect and residual error) is fixed at 1000 for all the conditions.

3) The teacher effect follows $N(0,\tau^2)$. The value of τ is fixed at 10. The difference between the average teacher effects for the effective and non-effective teacher groups is one unit of the standard deviation. Then the teacher effects for the effective teachers are generated from $N(5,10^2)$ and the teacher effects for the non-effective teachers are generated from $N(-5,10^2)$

4) For the general and CC model, the random student effect follows $N(0,v^2)$. The value of v is chosen to be 5.

5) The teacher effect persistence parameters ϕ_{21}, ϕ_{31} and ϕ_{32} are 0.2, 0.3 and 0.3 for the general and PS model; they are 1 for the GS and LA model.

6) For student i, the variance-covariance for the residual error is a 3 by 3 matrix. A random matrix is created for the general and the PS model. The correlation of scores across years within student i is 0 for the GS and CA model, and is 0.7 for the CC and LA model.

Table 4.1 Models Used for Generating and Fitting Simulation Data

		Data Fitted					
		General	GS	CA	CC	LA	PS
Data Generated	General	√	√	√	√	√	√
	GS	√	√				
	CA	√		√			
	CC	√			√		
	LA	√				√	
	PS	√					√

Table 4.2 Teacher Arrangement for Each Class

School	% of advantaged students	Class	Year 1	Year 2	Year 3
A	80%	NNN	N	N	N
		NNE	N	N	E
		EEN	E	E	N
		EEE	E	E	E
B	50%	NNN	N	N	N
		NNE	N	N	E
		EEN	E	E	N
		EEE	E	E	E
C	20%	NNN	N	N	N
		NNE	N	N	E
		EEN	E	E	N
		EEE	E	E	E

Analysis and Comparison of Model Estimation

For each model that will be analyzed, the MCMC algorithm shown in Chapter 3 is implemented to generate a sequence or chain of parameters sampled from the posterior distribution of that model. Constraints are put on specific parameters when the MCMC algorithm is implemented for estimating the reduced model. For example, the persistence parameter ϕ_{ij} is fixed to be 1 at each step when the data are estimated by the LA model. The convergence of the chains is diagnosed using the Gelman-Rubin diagnostic (Gelman & Rubin, 1992).

Model comparison is required for a diversity of activities, including variable selection in regression, determination of the number of components in a mixture model or the choice of parametric family. As with frequentist analogues, Bayesian model comparison will not inform about which model is "true", but rather about the preference for a model given the data and other information. In the Bayesian arena, common methods for model comparison are based on the following: separate estimation including posterior predictive distributions, Bayes factors and approximations such as the Bayesian information criterion (BIC) and deviance information criterion (DIC); comparative estimation including distance measures such as entropy distance or Kullback-Leibler divergence; and simultaneous estimation, including reversible jump MCMC and birth and death processes (Alston, Kuhnert, Low Choy, McVinish & Mengersen, 2005).

Researchers have shown that, as an approximation to the Bayesian factor, the DIC is a popular method for model comparison, especially for models that involve many random effects, large numbers of unknowns or improper priors. DIC penalizes against higher dimensional models (Spiegelhalter et al, 1999): with deviance denoted by D,

$$DIC = E[D(\theta)|y] + E[D(\theta|y)] - D(E[\theta|y]) = D^{*}(\theta) + p_D \qquad (41)$$

where

$$D^{*}(\theta) = E_{\theta}[-2\log p(y|\theta)] + 2\log p(y) \qquad (42)$$

and p_D denotes the effective number of parameters. It can thus be seen that the DIC comprises of terms that are a function of the data alone (e.g., $2\log p(y)$) and a measure of the complexity of the model (e.g., $E_{\theta}[-2\log p(y|\theta)]$). In this study, DIC is used for the model comparison in terms of the overall goodness of fit. Within the same simulated data, the overall goodness of fit is compared among the models being used. Each year's estimated mean scores are compared to true mean scores. The bias of the estimated mean score is examined under all the conditions. Covariate variable is included in the general, GS and CA model. Specifically, a binary variable is used to indicate the student's SES (advantaged or disadvantaged). The impact of inclusion of the covariate variable is discussed.

Several measures are considered for the estimated teacher effects, for example, the estimates of individual teacher effects and the overall contributions of teacher to variability in student outcomes. Both measures are compared to their true values and the estimation accuracy is compared among different conditions, specifically, the impact of the inclusion of covariates on the estimation accuracy will be discussed. Teacher effect persistence parameters estimation accuracy is also checked.

Results

For each of the 16 conditions, four chains started at random are run. All the chains have the same number of burn-in, 5,000, but have different chain lengths. The chain

lengths are determined to ensure that all the parameters have converged. The resulting total number of iterations range from 15,000 to 25,000. The initial estimates for all the parameters are obtained based on the draws after the burn-in of each chain. The final estimates are obtained by averaging the estimates across the four chains. In addition, the posterior variance of the estimates is computed using the sample variance of the iterations after subtracting the burn-in.

Overall Model Fit

Tables 4.3-4.5 summarize the DIC value provided by all the models using different generated data for School A, B and C, respectively. The row label indicates the model by which the data were generated; the column label indicates the model by which the generated data were estimated. For each dataset, the estimating model is called “correct’ model when it corresponds to the generating model. Comparing these three tables, same pattern can be found, although the DIC values in the same cell across different tables are slightly different. Therefore, the following discussion is based on the results from by School A (Table 4.3).

The most salient result from Table 4.3 is that, for each data, the correct models consistently provide better model fit, which is indicated by the smaller DIC values. To be more specific, when the data were generated by the general model, the correct model provides smaller DIC than any other models. Among the other models, the PS model obtains the closest DIC to the general model DIC (Difference is 51.). This is because the PS model is most similar to the general model. DIC values provided by the CC and LA model are close to each other (12278 vs. 12264) and larger than those provided by other models. When the data were generated by the reduced model, the correct model provides

smaller DIC than the general model does. However, the difference of DIC values between the correct model and the general model is quite small. As the PS model differs from the general model only by one parameter (the student random effect), the DIC value from the general model is just higher by 1. Even the largest difference, which occurs in the GS model case, is only 26.

Table 4.3 DIC Obtained from All the Models Using Different Generated Data (School A)

		Fitted Model					
		General	GS	CA	CC	LA	PS
Generating Model	General	11475	12207	12223	11533	11539	11526
	GS	11516	11490				
	CA	11474		11465			
	CC	11688			11680		
	LA	11685				11682	
	PS	11481					11480

Table 4.4 DIC Obtained from All the Models Using Different Generated Data (School B)

		Fitted Model					
		General	GS	CA	CC	LA	PS
Generating Model	General	11471	12264	12278	11535	11541	11522
	GS	11525	11491				
	CA	11476		11463			
	CC	11759			11748		
	LA	11781				11770	
	PS	11488					11484

Table 4.5 DIC Obtained from All the Models Using Different Generated Data (School C)

		Fitted Model					
		General	GS	CA	CC	LA	PS
Generating Model	General	11472	12299	12313	11534	11540	11524
	GS	11517	11491				
	CA	11484		11465			
	CC	11705			11695		
	LA	11706				11696	
	PS	11481					11481

To summarize the findings, for the School A data, the comparison of the DIC values show that the general model provides the best model fit for the general-model-generated data; the PS model provides the closest result, which is just slightly worse. The CC and LA model perform much worse. For the reduced-model-generated data, compared to the general model, the reduced models provide better model fit. However, the performance of the general model is not much worse than that of the correct models. This conclusion can be generalized to School B and C which indicates that, in terms of the overall model fit, the school composition has no impact on the performance of the general model and the relationship between the general and reduced models. For example, for School B, which represents a mix balance of advantaged and disadvantaged students, the DIC values also support that the general model performs the best for the general-model-generated data and performs slightly worse than the correct models for the reduced-model-generated data.

Fixed Effect Estimation

Only one measure of the fixed effect estimation is presented here – the absolute bias of the estimated fixed effect. Choosing to present the absolute bias instead of the bias is just for the purpose of conveniently comparing the magnitude of the bias. For the CC and LA model, the estimated fixed effect only includes estimated mean score for each year. For all the other models, the estimated fixed effect also includes the estimated SES effect. For example, the estimated fixed effect for the general model is $\hat{m} + \hat{\varphi}\overline{X}$. $\overline{X}$, the average of the SES variable, is 0.2 for School A, 0.5 for School B and 0.2 for School C. It should be noted that previous year's mean score is also included for the GS and CA model. The absolute bias of the estimated fixed effect is the absolute value of the difference between the estimated fixed effect and the true mean score. Table 4.6 presents the true mean score for each year generated using different models. The purpose of this table is just to show a general picture of the generated scores used in this simulation study. It shows that, from year 1 to year 3, the mean score grows from around 215 to around 251 for School A, around 209 to around 240 for School B and around 203 to around 227 for School C. The higher percentage of advantage students leads to higher mean score and higher gains.

Tables 4.7-4.9 present the absolute bias of the estimated fixed effect for School A, B and C, respectively. All the absolute biases shown in these three tables are smaller than 0.50, which indicates that all the models can provide accurate fixed effect estimates for any of the data generated using different models. Moreover, to investigate the stability of the fixed effect estimates, the posterior standard deviations of the estimates were computed. The small posterior standard deviations (around 5 for all the estimates) show that all the models can provide precise fixed effect estimates.

Although the absolute biases vary across conditions, no significant differences can be found ($Z<1.96$). In other words, none of the differences between the estimated absolute biases is large enough to draw the conclusion that one model gives more accurate estimates than another. And the differences can only be accounted for by the random errors.

Table 4.6 The True Mean Scores Generated under Various Conditions for Three Years

		Year 1	Year 2	Year 3	Average
School A	General	215.4	231.7	251.3	232.8
	GS	215.5	232.0	251.5	233.0
	CA	215.5	231.9	251.4	233.0
	CC	217.3	233.7	253.3	234.8
	LA	217.4	233.9	253.4	234.9
	PS	215.6	231.8	251.4	232.9
School B	General	209.8	223.3	240.1	224.4
	GS	209.9	223.6	240.3	224.6
	CA	209.9	223.5	240.2	224.6
	CC	211.7	225.3	242.1	226.4
	LA	211.8	225.5	242.2	226.5
	PS	209.9	223.40	240.2	224.5
School C	General	203.4	213.7	227.3	214.8
	GS	203.5	214.0	227.5	215.0
	CA	203.5	213.9	227.4	215.0
	CC	205.3	215.7	229.3	216.8
	LA	205.4	215.9	229.4	216.9
	PS	203.6	213.8	227.4	214.9

Table 4.7 Absolute Bias of the Estimated Fixed Effect for Each Year from Different Models (School A)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	0.35	0.36	0.25	4.91	5.12	5.01
	GS	0.31	0.19	0.15	5.32	5.44	5.41
	CA	0.08	0.08	0.32	5.41	5.39	5.52
	CC	0.06	0.18	0.11	5.21	5.30	5.19
	LA	0.18	0.12	0.24	5.23	5.19	5.27
	PS	0.11	0.10	0.03	4.81	5.15	5.10
GS	General	0.23	0.34	0.25	5.43	5.29	5.33
	GS	0.23	0.44	0.33	5.19	5.24	5.25
CA	General	0.04	0.25	0.06	5.39	5.44	5.28
	CA	0.18	0.05	0.17	5.14	5.15	5.29
CC	General	0.07	0.24	0.21	5.31	5.29	5.24
	CC	0.18	0.34	0.37	5.09	5.12	4.87
LA	General	0.07	0.04	0.01	5.28	5.19	5.24
	LA	0.02	0.28	0.49	5.15	5.10	4.88
PS	General	0.02	0.13	0.06	5.16	5.20	4.89
	PS	0.02	0.20	0.19	5.14	5.20	4.90

In summary, both the general and reduced VAMs can provide accurate and precise fixed effect estimates whether or not the model assumptions fit the data structure or not. Choosing different models to estimate the same generated data does not yield significantly different estimates. In other words, the fixed effect estimation is not sensitive to the model choice. Meanwhile, there are also no significant differences in the results across three And no significant differences in estimates can be observed when various models are used to fit the same data. Therefore, the fixed effect estimation is not affected by different school compositions. For both School B and School C, the general

and reduced model can provide accurate and precise fixed effect estimates for all the generating and fitting model combinations.

Table 4.8 Absolute Bias of the Estimated Fixed Effect for Each Year from Different Models (School B)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	0.25	0.19	0.14	4.89	5.14	5.11
	GS	0.22	0.19	0.18	5.34	5.42	5.39
	CA	0.07	0.06	0.18	5.40	5.44	5.48
	CC	0.04	0.13	0.16	5.22	5.34	5.20
	LA	0.19	0.14	0.20	5.19	5.27	5.23
	PS	0.13	0.22	0.11	4.88	5.09	5.14
GS	General	0.13	0.24	0.30	5.41	5.31	5.34
	GS	0.19	0.42	0.46	5.16	5.23	5.24
CA	General	0.17	0.24	0.06	5.38	5.39	5.38
	CA	0.16	0.04	0.09	5.20	5.18	5.22
CC	General	0.05	0.14	0.16	5.32	5.33	5.29
	CC	0.23	0.34	0.37	5.14	5.15	4.88
LA	General	0.10	0.15	0.15	5.25	5.22	5.22
	LA	0.05	0.20	0.35	5.18	5.09	4.92
PS	General	0.10	0.10	0.15	5.17	5.22	4.87
	PS	0.12	0.42	0.16	5.09	5.15	4.93

Table 4.9 Absolute Bias of the Estimated Fixed Effect for Each Year from Different Models (School C)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	0.14	0.02	0.03	4.93	5.13	5.10
	GS	0.12	0.18	0.21	5.40	5.32	5.44
	CA	0.06	0.04	0.04	5.40	5.42	5.42
	CC	0.01	0.08	0.2	5.12	5.17	5.29
	LA	0.19	0.15	0.16	5.32	5.29	5.37
	PS	0.14	0.31	0.19	4.89	5.10	5.20
GS	General	0.03	0.13	0.35	5.40	5.39	5.30
	GS	0.14	0.16	0.28	5.25	5.19	5.21
CA	General	0.29	0.22	0.05	5.35	5.40	5.31
	CA	0.13	0.02	0.01	5.18	5.17	5.28
CC	General	0.03	0.04	0.11	5.28	5.33	5.25
	CC	0.28	0.34	0.37	5.11	5.22	4.84
LA	General	0.13	0.26	0.29	5.26	5.21	5.18
	LA	0.08	0.12	0.18	5.12	5.14	4.89
PS	General	0.17	0.07	0.24	5.17	5.21	4.91
	PS	0.21	0.34	0.12	5.12	5.16	4.96

Teacher Effects Estimation

Before evaluating the teacher effects estimation, a brief description of the generated teacher effects across three years is first presented here. For year 1, the generated teacher effects range from -18.5 to 15.6 with a mean of 0.4 and standard deviation of 10.1; for year 2, they range from -16.1 to 13.5 with a mean of -0.6 and standard deviation of 9.7; and for year 3, they range from -19.0 to 13.9 with a mean of 0.38 and standard deviation of 10.2. To investigate the feasibility of the general model and to compare the general and the reduced models with respect to the teacher effects estimation, two measures of estimated teacher effects were computed for different combinations of data and models.

First, the correlation between the true teacher effects θ's and the estimated teacher effects $\hat{\theta}$'s was computed. The major purpose of estimating teacher effects using VAMs in the educational practice is to rank-order the involved teachers. Therefore, a model can provide accurate estimates in practice if the estimated teacher effects using this model have high correlation with the true teacher effects. Second, the absolute bias of the estimated teacher effects $Bias(\hat{\theta})$ was computed. Besides these two measures, the correlation between estimated teacher effects obtained from various models was also computed to investigate the interrelationship among the general and reduced models regarding the random effect estimation.

The Correlation between the True and Estimated Teacher Effects

Tables 4.10-4.12 show the correlation between the true and estimated teacher effects for all the data and model combinations across three years for three schools. In this study, the correlations reported are the spearman's rank correlation coefficients because the rank order, instead of the absolute value, of the teacher effects is of primary interest in decision making practice. The range of the correlations is from 0.81 to 0.94. The lowest correlation 0.81 is observed when the CA model was used to estimate the School C's general-model-generated data in year 3. The highest correlation 0.94 is always found when the general model was used to estimate the general-model-generated data, for example, the School A's data in year 2 and year 3, School B's data in year 3 and School C's data in year 1. Therefore, the correlation measure supports that, in general, all the models can provide acceptable teacher effect estimates under various assumptions. In order to further examine the different model performances, the following analysis focuses on the pattern observed within each school.

Table 4.10 Correlation Between estimated and true teacher effects for Each Year from Different Models (School A)

Generated	Fitted	Year 1	Year 2	Year 3
General	General	0.93	0.94	0.94
	GS	0.82	0.83	0.84
	CA	0.82	0.83	0.83
	CC	0.85	0.84	0.84
	LA	0.84	0.85	0.85
	PS	0.92	0.91	0.91
GS	General	0.86	0.86	0.87
	GS	0.89	0.91	0.91
CA	General	0.86	0.85	0.85
	CA	0.92	0.91	0.89
CC	General	0.85	0.86	0.86
	CC	0.86	0.86	0.87
LA	General	0.86	0.84	0.87
	LA	0.87	0.86	0.89
PS	General	0.91	0.91	0.92
	PS	0.90	0.92	0.92

For School A, when the data were generated using the general model, as expected, the general model itself gives the highest correlation. Meanwhile, the correlation obtained by the PS model is only lower by 0.01or 0.02. Again, this supports our assumption that the PS model has the closest results to the general model because it has the most similar model specification as the general one. The CC and LA model have the relatively close results - the correlation is about

Table 4.11 Correlation Between estimated and true teacher effects for Each Year from Different Models (School B)

Generated	Fitted	Year 1	Year 2	Year 3
General	General	0.92	0.93	0.94
	GS	0.84	0.85	0.86
	CA	0.86	0.85	0.84
	CC	0.92	0.90	0.90
	LA	0.91	0.91	0.92
	PS	0.91	0.92	0.93
GS	General	0.88	0.89	0.88
	GS	0.88	0.90	0.90
CA	General	0.91	0.89	0.88
	CA	0.91	0.90	0.89
CC	General	0.89	0.88	0.88
	CC	0.90	0.91	0.90
LA	General	0.88	0.89	0.88
	LA	0.90	0.91	0.89
PS	General	0.90	0.92	0.91
	PS	0.91	0.92	0.93

0.84 or 0.85. And the GS and CA model have the relatively close, but the lowest, correlation. When the data were generated using the reduced models, the reduced models themselves perform very well - even the lowest correlation is 0.86, which is observed when the correct models were used to estimate the CC-model-generated data and LA-model-generated data. However, the general model does not perform equally well under various conditions. When the data were generated using the CC, LA and PS model, the general model results are as good as those obtained from the correct models. While when the data were generated using the GS and CA model, the correlation of the general model is much lower than that of the correct one, for example, the differences between the

Table 4.12 Correlation Between estimated and true teacher effects for Each Year from Different Models (School C)

Generated	Fitted	Year 1	Year 2	Year 3
General	General	0.94	0.93	0.93
	GS	0.84	0.85	0.85
	CA	0.83	0.82	0.81
	CC	0.82	0.83	0.82
	LA	0.85	0.84	0.84
	PS	0.92	0.92	0.90
GS	General	0.86	0.86	0.88
	GS	0.90	0.92	0.92
CA	General	0.86	0.85	0.86
	CA	0.92	0.91	0.91
CC	General	0.84	0.86	0.87
	CC	0.86	0.86	0.88
LA	General	0.86	0.85	0.88
	LA	0.87	0.88	0.89
PS	General	0.91	0.90	0.91
	PS	0.91	0.91	0.92

correlations obtained using the general model and the correct model when the data were generated using the CC model are 0.06 for the year 1 and year 2 data. Based on the above analysis, a conclusion can be drawn that the estimation for the general-model-generated data and the estimation for the reduced-model-generated data show the same pattern. That is, with respect to the teacher effect estimates, the general and the PS model have the close results; the CC and LA model have the similar results; and results obtained by

the GS and CA model are close to each other, but have the most differences with the general model.

The results for the School C data show the same pattern with those for the School A data, although they are slightly different in magnitude. That is, when the data were generated using the general model, the general model provides the highest correlations between the true and estimated teacher effects for all three years; the CC and LA model provide the similar and worse results; the GS and CA model results are similar and worse than the CC and LA model results. When the data were generated using the reduced models, the correct models provide better estimates; the general model provides slightly worse results for the GS and CA model and much worse results for the CC and LA model.

However, the School B data tell a different story. Two types of improvements can be observed when switching from the School A or C data to the School B data. First, for the general-model-generated data, not only the PS model estimates, but also the CC and LA model estimates are as good as the general model estimates - the CC and LA model correlations are all greater than 0.90. While the GS and CA model results, which are around 0.85, are still worse than the general model one. Second, when the data were generated using the CC and LA model, higher correlation can be obtained from using both the general and the correct model. Specifically, switching from the School A or C data to the School B data, the correlations increase from around 0.85 to around 0.90 for the CC-model-generated and the LA-model-generated data by using both the general and the correct model. These two types of improvements indicate that the CC and LA model are sensitive to the school composition. To be specific, the CC and LA model perform noticeably better for the school that has roughly the same mix of advantaged and

disadvantaged students (as School B) than for those have unbalanced mix (as School A or C). The performance of the general model for estimating the CC-model-generated or LA-model-generated data also improves when switching from the balanced mixed school to unbalanced mixed school.

The Mean Absolute Bias of the Estimated Teacher Effects

Table 4.13 Mean Absolute Bias of the Estimated Teacher Effects for Each Year from Different Models (School A)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	4.40	3.02	3.91	5.91	6.16	6.03
	GS	3.28	7.73	6.47	6.60	6.75	6.61
	CA	3.19	7.91	6.53	6.68	6.59	6.64
	CC	3.01	7.76	6.43	6.37	6.48	6.55
	LA	4.91	3.11	5.33	6.30	6.45	6.44
	PS	4.71	3.24	3.94	5.79	6.20	6.14
GS	General	5.00	4.11	4.97	6.54	6.37	6.42
	GS	5.03	3.98	4.96	6.25	6.31	6.32
CA	General	5.10	4.08	4.91	6.49	6.55	6.36
	CA	5.23	4.05	4.85	6.19	6.20	6.37
CC	General	3.52	3.92	4.16	6.39	6.37	6.31
	CC	3.52	4.13	4.33	6.13	6.16	5.86
LA	General	4.85	3.38	4.04	6.36	6.25	6.31
	LA	4.88	3.37	3.89	6.20	6.14	5.88
PS	General	4.45	2.99	3.96	6.21	6.26	5.89
	PS	4.53	3.02	3.99	6.19	6.26	5.90

Tables 4.13-4.15 show the mean absolute bias and the posterior standard deviation of the estimated teacher effects for each year from different models for School A, B and C, respectively. There are 16 teachers being evaluated within each school for each year.

The mean absolute bias presented in the tables is calculated as $\frac{1}{16}\sum_{n=1}^{16}\left|\hat{\theta}_n - \theta_n\right|$.

Generally speaking teacher effects are not estimated with great accuracy – the mean absolute bias in these three tables ranges from 2.91 to 7.95. Recent literatures (e.g., McCaffrey et al; RAND 2003) show that there exist several sources of error in estimated teacher effects. It is very difficult to make any meaningful inference based on the magnitude of the teacher effects using the current methodologies so that the magnitude of the estimated individual teacher effect is not of primary interest in the practical

Table 4.14 Mean Absolute Bias of the Estimated Teacher Effects for Each Year from Different Models (School B)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	4.46	3.00	3.87	5.91	6.34	6.17
	GS	3.06	7.74	6.45	6.61	6.92	6.58
	CA	4.86	3.08	5.29	6.85	6.59	6.57
	CC	3.28	7.95	6.60	6.41	6.63	6.70
	LA	3.28	7.94	6.54	6.25	6.37	6.47
	PS	4.41	2.95	3.89	5.97	6.22	6.24
GS	General	5.03	4.05	4.99	6.59	6.34	6.36
	GS	5.15	4.03	5.06	6.34	6.46	6.49
CA	General	5.14	4.08	4.90	6.46	6.55	6.28
	CA	5.27	4.01	5.06	6.26	6.40	6.50
CC	General	3.63	3.90	4.15	6.48	6.32	6.27
	CC	3.55	3.93	4.35	6.27	6.15	5.85
LA	General	4.89	3.39	3.94	6.56	6.24	6.46
	LA	4.88	3.41	3.94	6.14	6.15	5.92
PS	General	4.48	2.93	3.98	6.23	6.43	6.08
	PS	4.51	2.95	3.97	6.31	6.27	5.83

accountability system. Therefore, in this simulation study, how accurate and precise the individual teacher effect can be estimated is also not of interest. The mean absolute bias in those three tables is mainly used to investigate the differences and similarities among the models.

Table 4.15 Mean Absolute Bias of the Estimated Teacher Effects for Each Year from Different Models (School C)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	4.46	3.03	3.88	5.99	6.36	6.03
	GS	3.19	7.84	6.61	6.67	6.88	6.74
	CA	3.26	7.63	6.56	6.68	6.62	6.59
	CC	3.12	7.55	6.46	6.48	6.44	6.50
	LA	4.92	3.01	5.31	6.29	6.65	6.62
	PS	4.42	3.02	3.86	5.76	6.27	6.33
GS	General	5.00	4.07	5.07	6.69	6.57	6.54
	GS	5.23	4.03	5.18	6.43	6.23	6.34
CA	General	5.09	4.08	4.92	6.53	6.72	6.41
	CA	5.42	4.05	5.15	6.32	6.24	6.29
CC	General	3.59	4.03	4.32	6.47	6.49	6.32
	CC	3.58	3.89	4.31	6.25	6.25	5.95
LA	General	4.86	3.39	3.96	6.35	6.34	6.41
	LA	4.88	3.41	3.96	6.33	6.07	5.84
PS	General	4.45	3.01	4.00	6.21	6.28	5.98
	PS	4.47	2.91	3.97	6.22	6.35	5.86

Comparing the results across three years, it is apparent that the relatively larger biases (> 7.00) or smaller biases (<3.00) tend to be observed from the year 2 results and the ranges of the biases from the year 1 and 3 results are relatively smaller. Specifically, the bias ranges from 3.01 to 5.42 for the year 1 data and from 3.86 to 6.61 for the year 2

data. The reason for this pattern might be that the range of the generated teacher effects for year 2 is smaller than those for year 1 and year 3.

Focusing on the three years results within each school, the following pattern can be found for School A data. When the data were generated using the general model, the general model has the lowest bias of 3.78 and the PS model has the second lowest bias of 3.96. While the GS, CA and CC model produce relative larger bias. When the data were generated using the reduced models, the correct model and the general model produce very close results (In some extreme cases, for example, when the data were generated using the LA model, the general model provides even smaller biases. However the improvement of the general model is very small, which might be attributed to the estimate errors since the teacher effect estimates themselves are of great accuracy). Therefore, the bias measure also supports our assumption that the general model performs best when it is the true model and it also provides the similar quality results as the reduced model even when the reduced model is the true model. In contrast to the correlation measure, this pattern is also true for the School B and School C data, and no evident impact can be found of the school composition on the model performance.

The Correlation between the Estimated Teacher Effects from Different Models

To further investigate the interrelationship and to compare the differences and similarities among all the models, the correlation between the estimated teacher effects from different models were also computed. Because the general model and the reduced model provide very close estimates (The correlation is consistently greater than 0.95.) when the correct model is the reduced model, we plot the estimated teacher effects obtained from the general model and the reduced model to see how these estimated

teacher effects are distributed. Figures 4.1-4.5 present the correlations obtained from using the five reduced-model-generated data, respectively. Each figure contains 9 panels. Within each panel the points represent the individual estimated teacher effect. Rows of panels correspond to three schools and columns of panels correspond to three years. It is apparent that all the points are almost on a straight line, which indicates highly correlated relationship between the estimated teacher effects from the reduced model and those from the general model. However, the distribution of the points within each panel varies from model to model, and it also varies from year to year within each model. For example, comparing the year 2 panels in Figure 4.1 and 4.2, it is easy to find that the points tend to be more separated along the scales (from -30 to 30) using the GS-model-generated data than using the CA-model-generated data. Furthermore, comparing the year 1, year 2 and year 3 panels in Figure 4.1, we can find that these three plots are also different in how the estimated teacher effects spread along the scales. Year 1 data show three outliers - two of them have much lower teacher effects and one of them has much higher teacher effect than most of the teachers. Year 3 data show only one outlier that is far left behind all the other teachers. The same pattern can be observed from Figure 4.1 to 4.5, which indicates that, with respect to identifying the outliers, all the models present similar results regardless of the true underlying data structure.

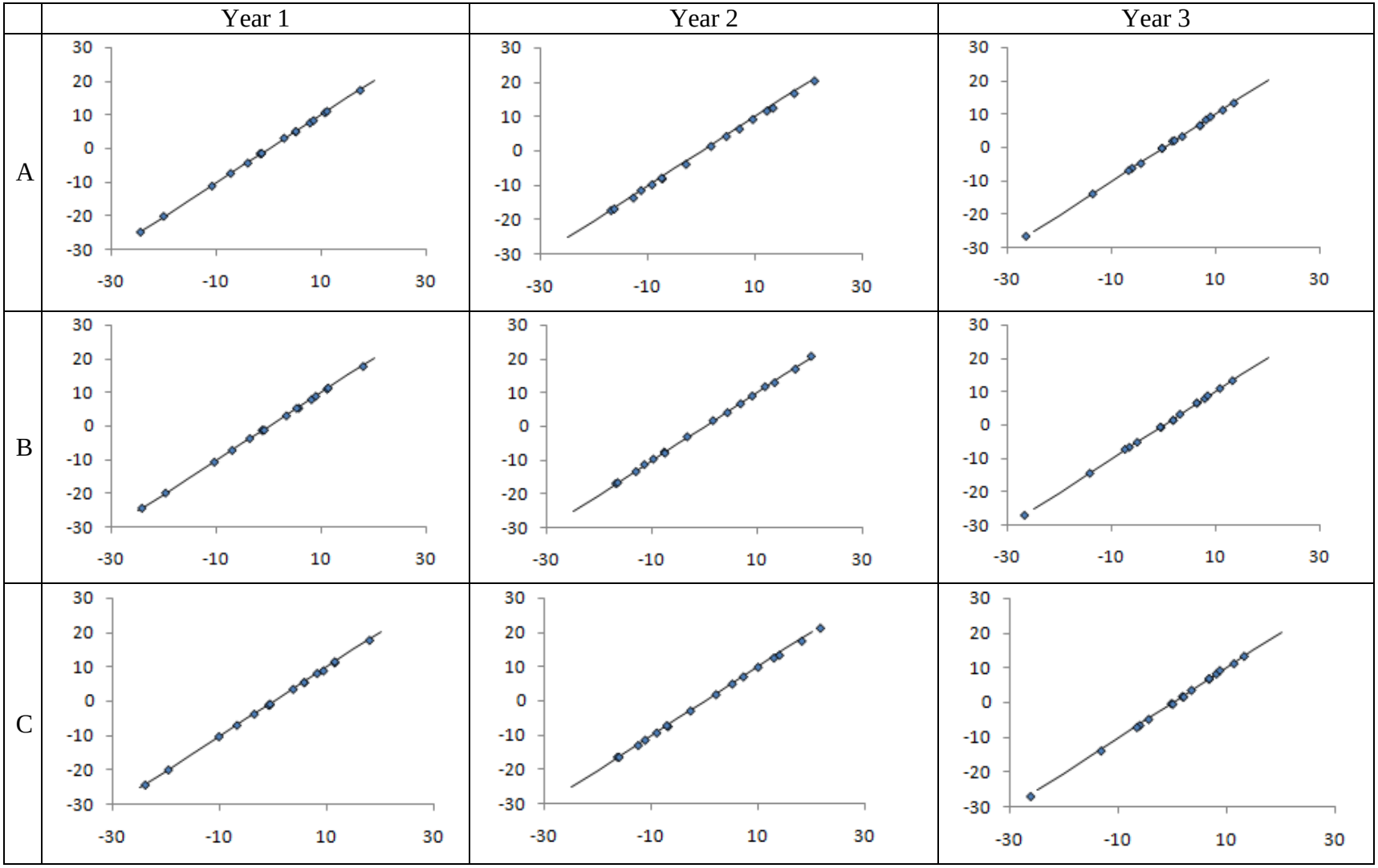

Figure 4.1 Correlation between the Estimated Teacher Effects from the General and the GS Model (GS Data)

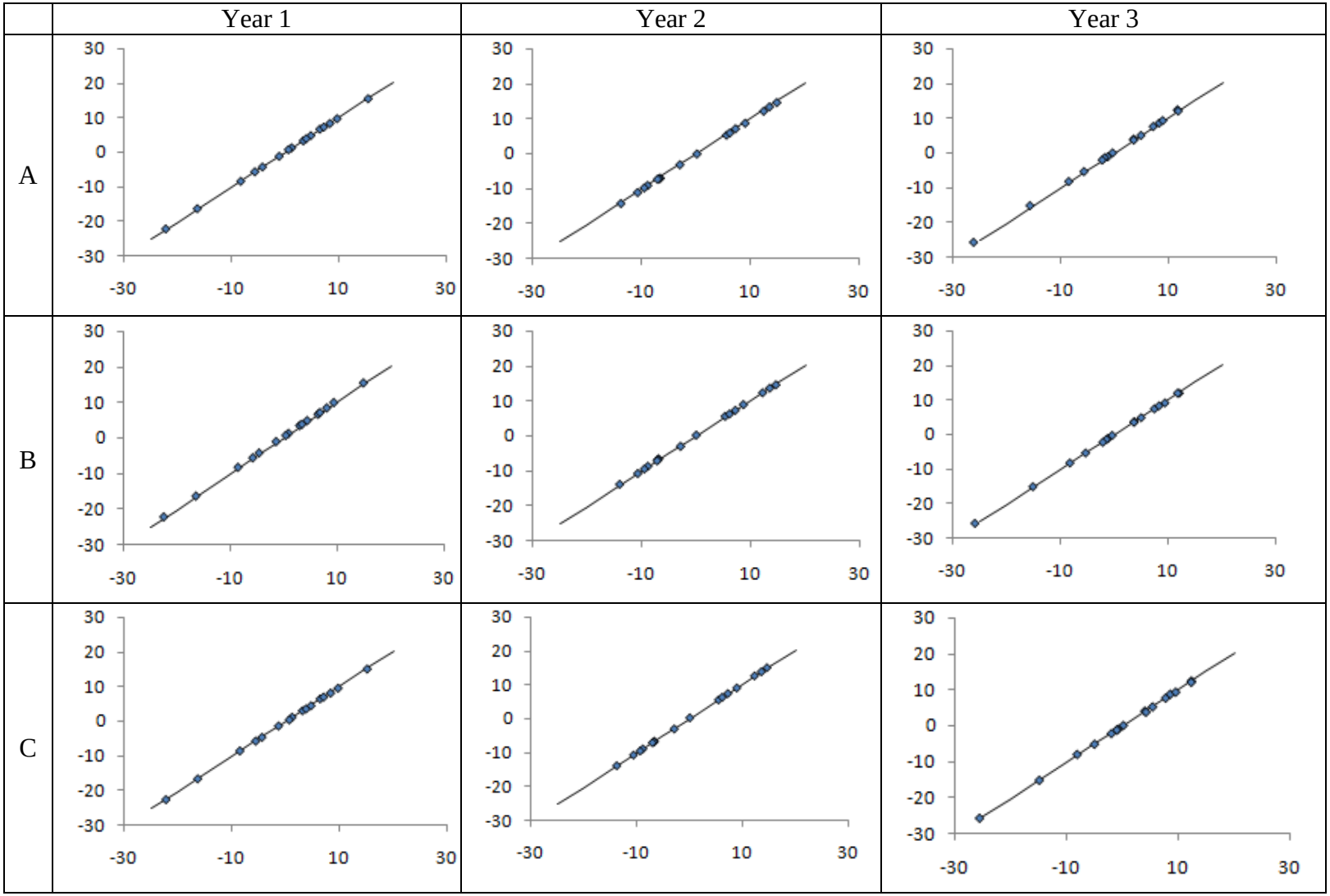

Figure 4.2 Correlation between the Estimated Teacher Effects from the General and the CA Model (CA Data)

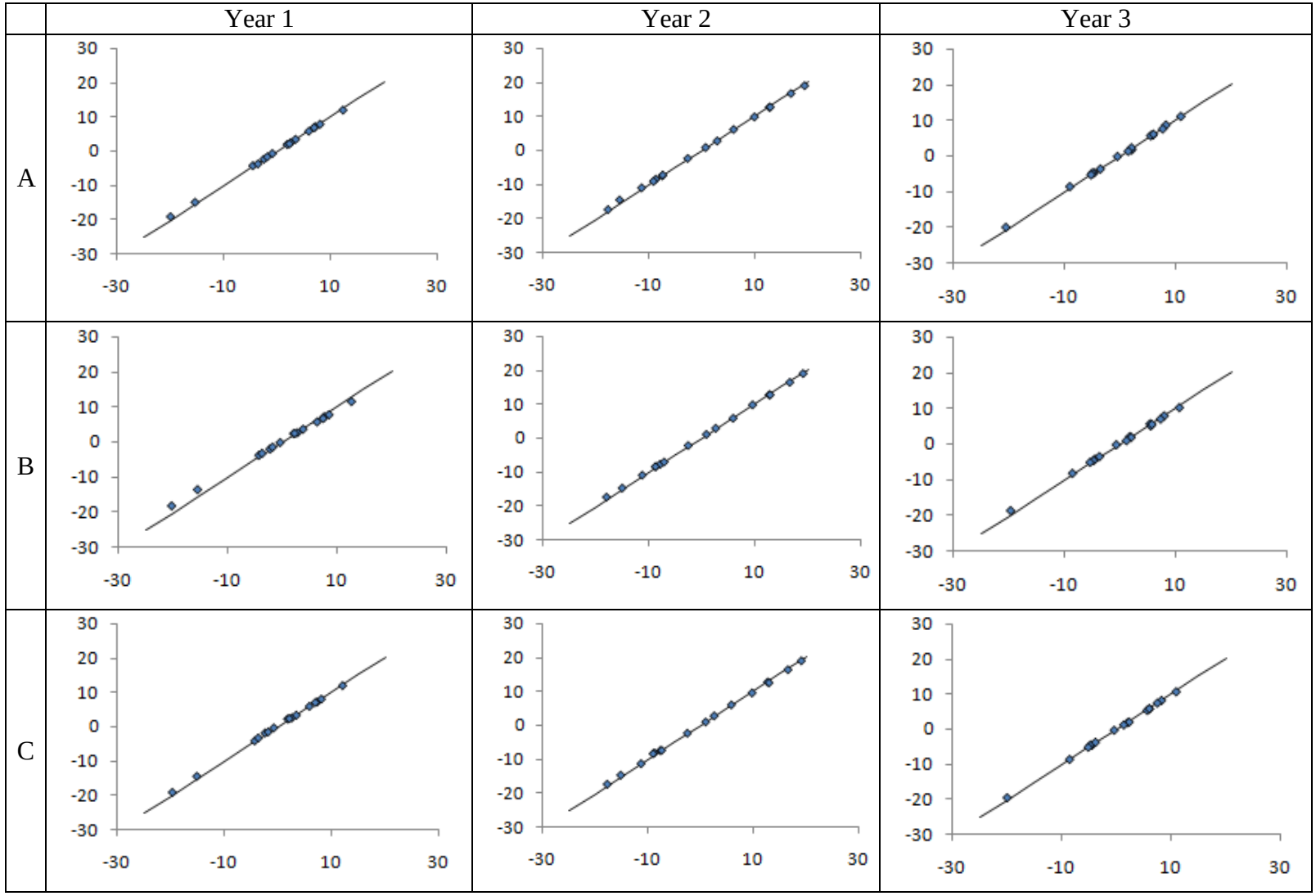

Figure 4.3 Correlation between the Estimated Teacher Effects from the General and the CC Model (CC Data)

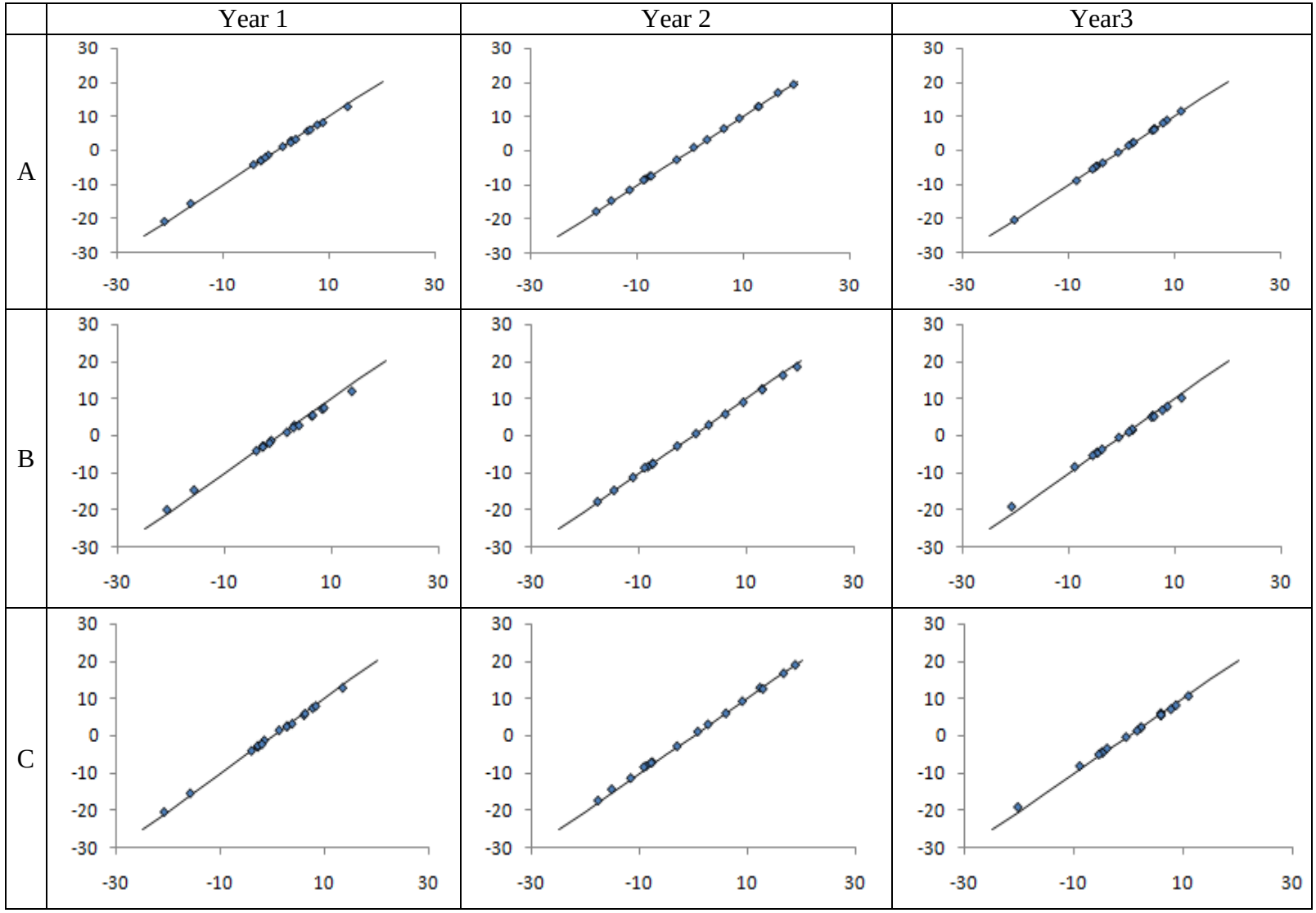

Figure 4.4 Correlation between the Estimated Teacher Effects from the General and the LA Model (LA Data)

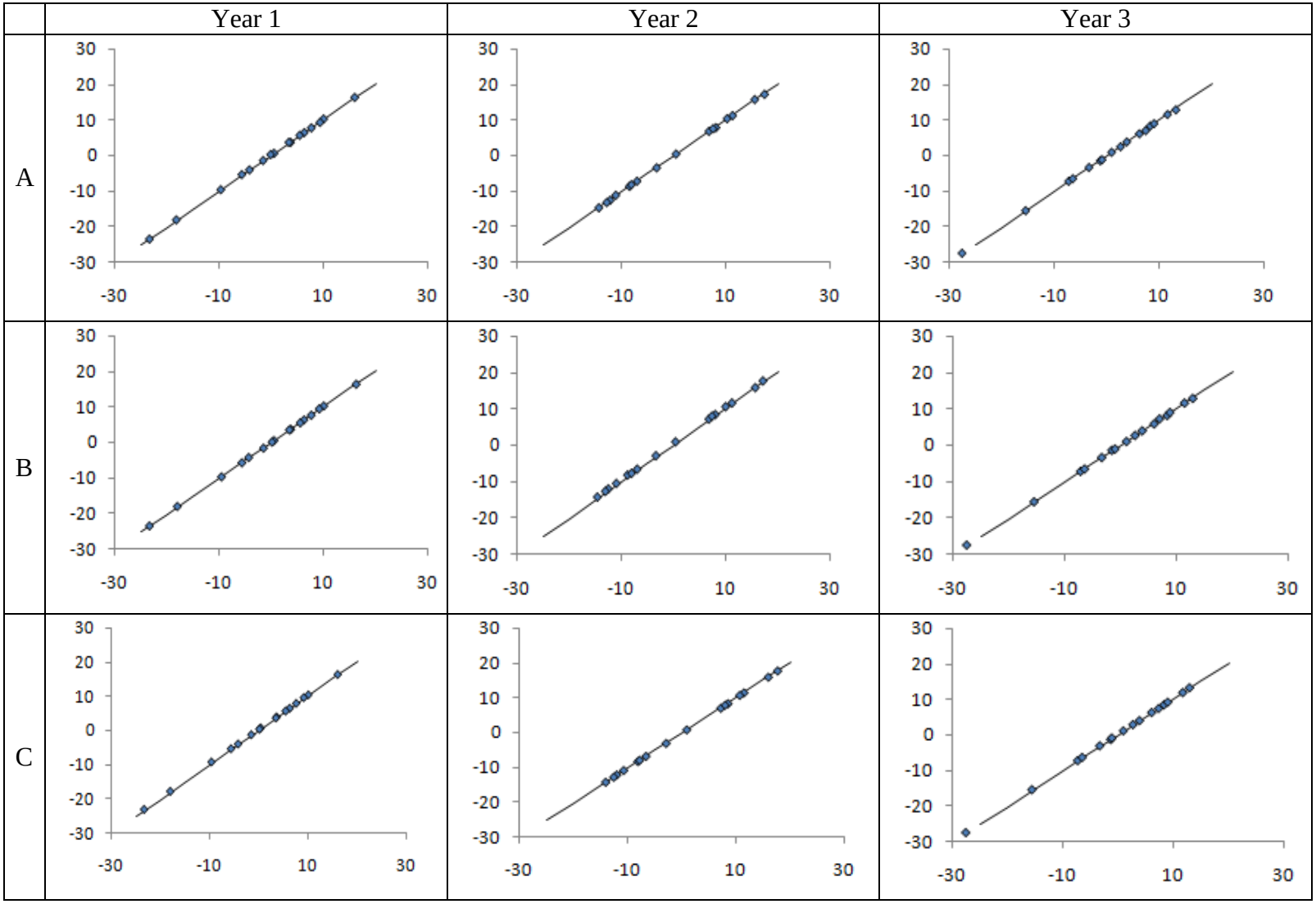

Figure 4.5 Correlation between the Estimated Teacher Effects from the General and the PS Model (PS Data)

We use Tables 4.16-4.18 to report the pair-wise correlation between the estimated teacher effects when the data were generated using the general model. The correlation ranges from 0.71 to 0.97 across three schools for three years. Within each school, data across three years provide very similar results. For School A data, the general-PS pair gives the highest correlation. On average, the general model has the highest correlation with the other models. This again proves the feasibility and advantage of the general model. The GS and CA model have a relatively high correlation with each other (greater than 0.90), whereas they have much lower correlations with the CC model and the LA

Table 4.16 Pair-Wise Correlation between the Estimated Teacher Effects from Different Models Using the General-Model-Generated data (School A)

		General	GS	CA	CC	LA	PS
Year 1	General	1.00	0.77	0.79	0.93	0.94	0.96
	GS		1.00	0.93	0.72	0.73	0.80
	CA			1.00	0.72	0.74	0.80
	CC				1.00	0.93	0.83
	LA					1.00	0.84
	PS						1.00
Year 2	General	1.00	0.77	0.78	0.92	0.91	0.94
	GS		1.00	0.89	0.73	0.72	0.80
	CA			1.00	0.72	0.72	0.80
	CC				1.00	0.91	0.83
	LA					1.00	0.83
	PS						1.00
Year 3	General	1.00	0.82	0.82	0.92	0.94	0.95
	GS		1.00	0.91	0.72	0.72	0.82
	CA			1.00	0.71	0.72	0.82
	CC				1.00	0.93	0.84
	LA					1.00	0.85
	PS						1.00

model (smaller than 0.75). School C data present the similar results with School A data. However, School B data present differences in the correlation of the PS-CC pair and the correlation of the PS-LA pair. Switching to the School B data, the PS-CC and PS-LA correlations increase from around 0.83 to around 0.94. This significant change indicates that the differences that exist between the PS and CC or LA model can be reduced if the data are from a balanced mixed school. It is natural to relate this phenomenon with the control of the

Table 4.17 Pair-Wise Correlation between the Estimated Teacher Effects from Different Models Using the General-Model-Generated data (School B)

		General	GS	CA	CC	LA	PS
Year 1	General	1.00	0.82	0.80	0.93	0.95	0.97
	GS		1.00	0.94	0.73	0.72	0.82
	CA			1.00	0.73	0.75	0.83
	CC				1.00	0.89	0.94
	LA					1.00	0.93
	PS						1.00
Year 2	General	1.00	0.83	0.81	0.92	0.94	0.96
	GS		1.00	0.93	0.74	0.73	0.81
	CA			1.00	0.74	0.72	0.83
	CC				1.00	0.90	0.94
	LA					1.00	0.94
	PS						1.00
Year 3	General	1.00	0.80	0.81	0.93	0.95	0.97
	GS		1.00	0.93	0.70	0.73	0.83
	CA			1.00	0.72	0.74	0.83
	CC				1.00	0.90	0.94
	LA					1.00	0.94
	PS						1.00

covariates because the major difference between the PS and the CC or LA model is that the PS model includes covariates but the CC and LA model don't. Therefore, one possible explanation is that the inclusion of the covariates does not make much difference if the characteristics described by the covariates distribute homogenously in the sample.

Table 4.18 Pair-Wise Correlation between the Estimated Teacher Effects from Different Models Using the General-Model-Generated data (School C)

		General	GS	CA	CC	LA	PS
Year 1	General	1.00	0.82	0.83	0.94	0.95	0.97
	GS		1.00	0.92	0.75	0.73	0.82
	CA			1.00	0.75	0.75	0.82
	CC				1.00	0.93	0.83
	LA					1.00	0.84
	PS						1.00
Year 2	General	1.00	0.81	0.82	0.90	0.90	0.96
	GS		1.00	0.91	0.74	0.74	0.82
	CA			1.00	0.75	0.73	0.83
	CC				1.00	0.92	0.84
	LA					1.00	0.84
	PS						1.00
Year 3	General	1.00	0.82	0.83	0.91	0.92	0.97
	GS		1.00	0.92	0.74	0.73	0.84
	CA			1.00	0.73	0.74	0.83
	CC				1.00	0.92	0.83
	LA					1.00	0.83
	PS						1.00

Teacher Variance Components Estimation

The MCMC algorithm designed for the general model also allows to simultaneously estimating the teacher variance component, which is the variance of the teacher effects within one year. For the convenience of comparing with the true standard deviation of the

teacher effects distribution, Tables 4.19 to 4.21 present the square root of the estimated teacher variance components and the posterior standard deviation of the estimates across three years for School A, B and C, respectively. All of the estimates are consistently greater than the true value 10. There might be many reasons why the teacher variance component estimates are larger than it would be. One of them is that the number of students linked to each teacher is small, which would introduce larger measurement errors. All of the posterior standard deviations for these estimates are between 2 and 3, which indicate that the precision of the estimation is acceptable.

Table 4.19 The Estimated Teacher Variance Components for Each Year from Different Models (School A)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	10.8	10.6	11.3	2.1	2.0	2.3
	GS	12.5	11.8	13.0	2.4	2.5	2.6
	CA	12.6	11.6	12.9	2.5	2.6	2.7
	CC	11.2	11.0	12.1	2.3	2.5	2.4
	LA	11.4	10.9	11.7	2.5	2.3	2.6
	PS	11.0	10.9	11.2	2.2	2.1	2.2
GS	General	12.2	11.7	12.8	2.2	2.3	2.4
	GS	12.2	11.5	12.7	2.2	2.1	2.2
CA	General	12.3	11.2	12.7	2.3	2.4	2.2
	CA	12.2	11.2	12.5	2.1	2.2	2.2
CC	General	11.3	11.1	11.3	2.3	2.3	2.4
	CC	11.1	11.1	11.2	2.2	2.3	2.1
LA	General	11.4	11.1	11.5	2.4	2.5	2.4
	LA	11.4	11.3	11.4	2.3	2.4	2.4
PS	General	11.3	11.3	11.5	2.3	2.5	2.2
	PS	11.3	10.9	11.2	2.1	2.3	2.1

Table 4.20 The Estimated Teacher Variance Components for Each Year from Different Models (School B)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	11.1	10.7	11.4	2.2	2.1	2.2
	GS	12.1	12.0	12.8	2.3	2.5	2.5
	CA	12.6	11.7	13.0	2.6	2.6	2.5
	CC	11.4	11.2	12.0	2.3	2.5	2.5
	LA	11.0	11.0	11.5	2.4	2.5	2.5
	PS	11.2	10.8	11.3	2.1	2.2	2.3
GS	General	12.1	11.8	12.7	2.3	2.2	2.3
	GS	12.1	11.7	12.6	2.1	2.3	2.3
CA	General	12.4	11.3	12.9	2.2	2.4	2.2
	CA	12.2	11.2	12.7	2.3	2.3	2.2
CC	General	11.4	11.1	11.5	2.4	2.4	2.3
	CC	11.0	11.1	11.1	2.2	2.2	2.1
LA	General	11.5	11.5	11.5	2.5	2.4	2.4
	LA	11.3	11.4	11.3	2.4	2.2	2.4
PS	General	11.3	11.2	11.4	2.4	2.4	2.2
	PS	11.2	11.0	11.2	2.2	2.2	2.1

Comparing across three tables, no significant difference can be observed among the results for the three schools, and there are several similar patterns can be found in all the three tables. First, the estimates for the year 2 data are consistently lower than those for the year 1 and year 3. This pattern can also be observed from the Figures 4.1-4.5, in which the point distribution shown by the year 2 data has lower level of dispersion than those shown by the year 1 and 3 data. Again, one possible explanation for this is that the generated teacher effects for year 2 have smaller range and smaller variance. Second, for the general-model-generated data, the general and PS model provide similar results and they are closest to the true value 10; the CC and LA model results are slightly more

biased; and the GS and CA model results are much more biased. When the data were generated using the reduced model, the general model result is just slightly larger than the result yielded by the correct model. Generally speaking, the estimated teacher variance components from all of the models for different data are acceptable - the largest bias is 3.2.

Table 4.21 The Estimated Teacher Variance Components for Each Year from Different Models (School C)

		Estimates			SDs		
Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year2	Year 3
General	General	11.0	10.7	11.3	2.2	2.2	2.3
	GS	12.4	11.8	13.2	2.5	2.6	2.6
	CA	12.5	11.8	12.9	2.4	2.7	2.7
	CC	11.3	11.1	12.0	2.4	2.6	2.4
	LA	11.3	11.0	11.6	2.4	2.4	2.6
	PS	11.2	10.9	11.4	2.2	2.3	2.2
GS	General	12.2	11.7	12.6	2.4	2.5	2.4
	GS	12.0	11.6	12.5	2.2	2.2	2.2
CA	General	12.4	11.2	12.7	2.3	2.4	2.3
	CA	12.3	11.3	12.6	2.2	2.3	2.4
CC	General	11.2	11.1	11.4	2.3	2.4	2.4
	CC	11.1	11.2	11.2	2.3	2.4	2.1
LA	General	11.5	11.2	11.6	2.5	2.5	2.4
	LA	11.4	11.4	11.5	2.4	2.4	2.2
PS	General	11.4	11.2	11.6	2.3	2.5	2.2
	PS	11.3	10.8	11.3	2.2	2.3	2.0

Teacher Effect Persistence Estimation

Table 4.22 shows the estimated teacher effect persistence parameters obtained from the only two models that assume the persistence of previous years' teacher effect is diminished--the general and the PS model. Comparing with the true value for generating

ϕ_{21}, ϕ_{31} and ϕ_{32} (0.2, 03 and 0.3, respectively), one can find that both the models can provide relatively accurate estimates; and all the posterior standard deviations are between 2 and 3. Moreover, it is apparent that the biases of the estimates are consistently positive. Especially, when generating and fitting model combination is given, biases for ϕ_{21} tends to be even larger than those for ϕ_{31} and ϕ_{32}. To some extent, this result supports McCaffrey's criticism of other researchers' exaggerate claims on the persistence effect. However, at this point, there is no clear explanation or interpretation for why the persistence parameter would be overestimated. Also, there is no evident impact can be observed from using different schools' data.

Table 4.22 The Estimated Teacher Effect Persistence Parameters from the General and PS model

			Estimates			SDs		
School	Generated	Fitted	$\phi_{21}(0.2)$	$\phi_{31}(0.3)$	$\phi_{32}(0.3)$	ϕ_{21}	ϕ_{31}	ϕ_{32}
A	General	General	0.26	0.32	0.32	0.03	0.05	0.05
		PS	0.31	0.34	0.35	0.06	0.06	0.06
	PS	General	0.26	0.34	0.35	0.04	0.05	0.03
		PS	0.25	0.33	0.32	0.04	0.04	0.04
B	General	General	0.26	0.34	0.35	0.02	0.04	0.05
		PS	0.32	0.35	0.36	0.05	0.05	0.06
	PS	General	0.26	0.34	0.35	0.03	0.04	0.05
		PS	0.25	0.33	0.32	0.03	0.05	0.04
C	General	General	0.26	0.34	0.33	0.03	0.04	0.05
		PS	0.30	0.34	0.35	0.05	0.05	0.06
	PS	General	0.27	0.35	0.37	0.04	0.05	0.05
		PS	0.26	0.33	0.32	0.05	0.05	0.04

Random Student Effects Estimation

Table 4.23 shows the only measure for students' own random effect - the estimated student effect component and their posterior standard deviations. Also, only the results

for the two models that take into account the students' own random effect are presented. All the estimates are consistently smaller than the true value 5. However, when the correct model is used to fit the data, the bias of the estimate is only around 1. For example, when the general model is chosen to estimate to general-model-generated data, the biases of the estimates for School A data are 0.9, 0.8 and 0.7 for three years. No apparent trend for the estimates over three years can be found. The impact of school composition on the estimation is also not clear for this simulation.

Table 4.23 The Estimated Student Effect Component from the General and CC model

			Estimates			SDs		
School	Generated	Fitted	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
A	General	General	4.1	4.2	4.3	2.1	1.9	1.9
		CC	3.4	3.6	4.0	2.3	2.2	2.0
	CC	General	3.8	3.9	4.1	2.3	2.4	2.4
		CC	4.2	4.1	4.3	2.1	2.3	2.3
B	General	General	4.2	4.0	4.2	2.2	2.0	1.9
		CC	3.6	3.5	4.0	2.2	2.2	2.1
	CC	General	3.9	3.8	4.2	2.4	2.4	2.3
		CC	4.2	3.9	4.3	2.1	2.4	2.2
C	General	General	4.2	3.9	4.4	2.1	2.1	2.0
		CC	3.7	3.6	3.9	2.4	2.2	2.1
	CC	General	3.9	4.1	3.8	2.4	2.3	2.4
		CC	4.1	4.2	3.9	2.2	2.2	2.3

Estimation of the Teachers' Contribution to Total Variance

The estimated teachers' contribution to total variance is the percentage of the estimated teacher variance component in the estimated total variance. The estimated total variance is the sum of the estimated teacher variance component, estimated student

variance component and estimated residual error. Results obtained using School A, B and C data only have slight differences in magnitude. School composition does not have significant impact on the teachers' contribution to the total variance estimation. However, it is easy to find that the estimated teachers' contribution is lower for the year 2 data than

Table 4.24 The Estimated Teachers' Contribution to Total Variance for Each Year from Different Models (%) (School A)

Generated	Fitted	Year 1	Year 2	Year 3
General	General	9.5	9.3	9.6
	GS	12.1	11.5	12.0
	CA	11.9	11.3	11.8
	CC	10.9	10.9	11.2
	LA	11.2	10.8	11.4
	PS	9.3	9.3	9.4
GS	General	8.8	8.7	8.9
	GS	11.8	11.7	11.8
CA	General	8.9	8.8	8.8
	CA	11.5	11.0	11.2
CC	General	9.2	9.0	9.1
	CC	10.4	10.5	10.5
LA	General	9.2	9.0	9.2
	LA	10.6	10.5	10.6
PS	General	9.2	9.2	9.3
	PS	9.3	9.3	9.4

for the year 1 and 3 data under every condition. This is consistent with the teacher effect estimation and teacher variance component estimation. The estimates range from 8.6 to 12.2. When the data are fitted by the general or the PS model, regardless of the generating model, the estimates are lower than the true value 10. On the other hand, when fitting the data using other models, regardless of the generating models, the estimates are

greater than 10. Therefore, we can conclude that the general and PS model tend to underestimate the teachers' contribution to the total variance, whereas the other models tend to overestimate that.

Table 4.25 The Estimated Teachers' Contribution to Total Variance for Each Year from Different Models (%) (School B)

Generated	Fitted	Year 1	Year 2	Year 3
General	General	9.5	9.2	9.6
	GS	12.1	11.5	12.1
	CA	11.8	11.4	11.9
	CC	11.0	10.9	11.1
	LA	11.3	10.8	11.3
	PS	9.5	9.4	9.5
GS	General	8.7	8.6	8.9
	GS	11.9	11.7	11.8
CA	General	8.9	8.7	8.8
	CA	11.5	11.1	11.2
CC	General	9.2	9.0	9.1
	CC	10.4	10.5	10.7
LA	General	9.3	9.2	9.3
	LA	10.6	10.5	10.7
PS	General	9.2	9.0	9.1
	PS	9.3	9.1	9.2

Table 4.26 The Estimated Teachers' Contribution to Total Variance for Each Year from Different Models (%) (School C)

Generated	Fitted	Year 1	Year 2	Year 3
General	General	9.4	9.4	9.5
	GS	12.2	11.3	11.9
	CA	11.8	11.3	11.9
	CC	11.0	10.9	11.3
	LA	11.2	10.7	11.5
	PS	9.5	9.3	9.5
GS	General	8.7	8.6	8.7
	GS	11.8	11.7	11.9
CA	General	8.8	8.8	8.9
	CA	11.3	11.0	11.3
CC	General	9.1	9.0	9.0
	CC	10.5	10.4	10.5
LA	General	9.2	9.1	9.3
	LA	10.6	10.5	10.7
PS	General	9.2	9.1	9.1
	PS	9.2	9.1	9.2

CHAPTER 5

REAL DATA AND ANALYSIS

Data

The data used for this study consist of 3 years of longitudinally linked student-level data from one cohort of 1,836 students from a large statewide achievement testing program. In addition to the scaled scores for Mathematics, the variable of interest in this study is free or reduced price lunch eligibility (FRL). The data contain no missing school-student linkage and no incomplete consecutive scores. To explore the impact of the data structure on model fit, the selected students are purposively divided into three samples according to the SES structure of the schools they attended. The three samples have 10, 12, 12 schools, respectively. In the first sample (Data 1), the chosen schools have similar proportions of FRL students. The FRL rates for Data 1 range from 11% to 23%.The second sample (Data 2) also contains schools with similar proportions of FRL, and the rates range from 61% to 75%. Compared to the first two samples, the third sample (Data 3) is highly heterogeneous with FRL rates ranging from 8% to 75%. The selected students may transfer schools, but they have to stay in the same sample for the duration of the study. Both the general and the five reduced models (the GS, CA, CC, LA and PS models) were used to fit the data, and all the model estimation and comparison were independently conducted for each of the three samples.

Table 5.1 summarizes the FRL rates and mean scores for both FRL students and non-FRL students school by school. For Data 1, the mean scores range from 471 to 492 for FRL students and range from 503 to 520 for non-FRL students; for Data 2, the mean scores range from 462 to 488 for FRL students and range from 503 to 518 for non-FRL

students; and for Data 3, the mean scores range from 469 to 491 for FRL students and range from 501 to 519 for non-FRL students. On average, the mean score for FRL students are around 30 less than the mean score for non-FRL students across all the schools of interest. This is true for all the three samples. From the descriptive analysis of the three samples, one can see that, the simulated data were purposively generated according to the real data structure, although they cannot be exactly the same. The similarities and differences between the simulated and real data are summarized as follows: First, both the simulated and real data show that student scores increase across years and the non-FRL students have higher scores and faster gains. The mean score is

Table 5.1 FRL Rate and Mean Score for FRL and non-FRL Students from Different Schools

	Data1			Data 2			Data 3		
School	FRL (%)	Mean Score		FRL (%)	Mean Score		FRL (%)	Mean Score	
		FRL	non-FRL		FRL	non-FRL		FRL	non-FRL
1	11	477.1	520.3	65	471.4	503.2	8	476.1	511.2
2	12	490.3	514.5	67	475.5	508.8	11	469.7	508.9
3	14	492.5	510.0	67	485.3	506.6	23	471.6	515.2
4	14	471.0	521.5	68	479.7	502.4	37	469.4	510.5
5	16	480.9	515.6	69	488.0	505.7	39	475.2	519.4
6	17	482.5	517.9	69	487.1	514.4	44	491.3	518.9
7	20	475.8	510.3	69	477.3	518.3	45	484.7	514.3
8	21	477.2	513.1	73	473.9	510.0	52	473.9	509.8
9	21	489.3	506.9	74	469.5	511.9	57	479.4	513.7
10	23	487.8	503.1	74	477.4	509.4	60	480.5	515.4
11				75	473.8	513.7	65	483.9	505.3
12				78	462.3	518.6	75	490.2	501.9

around 220 for the simulated data and is around 500 for the real data. Second, for both data, three different samples are created to represent different teacher or school compositions. However, for the simulated data, all the classrooms within each sample have the same proportion of the non-FRL student, whereas in the real data, different schools have different proportions of the non-FRL students, especially, the third sample is highly heterogeneous. Third, the simulated data have 400 students' scores and 16 teachers of interest for each year, whereas the real data have 1836 students' scores and about 12 schools for each year.

Analysis and Comparison of Model Estimation

In the real data analysis, DIC is used for the model comparison in terms of the overall goodness of fit. Within the same data, the overall goodness of fit will be compared among general and all the reduced models. For the fixed-effect variables, such as mean scores and SES variable, the posterior mean and standard deviation obtained from the MCMC algorithm are reported as the estimated coefficients for each year and each subject and their posterior standard deviations.

For the estimated school effects, several measures are considered: estimates of individual school effects and the overall contributions of school to variability in student outcomes. The MCMC algorithm provides the estimate of each individual school's effect for all the models of interest. The spearman's rank correlation between the estimated school effects for each year from different models are computed to show their relationships. The variance components for school effects and their ratios to the overall variability in outcomes, which describe the schools' contribution to total variance, can

also be obtained directly from the MCMC algorithm. The school's contribution for each year and each subject obtained from different models are compared.

For the school persistence parameter, the analysis is based on the posterior mean and standard deviation. The assumption that the school effect persists into the students' future performance are examined according to the value of the estimated school persistence parameter. Whether the persistence is diminished or undiminished can also be found through the value of the persistence parameter and the overall model fit.

Results

Overall Model Fit

Table 5.1 summarizes the DIC value provided by all the models using three different data, respectively. It should be noted that the DIC values provided from different data are not comparable. Data 1 result shows that the general model yields the best overall model fit, which is indicated by the smallest DIC value. The PS model provides the second best overall model fit. The general and PS model are more complex than the other models, so this result suggests that the structure of Data 1 requires a complex model to obtain a good fit. The GS and CA model results are very close to each other and have the two largest DIC values. Data 2 yield relatively similar pattern to the Data 1 results. That is, for Data 2, the best model fit is also provided by the general model, which is followed by the PS model. And the GS and CA model perform the worst compared to the other models, but they two give the close results. There are also differences existing between Data 1 and Data 2 results. For Data 1, the CA model performs better than the GS model, and the LA model performs better than the CC model. However, for Data 2, the relationship between the CA and GS models or between the LA and CC models changes - the GS model

performs better than the CA model, and the CC model performs better than the LA model. The pattern shown by the Data 3 results is different from that shown by the Data 1 and 2. Although the general and PS model are still the best ones, the other four models give relatively close DIC values. That is, for Data 3, the disadvantage of using the GS and CA model is not that apparent compared to the Data 1 and 2.

Table 5.2 DIC Obtained from All the Models Using the Real Data

	Models					
	General	GS	CA	CC	LA	PS
Data 1	7977	8313	8298	8204	8107	8011
Data 2	7662	7842	7877	7738	7766	7695
Data 3	7842	8109	8224	8211	8143	7992

Fixed Effect Estimation

For the real data study, the fixed effect includes the overall mean for all the models, and one student level covariate – SES for all the models except the CC and LA model. Table 5.2 shows the overall mean estimates and their posterior standard deviations for each year from all the models. All the posterior standard deviations are around 5, which indicates that the precision of the overall mean estimates. It should be noted that the overall mean for the GS and CA model is actually the average growth from Year 1 to Year 2 and from Year 2 to Year 3. Comparing across the three data, we can find that when the same model being used the Data 1 has the highest overall mean estimate whereas the Data 2 has the lowest one. This is in accordance with our expectation since the Data 1 only has a small portion of disadvantaged students whereas the Data 2 has a large portion. Comparing across three years, we can find that the overall mean estimates increase over years. However, the three data show different rates of growth. The gain

from year 1 to year 2 is approximately 33 for Data 1, 20 for Data 2 and 27 for Data 3. The gain from year 2 to year 3 is approximately 28 for Data 1, 15 for Data 2 and 26 for Data 3. This result supports the assumption that the advantaged students not only have higher mean scores, but also have higher gains over years. When using the same data, no significant difference can be observed for the overall mean estimates from different models.

Table 5.3 Estimated Overall Mean for Each Year from Different Models Using the Real Data

		Estimates			SDs		
Data	Fitted	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Data 1	General	478.8	511.2	539.5	4.7	5.1	5.1
	GS	--	33.9	28.1	--	5.5	5.6
	CA	--	34.5	27.9	--	5.4	5.8
	CC	474.0	506.1	535.4	4.9	5.3	5.5
	LA	473.5	507.2	533.9	5.1	5.4	5.7
	PS	480.2	513.0	541.2	4.7	5.2	5.3
Data 2	General	464.3	486.4	500.2	4.8	5.3	5.2
	GS	--	11.5	15.4	--	5.6	5.5
	CA	--	10.1	14.3	--	5.3	5.6
	CC	460.0	481.2	493.7	5.1	5.5	5.6
	LA	458.8	479.6	491.9	5.2	5.6	5.5
	PS	462.6	482.8	498.9	4.9	5.1	5.4
Data 3	General	477.5	507.6	532.7	4.5	5.0	5.1
	GS	--	28.2	26.9	--	5.3	5.3
	CA	--	27.3	26.1	--	5.5	5.8
	CC	468.9	495.8	518.4	4.6	5.6	5.4
	LA	467.0	496.2	519.9	5.0	5.2	5.5
	PS	476.2	506.3	531.9	4.9	5.3	5.1

Table 5.4 Estimated Coefficients for SES for Each Year
from Different Models Using the Real Data

		Estimates			SDs		
Data	Fitted	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Data 1	General	29.9	28.7	33.4	4.6	4.4	4.3
	GS	--	30.1	31.9	--	5.3	5.2
	CA	--	31.2	34.6	--	6.4	5.2
	CC	--	--	--	--	--	--
	LA	--	--	--	--	--	--
	PS	29.4	27.9	33.9	5.3	5.4	6.3
Data 2	General	21.5	20.2	25.9	4.2	4.7	5.2
	GS	--	22.4	24.3	--	5.8	5.2
	CA	--	21.0	26.2	--	6.1	6.4
	CC	--	--	--	--	--	--
	LA	--	--	--	--	--	--
	PS	22.3	19.1	25.1	4.3	5.2	5.2
Data 3	General	23.7	23.4	26.8	4.4	5.8	5.1
	GS	--	22.1	24.4	--	5.5	5.4
	CA	--	23.9	27.6	--	6.1	5.4
	CC	--	--	--	--	--	--
	LA	--	--	--	--	--	--
	PS	23.0	22.2	26.4	5.6	5.2	5.6

The estimated coefficients for the SES variable for each year from different models are shown in Table 5.3. Although he impact of the SES variable changes over years and also changes across different data, all the coefficients are positive and statistically significant. Therefore, we can conclude that the advantaged students perform better than the disadvantaged students.

The Correlation between the Estimated Teacher Effects from Different Models

The accuracy of the school effects estimates cannot be evaluated for the real data. Therefore, in this section, the investigation focuses on the interrelationship among the school effects estimates from different models. Tables 5.4-5.6 report the pair-wise correlation between the estimated school effects from all the models for three data, respectively. The correlation ranges from 0.70 to 0.95 across three data for three years. Within each data, three years results are relatively close except for the correlation between the general and the CC model. For the general-CC pair, the correlation is much lower in Year 1 than in Year 2 and 3. This is true for all three data. The reason for this pattern remains unclear at this moment. For Data 1, the general and PS model give the highest correlation. On average, the general model has the highest correlation with the other models. This again proves that the general model is more reliable when the correct model is unknown. The GS and CA model have a relatively high correlation with each other (around 0.90), whereas they have much lower correlations with the CC model and the LA model (around 0.70). Data 1 presents the similar results with Data 2. However, Data 3 presents differences in the correlation of the PS-CC pair and the correlation of the PS-LA pair. Switching from Data 1 to Data 3, the PS-CC and PS-LA correlations increase to 0.88. This pattern shown by the real data is consistent with the pattern shown by the simulated data, although the latter is more apparent than the former. The simulated data result is more apparent might be because the data were generated using the general model, but for the real data the true underlying data structure is unknown. Therefore, again, it is natural to believe that the impact of the inclusion of the covariates depends on how the characteristics described by the covariates distribute among the sample.

Table 5.5 Pair-Wise Correlation between the Estimated School Effects from Different Models Using the Real Data (Data 1)

		General	GS	CA	CC	LA	PS
Year 1	General	1.00	--	--	0.87	0.92	0.94
	GS		--	--	--	--	--
	CA			--	--	--	--
	CC				1.00	0.92	0.80
	LA					1.00	0.82
	PS						1.00
Year 2	General	1.00	0.78	0.78	0.91	0.91	0.94
	GS		1.00	0.86	0.72	0.72	0.78
	CA			1.00	0.71	0.73	0.80
	CC				1.00	0.92	0.82
	LA					1.00	0.83
	PS						1.00
Year 3	General	1.00	0.82	0.79	0.92	0.91	0.94
	GS		1.00	0.93	0.71	0.71	0.82
	CA			1.00	0.72	0.74	0.82
	CC				1.00	0.91	0.83
	LA					1.00	0.82
	PS						1.00

Table 5.6 Pair-Wise Correlation between the Estimated School Effects from Different Models Using the Real Data (Data 2)

		General	GS	CA	CC	LA	PS
Year 1	General	1.00	--	--	0.87	0.91	0.94
	GS		--	--	--	--	--
	CA			--	--	--	--
	CC				1.00	0.90	0.78
	LA					1.00	0.83
	PS						1.00
Year 2	General	1.00	0.76	0.77	0.92	0.91	0.94
	GS		1.00	0.84	0.71	0.72	0.82
	CA			1.00	0.72	0.71	0.82
	CC				1.00	0.93	0.83
	LA					1.00	0.84
	PS						1.00
Year 3	General	1.00	0.82	0.81	0.91	0.90	0.94
	GS		1.00	0.92	0.72	0.70	0.84
	CA			1.00	0.72	0.72	0.84
	CC				1.00	0.91	0.84
	LA					1.00	0.82
	PS						1.00

Table 5.7 Pair-Wise Correlation between the Estimated School Effects from Different Models Using the Real Data (Data 3)

		General	GS	CA	CC	LA	PS
Year 1	General	1.00	--	--	0.83	0.91	0.95
	GS		--	--	--	--	--
	CA			--	--	--	--
	CC				1.00	0.91	0.90
	LA					1.00	0.90
	PS						1.00
Year 2	General	1.00	0.75	0.81	0.90	0.91	0.93
	GS		1.00	0.92	0.71	0.72	0.83
	CA			1.00	0.72	0.71	0.84
	CC				1.00	0.93	0.89
	LA					1.00	0.89
	PS						1.00
Year 3	General	1.00	0.78	0.82	0.91	0.90	0.95
	GS		1.00	0.89	0.70	0.72	0.83
	CA			1.00	0.72	0.73	0.81
	CC				1.00	0.90	0.88
	LA					1.00	0.88
	PS						1.00

School Variance Components Estimation

The school variance components estimate is another measure of the school effect estimation. As mentioned above, it is impossible to evaluate the accuracy of the estimates for the real data. Therefore, only the similarities and differences among all models from three data will be discussed. The school variance components estimates obtain from different data vary. They range from 6.5 to 15.2 for Data 1, from 9.9 to 18.9 for Data 2, and from 11.8 to 16.9 for Data 3. Moreover, comparing across three data, the estimated school variance components show different trends over three years. For Data 1, the

estimates decrease from Year 1 to Year 2, whereas they increase from Year 2 to Year 3. For Data 2, the estimates decrease from Year 1 to Year 3. However, for Data 3, the estimates from different models show different trends and the pattern of the trends is not quite clear. Next, we will examine the interrelationship among all the

Table 5.8 The Estimated School Variance Components for Each Year from Different Models Using the Real Data

		Estimates			SDs		
Data	Fitted	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Data 1	General	14.1	9.4	13.5	2.0	1.9	1.9
	GS	--	12.1	14.5	--	2.1	2.5
	CA	--	13.0	15.2	--	2.4	2.3
	CC	10.3	6.9	11.4	--	--	--
	LA	9.3	6.5	10.8	--	--	--
	PS	14.4	10.8	12.9	2.3	2.2	2.1
Data 2	General	13.9	13.2	11.6	2.1	2.1	2.2
	GS	--	18.9	15.7	--	2.2	2.4
	CA	--	17.4	16.0	--	2.3	2.5
	CC	11.4	10.2	9.9	--	--	--
	LA	12.8	11.6	10.4	--	--	--
	PS	14.4	13.1	10.9	2.2	2.2	2.2
Data 3	General	14.1	15.0	16.2	2.2	2.3	2.2
	GS	--	14.3	15.1	--	2.4	2.4
	CA	--	14.1	15.4	--	2.5	2.3
	CC	12.3	12.6	11.8	--	--	--
	LA	12.0	11.9	13.0	--	--	--
	PS	14.0	14.1	16.9	2.4	2.2	2.2

models within the same data. For Data 1, the estimates obtained from the general and the PS model are very close, those obtained from the GS and the CA model are close to each other and higher than the general model estimates, and those obtained from the CC and LA model are close to each other and lower than the general model estimates. The

interrelationships among all the models remain the same for Data 2. For Data 3, compare to the Data 1 and 2, the estimates obtained from the GS and the CA model are closer to the general model estimates with other patterns remaining the same. The changes occur to the GS and CA model for analyzing Data 3 allow us to relate the impact of explicitly modeling the intra-student correlation on the school variance components estimation to the structure of the data. We infer that ignoring the intra-student correlation (as the GS and CA model do) does not strongly affect the school variance components estimation when the students are heterogeneously grouped.

School Effect Persistence Estimation

All of the estimated school effect persistence parameters shown in Table 5.8 are larger than 0 and smaller than 0.5. This range is consistent with those reported in other studies using different empirical data. And this means that the previous years' teacher effects persist into the students' future achievement, although the persistence diminished over years. The general and PS model estimates are different but very close to each other. Over three years, the trends of the estimates show differences across three data. For Data

Table 5.9 The Estimated School Effect Persistence Parameters from Different Models Using the Real Data

		Estimates			SDs		
Data	Fitted	ϕ_{21}	ϕ_{31}	ϕ_{32}	ϕ_{21}	ϕ_{31}	ϕ_{32}
Data 1	General	0.21	0.15	0.32	0.04	0.04	0.03
	PS	0.25	0.20	0.26	0.04	0.05	0.04
Data 2	General	0.33	0.34	0.17	0.04	0.03	0.05
	PS	0.32	0.31	0.20	0.05	0.04	0.06
Data 3	General	0.16	0.25	0.28	0.05	0.04	0.04
	PS	0.21	0.24	0.32	0.06	0.04	0.05

1, the lowest estimates obtained in ϕ_{31} , whereas for Data 2 and 3, ϕ_{32} has the lowest estimates.

Random Student Effects Estimation

The estimated student own random effect components from the general and CC model are presented in Table 5.9. All the student effect estimates are significantly larger than 0, which supports our assumption on the existence of the student's own random effect. The estimates show the widest range in Data 1, which is from 4.2 to 7.5. The estimates obtained from the general and the CC model do not have significant differences except under three conditions-- the Year 1 result in Data 1 and Data 3 and Year 3 result in Data 2. Furthermore, no apparent pattern can be observed in terms of the changes of the estimates over years. For example, for Data 1, the general model estimates decrease from Year 1 to Year 3, but the CC model estimates increase. However, this pattern cannot be observed for Data 2 or Data 3.

Table 5.10 The Estimated Student Effect Component for Each Year from Different Models Using the Real Data

		Estimates			SDs		
Data	Fitted	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Data 1	General	7.5	5.1	4.2	2.5	2.7	2.8
	CC	4.3	4.9	5.1	2.8	3.1	2.9
Data 2	General	5.2	5.9	4.4	2.6	2.9	2.5
	CC	5.9	5.4	6.1	2.9	3.0	2.8
Data 3	General	6.3	5.0	5.9	2.4	2.6	2.7
	CC	4.1	6.2	5.1	2.8	2.7	3.0

Estimation of the Schools' Contribution to Total Variance

Table 5.10 shows the schools' contribution to total variance using three different data. We can observe that the variability of the estimates is higher than that of the school variance components estimates shown in Table 5.7. The estimates range from 3.5 to 21.4 for Data 1, from 8.1 to 21.0 for Data 2, from 10.1 to 20.7 for Data 3. Moreover, comparing across three data, the estimates show different trends over three years. For

Table 5.11 The Estimated Schools' Contribution to total variance for Each Year from Different Models Using the Real Data (%)

Data	Fitted	Year 1	Year 2	Year 3
Data 1	General	16.3	6.6	13.2
	GS	--	10.8	15.6
	CA	--	11.6	15.8
	CC	9.0	3.8	9.7
	LA	7.5	3.5	8.9
	PS	17.2	8.9	12.3
Data 2	General	14.3	14.2	9.6
	GS	--	21.7	8.8
	CA	--	15.1	12.0
	CC	11.7	8.7	8.1
	LA	12.4	11.0	8.9
	PS	16.5	12.3	8.3
Data 3	General	16.3	16.7	19.1
	GS	--	18.8	17.2
	CA	--	18.1	20.5
	CC	12.4	11.8	10.1
	LA	11.8	10.5	12.3
	PS	16.1	14.8	20.7

Data 1, the estimates decrease from Year 1 to Year 2, whereas they increase from Year 2 to Year 3. For Data 2, the estimates decrease from Year 1 to Year 3. However, for Data 3, the estimates from different models show different trends. This is the same pattern as

shown by the school variance components. In addition, the interrelationships among all the models observed from the schools' total contribution estimates are also the same with that observed from the school variance components estimates. That is, for Data 1 and 2, the estimates obtained from the general and the PS model are very close, those obtained from the GS and the CA model are close to each other and higher than the general model estimates, and those obtained from the CC and LA model are close to each other and lower than the general model estimates. For Data 3, compare to Data 1 and 2, the estimates obtained from the GS and the CA model are closer to the general model estimates.

CHAPTER 6

DISCUSSION AND CONCLUSION

Under NCLB, there is pressure to provide evidence to support the adequacy of teachers and schools in regards to student learning. VAM is being used as a tool to help illuminate which variables are in fact contributing to student learning, by isolating related factors, such as teacher and school effects. Although many researchers that have used VAM have shown promising results, additional research is needed in this area given the fact that mistakes in model misclassifications may have significant impact on teachers and schools, more research is needed. This study reviews several VAM approaches that are currently being implemented or reviewed for accountability purposes. Similar to McCaffrey et al. (2004), we intend to investigate the validity and reliability of several VAMs, by providing a general VAM framework and applying both the general and reduced models to the simulated and real data and then comparing the differences and similarities, given each model's basic assumptions. Compared to the general model proposed by McCaffrey et al., the general model proposed in this study is definitely more complex, in both formulation and estimation, in its attempt to explicitly parameterize and estimate the teacher effect persistence that has been proved to be necessary in describing the empirical data. In addition to proposing a new general model, an accompanying MCMC code for parameter estimation is also developed for this work.

The simulation study shows that the MCMC algorithm developed under a Bayesian framework functions very well for estimating the parameters involved in both the general and the reduced models. The fixed effect parameters can be accurately estimated using all the different models for generated data with different structures even when the model

specification does not match the underlying assumption of the data structure. The random effects investigated in the simulation study includes teacher effect and students' own random effect. The estimated teacher effects are acceptable, although their accuracy and precision are not ideal. As other studies have pointed out that VAMs are not capable of providing teacher effect estimation with any precision, the simulation study shows that the teacher effect estimates have relatively large biases. However, this does not affect the usage of the teacher effect estimates for accountability purposes. In practice, the magnitude of the teacher effect is not of the most importance. On the contrary, the rank-ordering teachers or identifying teachers at the extremes of the performance distribution is the objective of applying VAMs. The estimated teacher effects from both the general and the reduced models have high correlation with the generated true teacher effects. Meanwhile, the students' own random effects can be accurately estimated by the general and the CC model. Beyond the fixed and random effects, all the models can recover the teachers' contribution to total variance, which also depends on the quality of the residual error term estimation.

In addition to the feasibility of the general model, the relationship between the general model and the reduced model, and the relationship among all the reduced models are also investigated through the simulation study. The following summaries are based on the DIC values and the evaluation of the quality of the different estimates. First, the general model has the best performance in terms of the overall model fit when the data are generated using the general model. Even when the data are generated using the reduced models, the performance of the general model is just slightly worse than those of the correct models. Second, compared with all the other reduced models, the PS model

provides the closest results to those provided by the general model. This is in accordance with our expectation because the general model and the PS model have exactly the same underlying assumptions on the teacher effects, teacher effect persistence and residual error and the only difference between the general model and the PS model is the inclusion of the student's own random effect. Although the real data results support the existence of the student's own random effect, its magnitude and its contribution to the total variation of student's score are relatively small compared to that of the fixed effect and other random effects. This might be the reason why the advantage of the use of the general model is quite mild over the use of the PS model. In the future, a simulation study with stronger student's random effect and more empirical studies are needed to investigate the similarity and difference between the general and the PS model. The Third, the GS and CC model tend to provide relatively similar results to each other under various conditions and they have the most apparent differences with the general model, which is supported by the largest distances existing between their estimates and the general model estimates. One possible explanation for the similarity between the GS and CC model is that both of them include the student's previous year score into the fixed effect part and assume no intra-student correlation. Forth, the CA and LA model often provide similar results under some conditions. This might be because both of them do not incorporate any covariates and assume constant correlation across years within the same student.

The impact of the school composition on the model performance and on the interrelationship among models can also be observed from the simulation results. School A and B, which have unbalanced mix of the advantaged and disadvantaged students,

show the same picture of the model performance pattern in terms of the overall model fit and quality of the estimates. However, School C data, which has balanced mix of the advantaged and disadvantaged students, sometimes tells a different story. For example, the performances of the CC and LA model are noticeably better for analyzing the School C data than for the School A or B data when all of the data are generated using the general model. The performance of the general for estimating the data generated using the CC or LA model also improves. These improvements can be supported by higher correlation measured between the estimated and the true teacher effects. In addition, the correlation between the teacher effects estimates obtained from the different models shows that the PS-CC correlation and PS-LA correlation apparently increase when switching from the School A or B data to the School C data. This result allows us to infer that the impact of the covariates on the teacher effect estimation is associated with the school composition because the most salient feature of the CC and LA model is that both of them exclude the covariates. As mentioned in Chapter 2, there have been hot debates on controlling for student background in value-added assessments of teachers. Some researchers, given what they know about the relationship of demographic characteristics of persons to their educational attainment, believe it is unreasonable to think that covariates would have no relationship at all to outcomes. However, according to our simulation results, this is true under certain conditions.

The real data study shows that, to the extent that it can be verified, the analysis of actual students' outcomes from a large scale statewide testing provides very similar results to those obtained in the simulation study. Hence, the real data, which are of very complex structure, requires a complex model similar to the proposed general model to be

analyzed and interpreted appropriately. Some differences from the simulation results are encountered in real data analysis. One possible explanation is that the measurement errors associated with the observed variables are inevitable in practice. It should also be noted that the school effects investigated in the real data analysis are not necessarily causal effects of schools. Rather, they account for unexplained heterogeneity at the school level. All the discussed models indicate that school effects account for a significant proportion of the variability in students' growth in achievement scores, although the proportions among different models vary in magnitude. The magnitude of school effects should be interpreted with great caution.

The teacher effect persistence (school effect persistence in the real data analysis) is another issue that has received great attention. However, there is still no universal agreement on to what degree the teacher effect persists into the future among researchers. Sanders and his colleague believe the high rates of persistence of teacher effects over several years. McCaffrey et al. (2004) criticized their claims and provided more modest persistence effect estimates using models with less stringent assumptions. One of the most important findings of our real data analysis is that the persistence parameters imply long-term persistence of past years teachers' effects or schools' effects decay in the strength over time. Thus, the general model and PS model assumption on the persistence parameter fits better for the data than the GS and LA model, which assume that teachers' (or schools') effects from the past years persist undiminished into the future. All estimates are positive but substantially smaller than one. This finding is consistent with the empirical result presented in McCaffrey et al. (2004) obtained with different data. This finding can also shed light on the practical meaning of teacher (or school) effects - it

suggests that the effects of poor teaching should be more remediable than it has been claimed.

A common concern and drawback discussed about the use of more complex model like our general model is the computational challenge. However, as proposed, the MCMC algorithm in a Bayesian framework can successfully estimate all the involved parameters. The most important property of the MCMC algorithm is that sampling the joint posterior distribution all the parameters can be realized by repeatedly sampling from the conditional posterior distributions of one parameter as related group of parameters given the data and current values of all other parameters. This makes it well suited to dealing with models with complex relational structures. For example, the estimation of the persistence parameters can be treated as the estimation of the unknown regression coefficients on known predictors conditional on the random effects. According to Lockwood et al. (2007), conditioning on random effects reduces the complex covariance matrices to simple, computationally tractable block diagonal forms. Moreover, using a program written in Ox (Doornik, 2002) to implement the general model and analyze 16 teachers' effect and 400 students' scores takes a 2 GHz machine only fifteen minutes to run 10000 iterations. And more importantly, MCMC remains open and viable because its flexibility and ease of implementation allow us to develop more complex problems in future research.

There are a few important limitations to both the simulation and real data study. First, the simulation data is designed to have no incomplete student scores and no missing teacher-student linkage. In addition, the real data is also intentionally selected without any missing records from a large-scale statewide testing data. However, in practice, the

missing data problem is inevitable. For example, in the entire data set, from which the real data analyzed in the work have been obtained, actually, only 15% of the students have complete testing scores over the 3 years. In addition to modeling student data, the missing teacher-student linkage is another serious problem. Students and teachers transfer during the years of testing. For the incomplete student scores, the Bayesian augmentation method allows us to estimate the missing value as the unknown parameter. But dealing with the missing teacher-student linkage can only be determined by positing a missing mechanism. Lockwood et al. (2007) implemented three procedures for treating the missing link information for three different missing pattern assumptions, respectively. They also analyzed an empirical testing data to investigate the sensitivity of the value-added measures to the missing pattern assumptions using the PS model. In the future study, we can extend their investigation to all the VAMs including our general model and use well-designed simulation study to examine the different missing patterns.

Second, in the simulation study, the assignment of students to teachers is random conditional on the student SES variable. The same assumption is made for the real data analysis. However, in reality, there is little reason to think that this is an adequate characterization of classroom assignments. For example, the principles or parents have a great deal of information beyond the prior test score that can affect the classroom assignments. Rothstein (2009) quantified the biases in estimates of teacher effect from several value-added models under varying assumptions about the assignment process and pointed out that even the best feasible value-added models may be substantially biased with the magnitude of the bias depending on the amount of information used in the assignment process. Therefore, a further investigation on the performance of the

proposed general model, especially, the teacher effect estimation given more complex assignment assumptions should be conducted.

Third, the only covariate involved in both the simulation and real data study is the SES variable. This is because the SES variable is the most debatable covariate, which is believed to be confounded with the teacher or school effect. However, researchers have shown that gender, ethnicity and some other indicators are also important predictors of students' future performance. Future work should include studies that compare the models when more covariates are included.

Fourth, in the simulation study, due to the time and resource limitation, only one dataset were generated for each condition. Future study should generate at least 100 dataset for each condition to make the findings more reliable.

REFERENCES

Ballou, D. (2002). Sizing up test scores. Education Next, 2(2), 10-15.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for students background in value added assessment of teachers. Journal of Educational and Behavioral Statistics, 29(1), 37-66.

Barton, P.E. (2004). Why does the gap persist? Educational Leadership, 62, 8-13.

Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. Princeton, NJ: Educational Testing Service. Retrieved May 9, 2008, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Browne, W. J., Draper, D., Goldstein, H., & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. Computational Statistics and Data Analysis, 39: 203-225

Bryk, A., Raudenbush, S., & Congdon, R. (1996). HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs. Chicago: Scientific Software International, Inc.

Carey, K. (2004). The real value of teachers: Using new information about teacher effectiveness to close the achievement gap. Thinking K-16, 8(1), 1-42.

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1996). Analysis of longitudinal data. New York: Oxford University Press.

Doran, H. C., & Lockwood, J. R. (2006). Fitting Value-Added models in R. Journal of Educational and Behavioral Statistics, 31(2), 205-230.

Drury, D. & Doran, H. (2003). The Value of Value-Added Analysis. NSBA Policy Research Brief. 3(1), 25-42

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7, 457-472.

Goldhaber, D. and Anthony, E. (2004). Can Teacher Quality Be Effectively Assessed?2004, University of Washington

Goldschmidt, P. K. Choi, F. Martinez (2003). Using Hierarchical Growth Models to Monitor School Performance Over Time: Comparing NCE to Scale Score Results, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), U.S. Department of Education, Office of Educational Research and Improvement

Gooden, M. A., & Nowlin, T. Y. (2006), The Achievement Gap and the No Child Left Behind Act: Is there a Connection. Advances in Education and Administration, 9, 231-247

Hershberg, T., Simon, VA, & Lea-Kruger, B. (2004). The revelations of value-added. School Administrator, 61(11), 10-12

Hibpshman, T.L. (2004a). Review of Evaluating Value-Added models for Teacher Accountability. Kentucky Education Professional Standards Board.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. Educational evaluation and policy analysis, 25(3), 287-298.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). Journal of the Royal Statistical Society, Series B: Statistical Methodology, 34, 1-41.

Lockwood J.R., Schervish M.J., Gurian P.L., & Small M.J. (2004), Analysis of contaminant co-occurrence in community water systems, Journal of the American Statistical Association, 99(465), 26-45

Lockwood, L.R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F.(2007). The sensitivity of Value-Added teacher effect estimates to different mathematics achievement measures. Journal of Educational Measurement, 44, 47-67.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability, MG-158-EDU. Santa Monica, CA: RAND.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29(1), 67-101.

McCaffrey, D. F., Lockwood, J. R., Mariano, L. T., and Setodji, C. (2005). Challenges for value added assessment of teacher effects. In R. Lissitz (Ed.), Value added models in education: Theory and practice (pp. 272–297). Maple Grove, MN: JAM Press.

Meyer, R. (1997). Value-added indicators of school performance, Economics of Education Review, 16, 183-301.

Rasbash J. & Browne W. J. (2002). Non-Hierarchical Multilevel Models. To appear in De Leeuw, J. and Kreft, I.G.G. (Eds.), Handbook of Quantitative Multilevel Analysis.

Raudenbush, S., & Bryk, A. (1986). A hierarchical model for studying school effects. Sociology of Education, 59, 1-17.

Raudenbush, S., & Bryk, A. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S.W. (2004). What are value-added models estimating and what does this imply for statistical practice? Journal of Educational and Behavioral Statistics, 29(1), 121-129.

Raudenbush, S.W., & Willms, J.D. (1995). The estimation of school effects. Journal of Educational andBehavioral Statistics, 20(4), 307-335.

Rothestein, J. (2009). Student sorting and bias in value-added estimation: selection on observables and unobservables. Education Finance and Policy, 4(4), 537-571.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value added assessment in education. Journal of educational and behavioral statistics, 29(1), 103-116.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association, 95, 573-585.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. Journal of Educational and Behavioral Statistics, 29(1), 103-116.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. Teachers College Record, 104 (8), 1525-1567.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. Journal of Personnel Evaluation in Education, 8, 299-311.

Sanders, W. L., Saxton, A., & Horn, S. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Shkolnik, J., Hikawa, H., Suttorp, M., Lockwood, J., Stecher, B., & Bohrnstedt, G. (2002). Appendix D: The relationship between teacher characteristics and student achievement in reduced-size classes: A study of 6 California districts. In G. W. Bohrnstedt, B. M. Stecher (Eds.), What we have learned about class size reduction in California Technical Appendix. Palo Alto, CA: American Institutes for Research.

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. Journal of Educational and Behavioral Statistics, 29(1), 11-36.

Wanker, W.P., & Christie, K. (2005). State Implementation of the No Child Left Behind Act. Peabody Journal of Education, 80 (2), 57-72.

Curriculum Vita

Yuan Hong

EDUCATION

Ph.D., Education: Educational Statistics, Measurement and Evaluation
Rutgers University, New Brunswick, NJ, expected January 2010
M.S., Statistics
Renmin University of China, Beijing, P.R. China, June 2005
B.A., Statistics
Renmin University of China, Beijing, P.R. China, June 2002

EXPERIENCE

2007~2009	*Principle Investigator*, evaluating school and teacher effect using general value-added modeling framework, project funded by CTB/McGraw-Hill
2009	*Guest Lecture*, Regression Analysis, Rutgers University
2007~2008	*Principle Investigator*, examining the differential impact of test format on group performance, project funded by the College Board
2008	*Guest Lecture*, Regression Analysis, Rutgers University
2007	*Research Intern*, CTB/McGraw-Hill

PUBLICATION

de la Torre, J., & Hong, Y. (In press). Parameter estimation with small sample size: A higher-order IRT approach. *Applied Psychological Measurement.*

de la Torre, J., Hong, Y., & Deng, W. (In press). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement.*