COMPARISON OF VALUE-ADDED MODELS FOR SCHOOL RANKING AND CLASSIFICATION: A MONTE CARLO STUDY

A Dissertation

by

ZHONGMIAO WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2006

Major Subject: Educational Psychology

COMPARISON OF VALUE-ADDED MODELS FOR SCHOOL RANKING AND CLASSIFICATION: A MONTE CARLO STUDY

A Dissertation

by

ZHONGMIAO WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	B. Thompson
	V. L. Willson
Committee Members,	O. Kwok
	M. Speed
Head of Department,	M. Bentz

December 2006

Subject Major: Educational Psychology

ABSTRACT

Comparison of Value-Added Models for School Ranking and Classification: A Monte Carlo Study. (December 2006)

Zhongmiao Wang, B.S., Beijing Normal University;
M.S., Beijing Normal University

Co-Chairs of Advisory Committee: Dr. B. Thompson
Dr. V. L. Willson

A “Value-Added” definition of school effectiveness calls for the evaluation of schools based on the unique contribution of schools to individual student academic growth. The estimates of value-added school effectiveness are usually used for ranking and classifying schools. The current simulation study examined and compared the validity of school effectiveness estimates in four statistical models for school ranking and classification. The simulation study was conducted under two sample size conditions and the situations typical in school effectiveness research. The Conditional Cross-Classified Model (CCCM) was used to simulate data. The findings indicated that the gain score model adjusting for students’ test scores at the end of kindergarten (i. e., prior entering to an elementary school) (Gain_kindergarten) could validly rank and classify schools. Other models, including the gain score model adjusting for students’ test scores at the end of Grade 4 (i. e., one year before estimating the school effectiveness in Grade 5) (Gain_grade4), the Unconditional Cross-Classified Model (UCCM), and the Layered Mixed Effect Model (LMEM), could not validly rank or classify schools. The failure of the UCCM model in school ranking and classification indicated that ignoring covariates would distort school rankings and classifications if no other analytical remedies were applied. The failure of the LMEM model in school ranking and classification indicated that estimation of correlations among repeated measures could not alleviate the damage caused by the omitted covariates. The failure of the Gain_grade4 model cautioned against adjustment using the test scores of the previous year. The success of the Gain_kindergarten model indicated that under some circumstances, it was possible to achieve valid school rankings and classifications with only two time points of data.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

TABLE Page

TABLE Page

CHAPTER I

INTRODUCTION

“Value-Added” Models (VAM) of school effectiveness are different statistical models that estimate value-added school effectiveness. The current study compared the performance of different VAM in school ranking and classification. These models use student achievement test scores to judge the quality of schools, which reflect the popular practice of school accountability around the United States. The new federal education law, No Child Left Behind Act, calls for more emphasis on student academic achievement, and has motivated many states to design their accountability systems to evaluate schools based on their contributions to students’ academic achievement.

The VAM investigated in the current study reflect the latest thinking about school effectiveness. The recent development in the definition of school effectiveness emphasizes three points. One is that school effectiveness should reflect the unique effect of school education on individual student achievement. The second is that school effectiveness should reflect the impact of schools on students’ achievement growth over time, but not achievement status at a single time point (Teddlie & Reynolds, 2000). The third is that school effectiveness should reflect the impact of schools on individual students, but not on the aggregate level of all the students.

____

This dissertation follows the style of *Educational and Psychological Measurement*.

The definition of value-added school effectiveness combines the three points, which can be defined as the unique effect of each school on individual students' achievement growth over time (Doran, 2003; Mayer, 1997; Rowan, Correnti, & Miller, 2002). Although students' achievement can be presented in different areas, such as academic achievement, personality maturity, arts achievement, most of the current school accountability systems emphasize student academic achievement, and hold schools accountable for students' scores on academic achievement tests. Thus, value-added school effectiveness usually focuses on school unique contribution to individual students' growth in academic achievement.

In terms of the unique contribution of schools, Raudenbush and Willms (1995) defined two types of school effectiveness: Type A and Type B. Type A effect is the effect of both school context and school practice on student achievement. Type B effect is the effect of only school practice on student achievement. School context variables are school characteristics that are out of control of a school, such as school location, school demographic composition, and aggregated student characteristics, for example, the school mean intake test score, the Socio Economic Status (SES) of student body. School practice variables describe some instruction and administration policies that can be controlled by a school (Raudenbush & Willms, 1995). The Type B school effect is of greatest interest to school administrators and policy makers because this effect addresses how much impact educational intervention programs or policies have on student learning, and it puts schools under evaluation for what they can control.

In common practice of value-added assessment of schools, the policy makers use school effectiveness estimates for two purposes. One is to rank schools; and the other is to identify exceptional schools either for providing special aid or for accumulating successful experience. Of greatest concern is whether a statistical model can provide valid estimates of value-added school effectiveness to fulfill the two purposes. The current study investigated the validity of estimates of value-added school effectiveness in different statistical models for school ranking and classification.

The three points in the value-added definition of school effectiveness and the concepts of the Type A effect and Type B effect imply some methodology requirements in estimating value-added school effectiveness. The first point requires adjustments for the factors that are out of control of schools; the second point requires modeling of students' achievement growth; and the third point requires the use of multilevel models. Different Value-Added Models respond to the three requirements with different strategies. By comparing these models, we can determine which strategy works well in satisfying a certain requirement, and which strategy does not perform as intended.

Questions about unique effect of school practice on student learning imply adjustment for covariates. The randomized experimental design is the best research method that can adjust for the covariates and answer questions about causal effect (Braun, 2005; Raudenbush, 2005; Rubin, 2004, Rubin, Stuart & Zanutto, 2004). The best estimate of Type B school effect results from a nested random block experimental design. Let us imagine that the practice in each school is a treatment. First, J schools having identical context are assigned to different treatment levels that vary in terms of practice. Next, blocks of J students of identical background and aptitude are assigned to the J schools. The average performance of the J students in the same block represents the average effectiveness of the J schools' practice. The discrepancy between a student's performance and the average performance of the J students in the same block is the Type B effect (Raudenbush & Willms, 1995).

However, the random assignment of students to schools and schools to different kinds of practice is almost impossible in education. Without the benefit of random assignment, researchers try to use statistical models to isolate the school practice effect from the effect of other confounding variables. In concept, the confounding variables that need to be controlled in estimating Type B school effectiveness include student background variables and school context variables. However, because of multicollinearity among the student background variables and the school context variables, specification of these covariates in statistical models may still result in biased estimates of value-added school effectiveness (Teddlie & Reynolds, 2000), and invalid school rankings and classifications. Furthermore, selection and measurement of the covariates in education also present many problems.

Considering the problems of specifying covariates in the models, researchers try to use other strategies to adjust for the effects of the covariates without specifying them in the models. Sanders and his research team in University of Tennessee claimed that the influence of other exogenous factors can be filtered out without directly measuring and specifying these factors (Sanders, 2000; Sanders, Saxon, & Horn, 1997). Sanders said: "By taking advantage of longitudinal data, each student serves as his/her own control. In other words, each student can be regarded as a blocking factor. By blocking for each student, many of the exogenous influences most often cited as influencing academic progress - educational attainment of parents, socioeconomic level, race, and so on –

could be partitioned without having direct measures of each one" (Sanders, Saxon, & Horn, 1997, P.138).

This claim aroused intensive enthusiasm among school effectiveness researchers. Some research supported this idea (Casteel, 1994; Cook, 1985; Mclean & Sanders, 1984); while, some research did not support this idea (Tekwe et al., 2004); and some research found that this idea was valid under certain circumstances (Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Koretz, Louis & Hamilton, 2004). One intention of the current study was to examine this strategy of adjusting for the covariates without specifying the covariates in the model. In order to fulfill this intention, the author compared the school rankings and classifications based on the Layered Mixed Effect Model (LMEM) that adopts this strategy with the known true school rankings and classifications in the simulated data.

Covariates Modeled

In order to simplify the study, not all the associated covariates found in literature were considered. The current study included one student level covariate and one school level covariate. As far as the student level covariate, results of previous research indicated that student prior attainment explains most of the total variance in student achievement outcome. Because of multicollinearity between prior attainment and other student background variables, adjustment using only prior attainment achieved similar total variance reduction as adjustment using all the student background information. Thomas and Mortimore (1996) found that when prior attainment was accounted for, the total variance reduced 58%, which was similar to that (63%) in the refined model in which all student background variables, including gender, ethnicity, parent education, Socio Economic Status, were adjusted. This result was consistent with other research (Cuttance, 1992; Gray, Jesson, & Sime, 1990; Goldstein et al., 1993; Sammons, Nuttall, Cuttance, & Thomas, 1995; Scheerens, 1992; Willms, 1992). Furthermore, when student prior attainment was accounted for, the difference between the schools with the highest and the lowest means reduced 70.6%, and the variance accounted for by school level factors reduced from 14% to 10%. These figures were similar to those in the refined model. Thus, the adjustment with other student background variables beyond prior attainment just marginally refined the results. Thus, based on the literature review, the simulation model that was used to generate student test scores in the current study adjusted for student prior attainment at the student level.

As far as the school context variable, previous studies suggested controlling for the effect of student body Socio Economic Status (SES). Thomas and Mortimore (1996) investigated the impact of SES of student body. They found that once student SES was adjusted at the student level, the impact of SES of student body could not be detected at the school level. However, if SES of individual student was not adjusted for at the student level, even if their prior attainment was adjusted, the SES of student body still had statistically significant impact on student achievement. In analysis of a school district data, Darandari (2004) found that the SES of student body was the proxy of other school context variables. Once SES was adjusted, other school context variables rarely had statistically significant effects on student achievement. Adjustment of student body SES also reflects the common practice in school effectiveness research (Cuttance, 1992; Fitz-Gibbon, 1997; Goldstein, 1997, 1984; Tedllier & Reynolds, 2000; Thomas, Sammons & Mortimore, 1997; Willms, 1986; Raudenbush & Willms, 1995). Thus, besides adjusting for student prior attainment at the student level, the correct model in the current study also adjusted for the SES of student body at the school level.

An interesting question about adjustment for prior attainment is which prior attainment should be used. The common practice is to adjust for test scores from the previous year, such as adjusting Grade 4 test scores when estimating Grade 5 school effectiveness. The practice is especially common in the research on elementary schools. This is because most states administer achievement tests from Grade 3. There typically are no tests for children at the end of kindergarten or at the beginning of Grade 1, which can be the points of entry to elementary schools. However, Sammons (1996) found that the use of baseline attainment or achievement data collected after a period of years in the same school was likely to lead to a reduction in the estimate of school effect. Cuttance (1985) also cautioned against the use of prior achievement as a control when prior scores are proximal to the point at which the school effects are measured. Preece (1989) commented on the potential problem of partialling out school effects in such cases. Sammons (1996) recommended adjustment using the test scores collected at the point of entry to a school. The test scores at the point of entry to a school were called as intake test scores in the current study. One intention of the current study was to examine the difference of the adjustments with the two kinds of prior attainment in school rankings and classifications. Therefore, the author compared the performance of two kinds of gain score models. One gain score model adjusted for the test scores collected at the end of kindergarten (i. e., the point of entry to an elementary school); the other gain score model adjusted for the test scores collected at the end of Grade 4 (i. e., one year before estimating the school effectiveness in Grade 5).

Another question about covariates adjustment is how seriously the school rankings and classifications will be hurt if the covariates are ignored. Many researchers have indicated that ignoring important covariates at one level will bias estimates of both the fixed effects and the random effects at all the levels. This is because the ignored covariates will be included in the residuals, which will cause correlation between the residuals and the predictors in the model or correlation between the residuals at different levels. This violates the independence assumption of multilevel level modeling (Darandari, 2004; Raudenbush & Bryk, 2002), which will cause inaccurate and statistically inefficient estimates of the fixed effects and the variance components. Because the Empirical Bayes (EB) estimates of regression coefficients for each school is determined by both the fixed effects and the variance components (Raudenbush & Bryk, 2002), thus the EB estimates of regression coefficients for each school and the associated residual will be biased too. However, it is still not clear whether the magnitude of bias is large enough to change school rankings and classifications. One intention of the current study was to investigate the effect of omitted covariates on school rankings and classifications. Therefore, the author compared the known true school rankings and classifications with the school rankings and classifications based on the Unconditional Cross-Classified Model (UCCM).

Longitudinal Models Versus Gain Score Models

In terms of the second point of the definition of value-added school effectiveness, many researchers suggested using longitudinal design with at least three years of data (Hill & Rowe, 1996; Mortimore, Sammons, Stroll, Lewix, & Ecob, 1988; Raudenbush 1989; Raudenbush & Bryk, 1989; Teddlie & Rehold, 2000). This is because school effectiveness is most likely to present over a long term, and the school effectiveness in the current year is influenced by effectiveness in the previous years (Sammons, Nuttall, Cuttance & Thomas 1995). On the other hand, some researchers noted the problems associated with longitudinal models. The biggest problem is that test scores across a wide span of grades may measure different knowledge and abilities, which may make the vertical equating of test scores across different grades invalid (Linn, 2005; Martineau, 2006). Furthermore, datasets of test scores across several years usually have more missing scores than the datasets of test scores involving only two years. One purpose of the current study was to investigate whether a gain score model with only two years of test scores could achieve similar school rankings and classifications as the more complicated longitudinal models. Therefore, the author compared the school rankings and classifications based on the two kinds of gain score models with the known true school rankings and classifications.

Research Questions

In summary, comparisons of the Value-Added Models for school rankings and classifications were conducted to answer six questions: (1) whether a gain score model adjusting for the kindergarden test scores (Gain_kindergarten) could recover the known true school rankings and classifications; (2) whether a gain score model adjusting for the Grade 4 test scores (Gain_grade4) could recover the known true school rankings and classifications; (3) whether the Unconditional Cross-Classified Model (UCCM) that ignored the covariates could recover the known true school rankings and classifications; (4) whether the LMEM model that estimated correlations among repeated measurements over time but not specified the covariates in the model could recover the true school rankings and classifications; (5) when a gain score model was used, whether adjustment using the Grade 4 test scores could achieve similar school rankings and classifications as the adjustment using the kindergarten test scores; (6) whether estimating the correlations among repeated measurements could alleviate the problem caused by omitted covariates in estimating school rankings and classifications?

Simulation Design

Monte Carlo simulation was used to answer the six research questions. CCCM was used to generate students' test scores from the end of Grade 1 to the end of Grade 5. The test scores at the end of Grade 1 were the intercepts of students' individual growth curves. In the CCCM, student test score at the end of kindergarten was the predictor of the intercept at the student level. Student body SES was the predictor of growth at the school level. The mathematical equation for the CCCM is:

$$Y_{tij} = \beta_{00} + \beta_{01} * kinder_{ij} + r_{0i} + u_{0j} \quad \text{(Intercepts of students' individual growth curves)}$$

$$+ (\beta_{10} + r_{1i}) * time \qquad \text{(students' natural growth rates)}$$

$$+ \sum_{t=1}^{t} \left(\gamma_{01} * SES_j + u_{tj} \right) \qquad \text{(School level effects)}$$

$$+e_{tij} \qquad \text{(residual at each grade)} \qquad (1.1)$$

where Y_{tij} is the test score of the ith student in the jth school at time t. For Grade 1, t = 0. β_{00} is the adjusted mean score of the typical students at the end of Grade 1. The typical students have the grand mean level on kindergarten test scores. β_{00} is also the grand mean of the intercepts of students' individual growth curves. β_{01} is the fixed effect of student kindergarten test score on the intercepts of the growth curves. γ_{01} is the fixed effect of student body SES. r_{0i} is the student random effect on the intercepts of the growth curves. The CCCM model assumes that each student has a natural growth curve which exists no matter whether or not the student attends a school. The natural growth curve is linear. β_{10} is the grand mean slope of the natural growth curves. r_{1i} is the student random effect on the slopes of the natural growth curves. u_{0j} is the school random effect on the intercepts, which represents the school selection effect (Ponisciak & Bryk, 2005). u_{tj} is the value-added school effectiveness of school j at time t. The schools are ranked and classified based on estimates of u_{tj}. e_{tij} is the measurement residual term for student i in school j at time t.

Value-added assessment with multilevel modeling has a set of requirements with respect to achievement tests. These requirements are that the achievement tests at each grade measure the same content and the test scores are vertically linked without linking errors. Because the current study did not investigate measurement problems in value-added assessment of school effectiveness, these measurement requirements were assumed to be satisfied in the current study.

The simulation was conducted under two different conditions. One condition was the Number of Schools (NS), which had two levels: 50 and 10. The other condition was the Number of Students per School (NSS), which also had two levels: 50 and 10. Thus, the simulation adopted a 2*2 design. The levels of the NS factor were decided partially based on the NCES 2003-2004 Public Elementary/Secondary Universe Survey Data and partially based on the typical number of schools in most school effectiveness research. The distribution of the number of elementary schools in a school district is extremely positively skewed in the United States (skewness = 31.78). Among the 13,479 school districts that contain regular elementary schools, 4 school districts have more than 300 regular elementary schools, 23 school districts have 100-300 regular elementary schools, 61 school districts have 50-100 regular elementary schools, 98 school districts have 30-50 regular elementary schools, and 98.65% of the school districts have less than 30 regular elementary schools. In school effectiveness research, 50 groups was a frequently occurring number, and 30 groups was mentioned as minimum (Mass & Hox, 2004). However, I used an even smaller number as the lowest level of the NS factor. This was done to enlarge the difference between the highest level and the lowest level of the NS factor, so that the influence of the NS factor would be more obvious if an effect really existed.

The levels of the NSS factor were selected based on literature and capacity of the computer available to the author. A size of 50 was chosen because 50 should be sufficient on the basis of literature (Mass & Hox, 2004). Although a group size of 30 is common in educational research (Mass & Hox, 2004), the author used 10 as the lowest level of the NSS factor in order to enlarge the difference between the two levels of the NSS factor. In a pilot study, the author found that increasing school size led to a dramatic increase of computing time and convergence problem. With 50 schools and 100 students per school, the estimation of the UCCM and the LMEM in SAS and R could not achieve convergence at all. Thus, I did not use a large school size, such as 100, although such school sizes are common in the elementary schools around the country.

Literature-Based Simulation Parameters

The current study used the CCCM model to generate the data. The CCCM included three fixed effects and five variance components. The three fixed effects were: (1) the effect of the kindergarten test scores on the intercepts which are the test scores at the end of Grade 1 (β_{01}), (2) the grand mean of students' natural growth rates (β_{10}), (3) the effect of student body SES (γ_{01}). The five variance components were: (1) the variances of measurement errors at each grade, which were equal across years (i.e., $\sigma^2_{e0}= \sigma^2_{e1}= \sigma^2_{e2}= \sigma^2_{e3}= \sigma^2_{e4}= \sigma^2_{e}$), (2) the variance of student random effect on the intercepts (i.e., σ^2_{r0i}), (3) the variance of school random effect on the intercepts (i.e., σ^2_{u0j}), (4) the variance of student random effect on the natural growth rate (i.e., σ^2_{r1i}), (5) the variances of school value-added effectiveness at each grade, which were constant across years (i.e., $\sigma^2_{u1j}= \sigma^2_{u2j}= \sigma^2_{u3j}= \sigma^2_{u4j}= \sigma^2_{uj}$) . Besides the parameter values in the CCCM model, four kinds of correlations that may influence the estimates of value-added school effectiveness were also considered. The four correlations were: (1) correlation between intercept and slope of student natural growth curve (i.e. r_{r0ir1i}), (2) correlation between student body SES and the aggregated kindergarten test score of each school, (3) correlations among school value-added effectiveness over time, (4) intraclass correlation of student kindergarten test scores.

The parameter values for the fixed effects, variance components, and the correlations were selected to reflect the findings in school effectiveness research, so that the study would have greater ecological validity. Therefore, the following parameters were used: student kindergarten test scores explained 58% of the total variance of test scores at Grade 1. School selection effect explained 10% of the total variance of test scores at Grade 1. Student random effect explained 22% of the total variance of test scores at Grade 1; measurement error explained 10% of the total variance of test scores at Grade 1 (Bosker & Witzier, 1995; Thomas & Mortimore, 1996).

In the current study, student kindergarten test scores were assumed to be standardized; and student test scores at Grade 1 were also assumed to be standardized. Given the variance accounted for by kindergarten test scores, the regression coefficient for the kindergarten test scores and the residual variances can be determined for generating student test scores at Grade 1.

According to the CCCM model, test scores at later years are the sum of student true scores at the previous year plus achievement growth plus error. In each year, the overall achievement growth of a student can be divided into two parts. One is the natural growth. This kind of growth may happen because of natural maturity or other environment influence except schools. CCCM assumes this part of growth is linear. The other part of growth is due to school. By attending a given school, a student may gain more beyond his natural growth in a year. In order to generate test scores at later grades, we need to separate individual students' natural growth from their growth due to school. Based on literature (Mortimore, et al., 1988; Raudenbush, 1989; Raudenbush & Bryke, 1989; Rowan, Correnti & Miller, 2002), the variance of growth due to school was set at

1.5 times of the variance of natural growth. In the current study, the distribution of individual student natural growth was adopted from the study of Ponisciak and Bryke (2005), which had a mean of 0.63 and a variance of 0.325. Thus, the variance of growth due to school was 0.487 (i.e., 0.325*1.5=0.487), and the total variance of growth was 0.812 (i.e., 0.325+0.487=0.812). The variance due to school was further divided into two parts. Based on the literature (Teddlie & Stringfield, 1993; Willms, 1987), 35% of the between school variance of growth (i.e. 0.487*0.35 = 0.1694) was explained by the difference in school SES. Given that school SES had a standardized normal distribution, and the total variance of student annual growth was 0.812, the fixed effect of school SES on student growth was -0.414. The effect was negative because school SES is usually measured by the percentage of students eligible for reduced or free lunch.

The correlation between student random effects on the intercepts (r_{0i}) and the natural growth rates (r_{1i}) was set at -0.21. This was done by referring to the study of Ponisciak and Bryk (2005). The correlation between school SES and school mean kindergarten test score was -0.93, which was adopted from the study of Darandari (2004). The correlations among value-added school effectiveness over time were set at 0.6. This was done by referring to studies addressing stability of school effectiveness over time (Bosker & Scheerens, 1989; Teddlie & Reynolds, 2000; Willms, 1987). The intraclass correlation of kindergarten test scores was set as 0.21 in data generation. According to the meta-analysis of Bosker and Witzier (1995), without any adjustment, the proportion of between school variance to the total variance of student test scores at a single time point was about 0.21 in the United States.

Statistical Analyses

The current study used Spearman rho^2 to evaluate consistency between the estimated school rankings based on different VAM and the known true school rankings in the simulation data. The current study also calculated the effect sizes (i.e., η^2) of each simulation condition and their interaction on the rho^2.

In order to investigate agreement of school classifications, the current study classified schools based on their effectiveness estimates in different models. Schools with effectiveness 1 SE below the mean were classified as “ineffective”; schools with effectiveness 1 SE above the mean were classified as “effective”; other schools were classified as average. The Kappa coefficient and the Kappa Z coefficient were used to evaluate the degree of agreement of school classifications. η^2 was used to evaluate the effect of each simulation condition and their interaction on the agreement of school classifications. In addition, by using frequency analysis, the current study also explored the pattern of misclassifications.

The importance of the current study was to provide guidance for model specification and data collection in future value-added assessment of school effectiveness. In particular, the current study sought to discover the methodological problems in common practices of the value-added assessment of schools, so that people may avoid or at least keep these problems in mind when using Value-Added Models to rank or classify schools.

CHAPTER II

LITERATURE REVIEW

The first section of this chapter reviews the various definitions of school effectiveness. In order to make the definition of value-added school effectiveness more clear, the definition of value-added school effectiveness is compared with other definitions of school effectiveness. The second part reviews commonly used statistical models in current practice invoking value-added assessment of schools. The models are the two-level Gain Score Model, the Cross Classified Model, and the Layer Mixed Effect Model (LMEM). The third section reviews research about covariate adjustment and the effect size of the commonly adjusted covariates. The forth section reviews the studies about the influence of number of time points of data on school effectiveness estimates.

Concept of School Effectiveness and Value-Added Assessment

Much research has found that part of student achievement difference can be attributed to differences in schools (Aitkin & Longford, 1986; Raudenbush & Willms, 1995; Teddlie & Reynolds, 2000; Thomas & Mortimore, 1996; Goldstein, 1997). Thus, educational researchers and policy makers advocate that schools should be held accountable for student academic achievement (Millman, 1997). To fulfill the requirement of No Child Left Behind Act, many states have established accountability systems which base school evaluation on student academic achievement, such as Adequate Yearly Progress. It has been a very common practice in the United States, and even in the world, to use student test scores to rank schools, or to identify exceptional schools. However, even in the general framework of using student test scores to evaluate schools, there are different definitions of school effectiveness. These different understandings have resulted in different school effectiveness indicators and different statistical models to estimate these indicators. Comparison of different definitions of school effectiveness can help to clarify what kind of school effectiveness a certain statistical model aims at estimating and what the statistical model actually estimates.

Before reviewing different definitions of school effectiveness, a framework used to classify these definitions is presented. The first dimension of this framework is whether the school effectiveness evaluates the unique effect of school practice, or instead reflects the combined effect of both school practice and other background variables, such as student prior attainment before entering a school or student body Socio Economic Status (SES). The second dimension is whether the school effectiveness indicates the school effect on individual students, or instead on the average achievement of all the students in a school. The third dimension is whether the school effectiveness indicates the school effect on achievement growth, or instead on achievement status at only a single time point.

In terms of the first dimension, Raudenbush and Willms (1995) further discriminated Type A school effect and Type B school effect. The Type A effect isolates only student background effect. In contrast, The Type B effect isolates both student and school background effects. For policy making, estimates of Type B effect should be used to rank and classify schools. This is because the policy makers are more concerned about the effect of educational interventions on students' achievement; furthermore, they think that schools should be evaluated for what they can control.

Definition 1

One definition defined school effectiveness as the unadjusted average achievement of all students in a school (Teddlie & Reynolds, 2000). The indicator of school effectiveness based on this definition is the mean test score of all the students in a school, or the percentage of students who pass a critical level. Schools that have higher means or higher passing percentage on an achievement test have higher ranks.

This definition ignores the effect of student and school background on student achievement. It also ignores the within school difference involving student achievement. In addition, the definition only provides a snapshot of student achievement at a single time point, but not any information about student achievement growth.

The evaluation based on school unadjusted means has been criticized by many scholars for its unfairness (Darandari, 2004; Meyer, 1997; Goldstein, 1997). The main deficiency of this indicator is that it favors schools serving advantaged students who perform at a higher level before they enter the schools, or it favors schools serving students in a more wealthy community that can provide more learning resources and opportunities out of schools. However, schools should not be rewarded or punished for the factors out of their control.

While no rigorous educational researchers would consider unadjusted school means as indicators of school effectiveness, some people argue that adjustment using covariates, as is done in some other definitions, excuses for disadvantaged students to have lower achievement than advantaged students. We should have the same level of expectation for all the students no matter what their backgrounds are. A neutral way to resolve these conflicting views is to include both adjusted and unadjusted school effectiveness in school accountability systems. However, many parents, informed government officials, and education critics often use the unadjusted mean score to evaluate schools without recognizing its limitations.

Definition 2

Another definition of school effectiveness defines school effectiveness as the impact of schooling on the average achievement of all the students in a school, adjusted for family background and/or prior achievement (Teddlie & Reynolds, 2000). This definition isolates school effectives from the impact of other background variables. However, the definition still ignores the within-school difference of student achievement, and bases school evaluation on student achievement outcome, instead on student achievement growth.

The statistical model used to estimate school effectiveness in terms of this definition is the school level linear regression model with aggregated scores. This model is also known as mean on mean regression (Aitkin & Longford, 1986; Rauderbush & Willms, 1995). The model is specified as:

$$\overline{Y}_j = \alpha + \beta_b{}' * \overline{X}_j + e_j \qquad (2.1)$$

School effectiveness estimates based on this model are the residuals of regressing school mean outcome on the school means of student background variables. The

Ordinary Least Squares (OLS) estimate of β_b' is calculated using the between group sum of squares and cross products:

$$\beta_b' = \frac{\sum n_j(\overline{X}_j - \overline{X})(\overline{Y}_j - \overline{Y})}{\sum (\overline{X}_j - \overline{X})^2} \tag{2.2}$$

Under the assumption that residual term has no relationship with predictor variables, β_b' is an unbiased estimate of β_b, which is the school level regression weight in a two level hierarchical model (Raudenbush & Willms, 1995). However, because the means in small schools have large sampling errors, and also because the variance of the means across the schools include both parameter variance of true school means and error variance, β_b' usually has a large Standard Error (SE) (Aitkin & Longford, 1986; Ballou, Sanders, & Wright, 2004; Raudenbush & Willms, 1995; Raudenbush & Bryk, 2002).

Definition 3

The third definition defines school effectiveness as measuring the unique effect of each school on individual students' achievement outcomes (Teddlie & Reynolds, 2000). The third definition isolates school effectiveness from the effect of other background variables, and focuses on each individual student in a school. However, the school evaluation in terms of this definition is based on student achievement status at a single time point, but not on student achievement growth.

Three traditional student level linear regression models aim at estimating school effectiveness in terms of this definition. However, no matter how perfectly the data are collected, because the three models use single level to model the hierarchical data, the three traditional regression models theoretically estimate a mixed effect of school context and school practice variables, which is the Type A school effect (Raudenbush & Willms, 1995).

The first student level regression model is the traditional ANCOVA model, which can be specified as

$$Y_{ij} = \alpha_j + \beta_w * X_{ij} + e_{ij} \tag{2.3}$$

This model specifies a set of parallel regressions of Y on X, with each school having its own regression line. The parallel regressions differ only on the intercepts. The school effectiveness is indicated by the difference between the intercept of a school and the average intercept of all the schools. In this model, the OLS estimate of β_w is calculated using within group sum of squares and cross products, which is

$$\beta_w = \frac{\sum (X_{ij} - \overline{X}_j)(Y_{ij} - \overline{Y}_j)}{\sum (X_{ij} - \overline{X}_j)^2} \tag{2.4}$$

This model is the Model 2 in Aitkin and Longford's article (1986). When there are no school context effects or the context effects are very small, the school effectiveness estimates in this model are very close to the estimates of school practice effectiveness in multilevel models (Aitkin & Longford, 1986; De Leeuw & Kreft, 1995). When school context effect is noteworthy, this model estimates the Type A school effect but not the Type B school effect (Raudenbush & Willms, 1995). Furthermore this model underestimates the standard error of β_W. This is because OLS estimation assumes that the random errors are independent and each individual student provides a unique piece of information. However, because students are clustered in schools, the random errors of the students in a school are correlated (Goldstein, 1991; Raudenbush & Bryke, 2002).

Thus, the total information we actually have is less than that when the individual students are independent.

The second regression model is to pool all the students' data together and ignore their school statuses. The mathematical equation of this regression model is:

$$Y_{ij} = \alpha + \beta_t * X_{ij} + e_{ij} \tag{2.5}$$

The effectiveness estimate of school j is the aggregated residual of the students in school j. β_t is calculated using total sum of squares and cross products:

$$\beta_t = \frac{\sum_j \sum_i \left(X_{ij} - \overline{X}\right)\left(Y_{ij} - \overline{Y}\right)}{\sum_j \sum_i \left(X_{ij} - \overline{X}\right)^2} \tag{2.6}$$

Alwin (1976) showed that

$$\hat{\beta}_t = \eta^2 \hat{\beta}_b + (1 - \eta^2)\hat{\beta}_w \tag{2.7}$$

η^2 is the proportion of the total variance in X_{ij} that lies between schools. When η^2 is zero, which means all schools have the same mean on X, or the school context effect is zero, β_t is $\beta_{w,}$ and the school effect estimated in the pooled regression model is the Type A effect. When η^2 is one, which means all the total variance in X_{ij} is the between school variance, β_t is β_b. Thus, the school effect estimated in this model is the Type B effect (Raudenbush & Willms, 1995). However, in general cases, the total variance includes both within school variance and between school variance, and thus the school effect estimated in this model is a mix of Type A and Type B effect.

The third student level regression model is an extension of the second student level regression model, which adds school-level predictors into student level equation. This model is:

$$Y_{ij} = \alpha + \beta_1 * X_{ij} + \beta_2 * W_j + e_{ij} \tag{2.8}$$

In this model, β_1 is the adjusted $\beta_{t,}$ and β_2 overestimates the effect of school-level variables (Aitkin & Longford, 1986). Actually, at the student level, school characteristics should have no effect because for the students in the same school, school characteristics are the same. Thus, the multiple R^2 in this model is spuriously inflated. Furthermore, the sampling variance of β_2 is usually large (Aitkin & Longford, 1986).

From the above explanation, we can see that all the single level models can not estimate the Type B school effect accurately and statistically efficiently. In order to solve the problems in single level regression models and better honor the hierarchical structure of school data, multilevel models should be used to estimate school effectiveness in terms of Definition 3.

The multilevel models used to estimate school effectiveness in terms of Definition 3 usually have two levels. One is the student level; the other is the school level. The school effectiveness indicator is the residual associated with each school after taking into account the variations on other background variables. Theoretically, the school effectiveness estimates in the multilevel models are unbiased if the models are correctly specified. However, because we usually do not know which student or school level covariates should be added into the model, misspecification of the model often results in biased estimate of school effectiveness (Raudenbush & Willms, 1995). The current study did not focus on the magnitude of the bias, but on validity of the estimates for school rankings and classifications. The logic is that even if the estimates are biased, the bias may be not large enough to compromise the validity of the estimates for school rankings and classifications.

Definition 4: Value-added School Effectiveness

Compared to the above three definitions of school effectiveness, value-added school effectiveness is defined as measuring the unique impact of school practices on individual student achievement growth over time (Doran, 2003; Mayer, 1997; Teddlie & Reynolds, 2000). This definition isolates the effects of both student and school background variables from the effect of school practice. Thus, it is actually the Type B effect (Raudenbush & Willms, 1995). Second, this definition emphasizes the achievement of each student, but not a group of students. Third, this definition perceives achievement growth but not achievement outcome as the most appropriate criterion for assessing school effectiveness. In order to estimate school effectiveness in terms of this definition, we need to follow the same group of students for several years. The model is not like the cross-sectional studies of school effectiveness, which compare achievement of successive cohorts of students, such as the Adequate Yearly Progress evaluation of schools, and which do not invoke longitudinal measurements.

There are three types of statistical models commonly used to estimate value-added school effectiveness. One is the two-level hierarchical linear model with annual gain score as the dependent variable. The second is the Cross Classified Model. The third is the Layered Mixed Effect Model (LMEM or TVAAS). These three models are explained in the next section.

Review of Value-Added Models to Estimate School Effectiveness

Value-Added Models are actually different statistical models used to estimate value-added school effectiveness. This section reviews three kinds of Value-Added

Models. They are the two-level gain score model, the Cross-Classified Model, and the Layered Mixed Effect Model. For each model, the specification of the model, the value-added school effectiveness indicator in the model, the application of the model in school accountability system, and the related research on or with the model are discussed.

Two-level Gain Score Model

The Hierarchical Linear Model (HLM) overcomes the problems in one-level linear regression models addressed previously. By incorporating variables in different levels in a manner that respects the hierarchical nature of data, HLM relaxes a crucial assumption of independence of residuals (Goldstein, 1987; Raudenbush & Bryke, 2002; Snijders & Bosker, 1999). HLM provides researchers the flexibility of using analysis units at more than one level simultaneously. Thus, HLM can decompose the total outcome variance into individual-level variance and group-level variance. However, the variance decomposition in HLM is different from that in the one-level general linear model, such as ANOVA, ANCOVA, and regression. In the one-level general linear model, the between group variance is based on observed group means which are influenced by random error associated with individual unit. In contrast, the group-level variance in HLM is the parameter variance, or in other words the variance of the true group means estimated without random error (Raudenbush & Bryk, 2002). Hence, theoretically, HLM can provide more accurate and precise estimates of coefficients, variance components, and residuals at the different levels. The superiority of HLM to the one-level general linear model makes it an attractive method for estimating school effectiveness.

The most often used HLM model in value-added assessment of schools is the two-level HLM with the difference of test scores from two successive years as the outcome variable. Although models with the current year achievement as outcome and previous achievement as one of the predictors are also used in value-added assessment of schools, the models with annual gain score as outcome explicitly model student achievement growth in a year.

When the two-level gain score models are used in school evaluation, the first level is the student level, and the second level is the school level. Although classroom can be another level, the HLM models commonly used in school evaluation do not include the classroom level. The classroom level is included only when teacher evaluation is desired. The most general form of the two-level HLM is:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} * X_{1ij} + \beta_{2j} * X_{2ij} + ... + \beta_{qj} * X_{qij} + e_{ij}$ (2.9)

Level 2: $\beta_{qj} = \gamma_{q0} + \gamma_{q1} * W_{1j} + \gamma_{q2} * W_{2j} + ... + \gamma_{qs} * W_{sj} + u_{qj}$ (2.10)

Where β_{qj} is the student-level regression coefficient, which can be (a) fixed, (b) non-randomly varying across schools, or (c) randomly varying across schools. γ_{q0} is the adjusted mean of the coefficients across all the schools; γ_{qs} (where s=1, 2, …s) is the fixed effect of school-level covariates on student-level regression coefficients. When X_{ij} is group-mean centered, β_{0j} represents the annual gain of a typical student who has all the average characteristics of the students in school j. The school-level residual of the intercept, $u_{0j,}$ is the estimate of Type B value-added school effectiveness (Raudenbuch & Bryke, 2002; Raudenbush & Willms, 1995).

If the effects of student-level covariates are different across schools, then random-slopes or non-randomly varying slopes HLM need to be specified. However, in current practice of using a two-level gain score model to estimate school effectiveness, the effects of student-level covariates are usually assumed to be constant for all the schools; only intercepts are assumed to be different across the schools. The Empirical Bayes (EB) estimate of residual of the intercept of school j represents the value-added effectiveness of school j.

The two-level gain score model is a special case of mixed models (Ferron, 1997; Goldstein, 1995a, 1995b; Raudenbush & Bryk, 2002). The variance-covariance matrix of between school residuals, which is the G matrix in a mixed model, has a blocking diagonal structure with an identical block for each school. The elements in each block are the variance of intercepts (i.e., τ_{00}), the variance of the regression coefficients (i.e., τ_{qq}), and the covariance between the intercepts and the regression coefficients (i.e., τ_{0q}). Because the residual scores are assumed to be independently and normally distributed with constant variances across schools, the variance-covariance matrix of the residuals, which is the R matrix in a mixed model, has a diagonal structure with an identical diagonal element for each school (Ferron, 1997).

Dallas ISD adopts a transformed two-level gain score model in its value-added school accountability system (Webster & Mendro, 1997) The statistical solution in its accountability system has two stages. In the first stage, residuals are obtained from solving a set of student-level regression equations designed to account for the effect of ethnicity, limited English proficiency, gender, Socio Economic Status, and their first order and second order interactions. The residuals are obtained on both posttest scores and pretest score. In the second stage, two-level HLM is used on the residual scores obtained from stage 1. School background variables are used as the school-level predictors in the school-level regressions (Webster & Mendro, 1997). In the HLM model, Dallas ISD used Empirical Bayes estimates of the residuals of the schools to represent the value-added school effectiveness. The study of this two-stage model indicated that the school effectiveness estimates had only small correlations ($r< 0.10$) with both the student and the school background variables (Webster, 2005). On the other hand, the researchers also found that the school effectiveness estimates in the two stage model that adjusted for both student and school background variables highly correlated with the model that did not adjust for the school background variables ($r > 0.9$).

Cross-Classified Model

The HLM gain score model was criticized for its inability to account for long-term effects and for its low reliability in measuring change. Some researchers pointed out that school influences can only be observed over a long period of time, because some abilities do not develop obviously in a short term (Teddler & Reynolds, 2002). Some researchers indicated that a gain score has unacceptable measurement error (Doran, 2003). Thus, longitudinal models that include more than two years of data should be used to estimate school impact on student achievement growth.

The Cross-Classified Model (CCM) has been used to estimate value-added school effectiveness with more than two years of data (Hill & Goldstein, 1998; Ponisciak & Bryk, 2005; Rowan, Correnti, & Miller, 2002). The Cross-Classified Model was created for the reason that a lower level unit may also belong to several higher level units at the same time. For example, a student may belong to a school, and at the same time, also be a member of a neighborhood. If a study aims at investigating both school effectiveness and the neighborhood effectiveness, a Cross-Classified Model in which a student nested in each cell of schools by neighborhoods cross-classification is necessary (Goldstein, 1995b; Raudenbush & Bryk, 2002). The Cross-Classified Model can also be used in modeling growth longitudinally. When modeling growth, the repeated measurements are regarded as nesting in the cross-classification of two higher level units, such as the cross-classification of students by schools.

One way to consider the longitudinal Cross-Classified Model is to perceive the model as the combination of two simpler models. The first simpler model is a two-level linear growth model, which represents the natural growth of individual students given no school effectiveness exists. The first level units of this model are repeated measurements, and the second level units of this model are students. The second simpler model specifies the effectiveness of schools on student development over time. In the longitudinal Cross-Classified model, the value-added school effectiveness in each year is perceived as the deflection from a student’s expected achievement in that year given the student’s initial status and the natural growth rate (Ponisciak & Bryk, 2005). Thus, the combination of the two simpler models represents a non-linear growth trajectory. Figure 1 presents the growth trajectory specified by a Cross-Classified Model, in which u_{tj} represents the value-added effectiveness of school j at time t. The model is not like the traditional three-level longitudinal models that specify a linear growth trajectory and estimate school effectiveness on average growth rate over several years. The Cross-Classified Model specifies unique school effectiveness at each time period, and the total school effectiveness over time is the sum of the unique school effectiveness at each time period.

Figure 1

Growth trajectory of a student specified by a Cross-Classified Model

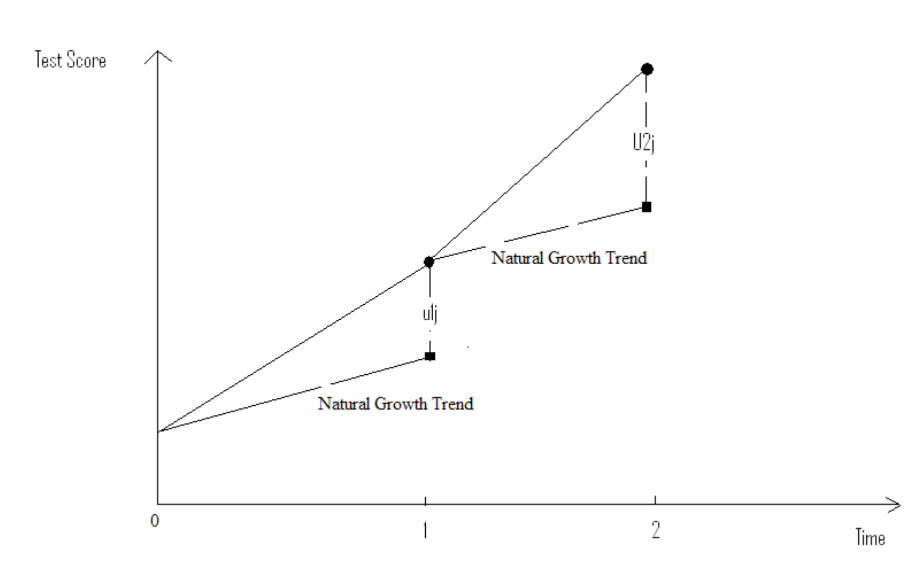

Figure 1 shows that each student has a natural growth represented by the dashed line. This growth exists even the student does not attend any schools. It is the school effectiveness that drives the student away from his natural growth trend, which makes the student either have more development or less development in a time period. Figure 1 represents the growth of a lucky student who attends a school that accelerates growth during each time period.

Depending on whether the effects of covariates are controlled, Cross-Classified Model can be classified as Unconditional Cross-Classified Model (UCCM) and Conditional Cross-Classified Model (CCCM). The UCCM is as:

$$Y_{tij} = \beta_{00} + r_{0i} + u_{0j} + (\beta_{10} + r_{1i})*t + \sum_{t=1}^{t} u_{tj} + e_{tij} \tag{2.11}$$

The CCCM is as:

$$Y_{tij} = \beta_{00} + \beta_{01} * X_{ij} + u_{0j} + r_{0i} + (\beta_{10} + \beta_{11} * X_{ij} + r_{1i}) * t + \sum_{t=1}^{t} (\gamma_1 * W_j + u_{tj}) + e_{tij} \qquad (2.12)$$

where

Y_{tij} is the test score of student i in school j at time t,

X_{ij} is the value of student i in school j on the student background variable X,

W_j is the value of school j on the school background variable W,

r_{0i} is the random effect of student i on the intercepts of the growth trajectories,

r_{1i} is the random effect of student on the natural growth rate,

u_{0j} is the random effect of school j on the intercepts of the growth trajectories,

β_{00} is the grand mean of the initial test scores of all the students,

β_{01} is the fixed effect of student background variable on their initial test scores,

β_{10} is the overall natural grow rate of all the students given no school effectiveness exists,

β_{11} is the fixed effect of student background variable on the natural growth rate,

γ is the fixed effect of school background variable on student development during a time period, and

u_{tj} is the value-added effectiveness of school j at time t.

Specifically, the test score of student i in school j from Grade 1 to Grade 3 is:

Test score at Grade 1 (time = 0):

$$Y_{0ij} = \beta_{00} + \beta_{01} * X_{ij} + r_{0i} + u_{0j} + e_{0i} \qquad (2.13)$$

Test score at Grade 2 (time = 1):

$$Y_{1ij} = \beta_{00} + \beta_{01} * X_{ij} + r_{0i} + u_{0j} \qquad \text{(Test score at Grade 1)}$$

$+\beta_{10}+\beta_{11}*X_{ij}+r_{1i}$ (Student natural growth)

$+\gamma_1*W_j+u_{1j}$ (School effect at Grade 2)

$+e_{1i}$ (Residuals at Grade2) (2.14)

Test score at Grade 3 (Time = 2):

$Y_{2ij}=\beta_{00}+\beta_{01}*X_{ij}+r_{0i}+u_{0j}$ (Test score at Grade1)

$+(\beta_{10}+\beta_{11}*X_{ij}+r_{1i})*2$ (Student natural growth)

$+2*\gamma_1*W_j+u_{1j}+u_{2j}$ (School effect at Grade3)

$+e_{2i}$ (Residuals at Grade 3) (2.15)

The Cross-Classified model is also a special case of mixed model. School effectiveness is assumed to be independently normally distributed with constant variance across years (McCaffery et al., 2004). Thus, G matrix for random school effects has a block diagonal structure with an identical block for each school; the elements on the diagonal of each block are the annual variances of school effectiveness. The G matrix for student random effects on intercepts and slopes also has a block diagonal structure with an identical block for each student. The elements in each block are the variances and covariance of the intercepts and slopes of students' natural growth curves. The variance-covariance matrix of residuals of student test score in each year, R, also has a block diagonal structure with an identical block for each student. The elements in each block are the residual variances in each year, which are equal across years.

So far, only the Chicago Public School District has used the Cross-Classified Model in school accountability (Ponisciak & Bryk, 2005). The model they used was a UCCM that did not adjust for any background variables. They used the test scores from

Grade 2 to Grade 8 to estimate the parameters in the model. In their model, they specified a linear trend for the value-added school effectiveness, so that $u_{tj}= u_{1j} + r_{uj}*t_{uj}$. Here, r_{uj} is the growth rate of the value-added school effectiveness, u_{1j} is the value-added school effectiveness in Grade 3, and t_{uj} is the time point for tracking the change of school effectiveness. The value-added effectiveness of schools on student development is observed one time point after the starting point, thus, $t_{uj}= t-1$. They found the average value-added school effectiveness over time to be highly correlated with the average gain scores ($r > 0.9$), and highly correlated with the NCLB proficiency percentage ($r > 0.8$). However, the value-added trend had a small correlation with the trend of gain score and the trend of NCLB proficiency percentage ($r < 0.3$). They didn't investigate the correlations between value-added estimates and student or school background variables.

The CCCM has not been used in the real practice of school accountability, but has been used in some research. Raudenbush (1993) conducted a study that investigated teacher effectiveness on student achievement from Grade 1 to Grade 4, which included teacher's education as a covariate at the teacher level. He found that when adding the teacher effect as a cross-classified random effect, the between student variance and the within student variance reduced, which resulted in the SEs of the parameter estimates being decreased also. Furthermore, he found that when adding teacher's education, the between teacher variance did not reduce, and the statistical test did not indicate statistically significant effect of teacher's education. He did not study the influence of controlling for teacher's education on the Empirical Bayes estimate of individual teacher's effectiveness.

Rowan, Correnti, and Miller (2002) used CCCM to analyze the data of 4,000 students, 300 teachers, and 120 schools in *Prospects: Congressional Mandated Survey of Educational Growth and Opportunity (1990 to 1994)*. Their CCCM model included four levels. The units of the four levels were repeated measurement, student, teacher, and school. They found that teacher effectiveness accounted for 60% of the covariate adjusted reliable variance of growth rate in reading, and 52% in mathematics. School effectiveness accounted for 55% and 53% of the covariate adjusted reliable variance of growth rate in reading and math, respectively. Both the teacher effectiveness and the school effectiveness in CCCM were much larger than those in the two-level HLM models with either posttest score or annual gain score as the dependent variable. Furthermore, they found that only 19% of the variance of student growth rate in mathematics after adjustment of the covariates was reliable, and the figure for reading was 28%. The small reliable variance of growth rate challenged the reliability of the annual gain score and its usage in the two-level HLM model.

Another important study on school or teacher effectiveness with CCCM was conducted by Hill and Goldstein (1998) with 59 primary schools, 365 teachers, and 6,678 students. The CCCM model only adjusted the student level covariates. The statistically significant covariates included gender, non-English speaking status, parents' occupation, critical events, and prior achievement.

Layered Mixed Effect Model (LMEM)

LMEM was developed by Willams Sanders at the University of Tennessee in 1980s'. In 1992, the Tennessee legislators passed the Education Improvement Act which adopted a set of statistical models to evaluate effectiveness of school systems, schools, and teachers on student academic gains. This set of statistical models is called Tennessee Value-Added Assessment System (TVAAS). LMEM is the most complex model in TVAAS. Originally, LMEM was used to estimate teacher effectiveness, and a Simple Fixed Effect Model (SFEM) was used to estimate school effectiveness (Sanders, Saxon & Horn, 1997). Currently, some researchers and school systems also use LMEM to estimate school effectiveness. In the current study, the author used LMEM to estimate school effectiveness.

An LMEM to estimate school effectiveness can be specified as (McCaffrey et al., 2004; Sanders & Horn, 1994; Sanders, Saxon, & Horn, 1997; Tekwe et al., 2004):

$$Y_{tij} = \mu_t + \sum_{t=1}^{t} \sum_{j=1}^{h} C_{tj} * S_{tj} + e_{tij} \qquad (2.16)$$

where Y_{tij} represents a test score for the i^{th} student in the j^{th} school at time t. μ_t represents the mean score of all the students in all the schools at time t. S_{tj} represents the value-added school effectiveness at time t. C_{tj} is the proportion of time a student spends in school j during the t^{th} time period. For example, if a student attended school j for a whole year, C_{tj} equals to 1; when a student did not attend school j at all in this year, C_{tj} equals to 0, otherwise C_{tj} equals to a fraction of a year. h represents the total number of schools.

The model is called layered model because the school effectiveness in the later years adds layers to the model for previous years (McCaffrey, et al, 2004; Sanders, Saxon, & Horn, 1997; Tekwe et al., 2004). For example, there are 22 schools available for the students to attend

Test Score at Grade 1 (time = 0) is:

$$Y_{0ij} = \mu_0 + \sum_{j=1}^{22} C_{0j} * S_{0j} + e_{0ij} \qquad (2.17)$$

Test Score at Grade 2 (time = 1) is:

$$Y_{1ij} = \mu_1 + \sum_{j=1}^{22} C_{0j} * S_{0j} + \sum_{j=1}^{22} C_{1j} * S_{1j} + e_{1ij} \qquad (2.18)$$

Test Score at Grade3 (time = 2) is:

$$Y_{2ij} = \mu_2 + \sum_{j=1}^{22} C_{0j} * S_{0j} + \sum_{j=1}^{22} C_{1j} * S_{1j} + \sum_{j=1}^{22} C_{2j} * S_{2j} + e_{2ij} \qquad (2.19)$$

LMEM is also a special case of the mixed models (McLean, Sanders & Stroup, 1991). The covariance matrix of school effectiveness (G) is assumed to have a block diagonal structure with an identical block for each school. Each diagonal element in each block is the variance of school effectiveness in a year. Unlike UCCM, the covariance matrix of residuals of repeated measurements (i.e., e_{tij}) within a student, which is the R matrix, is unstructured, which means that the correlations among the repeated measurements within a student are taken into account in model specification.

Currently, school district or statewide educational accountability systems in seven states have used the TVAAS as a component of their school accountability systems. The seven states are Tennessee, Iowa, Ohio, Pennsylvania, New York, Colorado, and Washington. The popularity of TVAAS even arouses enthusiasm in statistical software companies. SAS Inc. has set up a special department called Educational Value-Added Assessment System (EVAAS) to develop software and systems for implementing TVAAS.

The enthusiasm toward LMEM arises because the model has four main advantages. First, LMEM does not require measurement and specification of covariates. Second, it does not require all students included in the analysis have complete records of test scores. Third, LMEM can estimate school or teacher effectiveness in several subject areas simultaneously. Fourth, the estimable linear function of the estimates of LMEM parameters can produce many other meaningful values (Sanders & Horn, 1994; Sanders, Saxon, & Horn, 1997), such as the predicted gain of a student in a school. Actually, the report card does not report the Empirical Bayes estimates of school effectiveness, but report the average predicted gain of the students in a school.

The claim that “the influence of other exogenous factors can be filtered out without directly measurement and specification of these factors” (Sanders, Saxon, & Horn, 1997, P. 138) arouses much excitement and also much concern about LMEM. Sanders et al. said that “By taking advantage of longitudinal data, each student serves as his/her own control. In other words, each student can be regarded as a blocking factor. By blocking for each student, many of the exogenous influences most often cited as influencing academic progress - educational attainment of parents, socioeconomic level, race, and so on – could be partitioned without having direct measures of each one” (Sanders, Saxon, & Horn, 1997, P. 138). This big advantage absorbs extensive interest in school effectiveness research. This is because it is almost impossible to identify, measure, and specify all possible exogenous factors. Even if it was possible to measure and specify all the exogenous factors, multicollinearity would still bias the Empirical Bayes estimates of the regression weights for individual schools and the associated school effectiveness. However, many researchers also doubt the ability of LMEM to control for unmeasured covariates. Thus, they call for empirical studies to evaluate LMEM. The following paragraphs review the studies either support the claim or diminish the claim about covariates adjustment in LMEM.

Most of the empirical studies that support the claim were conducted by Sanders and his colleagues. The three pilot studies conducted in Knox County, Blount County, and Chattanooga City of Tennessee by Sanders and his colleagues concluded that (1) differences in student test score gains could be significantly explained by the differences in school effectiveness and teacher effectiveness (Mclean & Sanders, 1984); (2) school effectiveness and teacher effectiveness estimates were consistent across years (Mclean & Sanders, 1984); (3) teacher effect estimates were highly correlated with subjective report of school supervisors (Cook, 1985; Mclean & Sanders, 1984); (4) estimated student gains were not correlated with previous achievement levels and ability measurements (Cook,1985; Mclean & Sanders, 1984); and (5) school effect estimates were not correlated with school location and racial composition (Casteel, 1994).

David Haville, who is a statistic professor specialized in mixed effect model reviewed the statistical model of TVAAS (Haville, 1995). He concluded that LMEM provides a very appropriate model for school and teacher evaluation. He further indicated that although LMEM is new for educational accountability, it has been successfully used in other areas. He anticipated that application of LMEM in evaluating individual schools and teachers will be as successful as its use in other areas.

One study that argued against LMEM was conducted by Baker and Xu (1995) from the Office of Educational Accountability in Tennessee. They found that the school effectiveness estimates in one school with two groups of students were statistically significantly different. This result happened in Scotts Hill School. In this school, some of the students came from Henderson County, and some came from Decatur County. Sanders argued that this was not because of the statistical model, but because this school tailored its curriculum to lower-achieving students most of whom came from Decatur County.

Three empirical studies were conducted to check the claims about the issue of covariates adjustment in LMEM and drew a different picture from the studies conducted by Sanders and his colleagues. McCaffrey and his colleagues from the Rand Cooperation conducted a Monte Carlo study (McCaffrey, et al. 2003, 2004). They found that the ability of LMEM in controlling covariates effects depended on the distributions of the covariate variables. When a covariate variable completely randomly distributed, teacher effect estimates in LMEM did not correlate with the class average of the covariate variable. When the means of the covariate variable varied across classes, the estimated teacher effect in LMEM moderately correlated with the class average of the omitted covariate variable (r = 0.47). When the means of the omitted covariate variable varied across schools, the estimated teacher effect in LMEM highly correlated with the class average on the omitted covariate variable (r = 0.79). The LMEM used to estimate teacher effect in McCaffrey's study include both teacher effect and school effect. However, the LMEM model used to estimate school effect in the current study and other related research (Tekwe et al., 2004) didn't include teacher effect. Thus, we should not uncritically generalize the results of McCaffrey's study about teacher effect estimates to school effect estimates.

Ballou, Sanders and Wright (2004) found that adding student-level covariate, such as eligible for free lunch, to the original LMEM model did not change the estimates of teacher effectiveness substantially; however, adding classroom level or school level covariates, such as percentage of students eligible for free lunch, produced a different picture. Specifically, student-level covariates have small coefficients, and the modified model resulted in the same identification of exceptional teachers as the original model. In contrast, there is less agreement between teacher effect estimates in the original model and the modified model with classroom level or school level covariates specified. Thus, LMEM seems to be able to control for the effect of student level covariates, but not classroom level or school level covariates. However, this conclusion is still doubted because the SE of the coefficient associated with the classroom level or school level covariate was large.

Although the previous two studies focused on teacher effectiveness estimation, we may expect similar results when using LMEM to estimate school effectiveness. However, because the sample size condition is usually different between teacher effectiveness studies and school effectiveness studies, we can not simply make conclusions for school effectiveness estimation based on teacher effectiveness studies.

The third study was conducted by Tekwe et al. (2004). They found that the school effectiveness estimates in LMEM were highly correlated with the school effectiveness estimates in the Simplest Fixed Effect Model (SFEM) in which effectiveness of school j is the unadjusted discrepancy from the overall mean of all the schools. Furthermore, the school effectiveness estimates in the LMEM were also highly correlated with the estimates in an unconditional two-level gain score model. Thus, they concluded that specification of the correlations among repeated measurements and the Empirical Bayes estimation in the LMEM did not have obvious advantage over the simpler model in school effectiveness estimates. However, they thought their results may be limited because only two years of test scores were used. Thus, the current study investigated the ability of LMEM to control for covariate effects with three years of test scores.

Covariates Adjustment and Their Effect Sizes

This section has three parts. In the first part, the rational for covariate adjustment is presented. In the second part, the student and school background variables that were typically adjusted for in value-added assessment of schools, and the effect sizes of the typically adjusted covariates are reviewed. The literature review in this part provides information for specifying the model and setting up the parameter values to generate data in the current simulation study. In the third part, the studies of the effects of ignoring covariates are reviewed.

Rational for Covariates Adjustment

The definition of the Type B value-added school effectiveness requires to isolate school effectiveness from all the background variables that beyond the control of schools. Thus, evaluation of schools should not benefit the schools with more advantaged students or those located in a wealthier community.

The question about unique effectiveness of school practice on student learning implies a causal effect inference, and randomized experimental design is the best research method that can answer questions about causal effect (Braun, 2005;

Raudenbush, 2005; Rubin, 2004). The best estimation of Type B school effect results from a nested random block experimental design. Let us imagine that the practice in each school is a treatment. First, J schools having identical contexts are assigned to different treatment levels that vary in terms of practice. Next, blocks of J students of identical background and aptitude are assigned to the J schools. The average performance of the J students in the same block represents the average effectiveness of the J schools' practices. The discrepancy between a student's performance and the average performance of the J students in the same block is the Type B effect (Raudenbush & Willms, 1995).

However, random assignment is almost impossible in education. Thus, the ultimate goal in estimating value-added school effectiveness is to obtain a causal effect inference using observational data (Rosenbaum, 2002; Rubin, 2004). Without the benefit of random assignment, researchers tried to use statistical models to isolate school practice effects from the effects of other background variables.

One way to control for the background variables is to explicitly specify covariates in the model. The conditional two-level gain score model and the conditional Cross-Classified Model adopt this method to control for the background variables. One question that must be answered when using this method is for which covariates adjustments should be made. Another question is how the school effectiveness estimates and the applications based on these estimates will be biased if some related covariates are ignored. We now turn to these issues.

Typical Covariates Adjusted for and Their Effect Sizes

Teddlie and Reynolds (2000) reviewed the covariates adjusted for in school effectiveness studies. They found that the five most frequently adjusted covariates at the student level included previous achievement status, eligibility for free or reduced lunch, ethnicity, proficiency of English, and parents' occupational status. The most adjusted school-level covariates were aggregated student characteristics, which include school mean of previous test scores, and percentage of students eligible for free or reduced lunch.

As far as student background variables, results of previous research indicated that student prior attainment explains most of the total variance in student achievement outcomes or annual gain scores. Because of multicollinearity between intake attainment and other student background variables, adjustment with only intake attainment achieved similar total variance reduction as adjustment with all student background information. Thomas and Mortimore (1996) found that when only intake attainment was accounted for, the total variance reduced 58%, which is similar to that (63%) in the more refined model in which all student background variables, including gender, ethnicity, parent education, Socio Economic Status, were adjusted. These results were consistent with other research (Cuttance, 1992; Gray, Jesson, & Sime, 1990; Goldstein et al., 1993; Sammons, Nuttall, Cuttance, & Thomas, 1995; Scheerens, 1992; Willms, 1992). Furthermore, when student prior attainment was accounted for, the difference between the schools with the highest and the lowest mean scores reduced 70.6%, and the variance accounted by school level factors reduced from 14% to 10%. These figures about the between school variance reduction were similar to the refined model. Thus, the adjustment with other student background variables beyond intake attainment just marginally refined the results.

As far as school background variables, most research conducted using Value-Added Models with multilevel analysis adjusted student body Socio Economic Status (SES), which is usually measured as the percentage of students eligible for free or reduced lunch (Cuttance, 1992; Goldstein, 1997; Raudenbush & Willms, 1989; Teddlier & Reynolds, 2000; Thomas, Sammons & Mortimore, 1997; Willms, 1986). In the United States, Louisiana state and Dallas Independent School District adjust SES of student body in their school accountability systems (Mendro, 1998; Teddlie & Stringfield, 1993; Webster, 2005; Webster & Mendro, 1997). Darandari (2004) indicated that SES of student body has a strong correlation with other context variables. Once the variation of SES of schools has been accounted for, other school context variables, such as percentage of minority, rarely have residual effects on student achievement. Thomas and Mortimore (1996) also investigated the impact of Socio Economic Status of the student body. They found that once individual student Socio Economic Status (SES) was adjusted at the student level, the impact of SES of student body could not be detected. However, if SES of individual student was not adjusted for, even if their intake attainments were adjusted, the SES of student body still had statistically significant impacts on student achievement.

Reviews of research about school effectiveness on student achievement outcome indicated that SES of student body explained more than 50% of the school level variance of achievement outcome (Willms, 1987). For student achievement progress, Teddlie and

Stringfield (1993) and Willms (1987) indicated that 35% of the between school variance in annual gain scores was explained by SES of the schools.

Influence of Ignoring Covariates

The independence assumption of multilevel modeling requires that the residuals at level 1 are independent from the predictors at both level 1 and level 2, and with the level 2 residuals. In addition, the residuals at level 2 are assumed to be independent from the level 1 and level 2 predictors (Darandari, 2004; Goldstein, 1995a, 1995b; Raudenbush & Bryk, 2002). When a related covariate is ignored, its effect will be included in the residuals. If this covariate correlates with other covariates, which is very common in educational research, the independence assumption will be violated. This problem is especially serious at school level. This is because the measurements of school practice and school context are very difficult and seldom conducted. Thus, it is not clear yet which school background variables influence school effectiveness and should be adjusted. The omitted school background variables become a component of the school-level residual. If the omitted school background variables are correlated with the specified school background variables, the independence assumption at the school level is violated. Even if all the related background variables are specified in the model, without specifying school practice variables, the dependence between predictors and residuals at the school level can not be avoided because school background usually correlates with school practice. For example, high social economic schools are more likely to attract highly competitive administrators and teachers.

Violation of the school-level independence assumption can bias the coefficients of school-level covariates associated with the intercepts (β_{0j}) (Kim, 1990; Krull, 1997). Furthermore, the violation can bias SEs of the grand mean of intercepts (γ_{00}) and the grand mean of the student level regression weights (γ_{01}) (Krull, 1997; Raudenbush & Bryk, 2002). Specifically, Donoghue and Jenkins (1992) found that violation of this assumption can increase SE of γ_{00}.

Violating this assumption can also bias the estimates of variance components at the school level. Krull (1997) found that violation of this assumption increased variance of intercepts across level-2 units, τ_{00}. However, although Donoghue and Jenkins (1992) found bias on $\tau_{00,}$ they did not find any particular pattern of bias. What they found is underestimation of variance of the slopes across level-2 units, τ_{11}.

In HLM, estimation of one parameter influences estimation of other parameters. In order to obtain estimation of residuals for each school, we need to obtain estimations of regression coefficients for each school. In HLM, a regression coefficient for each school is the weighted mean of the grand mean of the regression coefficients across all the schools and the OLS estimate of the regression coefficient in that school (Raudenbush & Bryk, 2002). This estimate is called as Empirical Bayess estimate, or shrinkage estimate, or Best Linear Unbiased Predictor (BLUP). Maximum Likelihood estimations of the grand means of the regression coefficients depend on estimation of the variance components. Thus, if variance components estimates are biased, the grand means of the regression coefficients are biased, and so, the estimated regression coefficients and the residual for each school are biased.

Given the fact that violating independence assumption at school level is almost unavoidable, the estimate of school practice effect is almost always biased. Given other assumptions are met, if some school background variables are omitted, the estimates of school effectiveness in traditional multilevel models are at most estimating the confounded Type A and Type B effects (Carter, 2004; Raudenbush, 2004; Raudenbush & Willms, 1995; Rubin, 2004). In this situation, the Type B effect is overestimated. Even if all the related school background variables are specified, the correlation between school background variables and school practice will pull down the estimates of school practice effectiveness. Thus, what is more important is not whether or not the value-added school effectiveness is biased, but is whether the magnitude of the bias is too serious to invalidate applications such as school ranking or classification. The reason for saying this is that validity is not for the measurement itself, but for the inference based on the measurement (Crocker & Algina, 1986). The estimates of school effectiveness in a value-added model may be invalid for evaluating the absolute effectiveness of the schools, but may still be valid for ranking schools, or selecting exceptional schools.

In contrast to prior studies that investigated the effect of omitted covariates on parameter estimates, the current study investigated the influence of omitted covariates on school rankings and classifications. Another difference between the current study and prior studies is that the current study also investigated the effects of omitted covariates in a longitudinal model with more than two time points of data. Most previous studies only explored the effect of omitted covariates in a gain score model with only two time points of data, as noted previously.

The magnitude of the effect of ignoring related covariates on the school effectiveness estimate is influenced by other factors. These factors include sample size, effect size of the omitted covariates, correlation between the omitted covariates and the specified covariates, and the intraclass correlation of the covariates if the school level covariate is an aggregated variable derived from the student level covariate.

In a simulation study with the traditional two-level HLM model, Busing (1993) found that the group variance tended to be underestimated when the number of groups was small. Having a large number of groups, in general, is more important than having a large number of individuals per group in estimating group level fixed effect and variance components. He further suggested that highly accurate estimates of group level variance components needed at least 100 groups. Kreft (1996) suggested a rule of thumb that if the interest is in the fixed part, at least 30 groups with 30 individuals in each were required. If the interested is in the random part, the number of groups should be at least 100 with at least 10 individuals per group. On the other hand, Brown and Draper (2000) found that with as few as six to twelve groups, Restricted ML estimation (RML) provided reasonable estimates of variance components, and with 48 groups, both RML and Full information ML (FML) could provide reasonable variance estimates. Mass and Hox (2004) found that given the intraclass correlation of the residuals was not too large, which was 0.1, 0.2, or 0.3 in their study, and the distributions of the residuals were normal, small number of groups turned out to be a problem only with respect to the SE of the second level variance. Both the fixed parameter estimates and the variance parameter estimates had negligible bias. The largest percentage relative bias was -0.3% when the number of groups was 30, group size was 5, and ICC was at the highest level (i.e., 0.3). They also conducted a small simulation with even smaller sample size. They found that with only 10 groups of group size 5, the estimates of fixed effects and variance components were negligibly biased; the largest bias happened on the second level variance components, which was 25% upward. However, the SEs of the variance components were seriously biased, and the largest coverage rate was only 30.4%. The study of Mass and Hox (2004) was consistent with the claim of Snijders and Bosker (1999, P. 44) that multilevel modeling became attractive only when the number of groups larger than 10. In a simulation study, Darandari (2004) also found that with 50 schools and 30 students per school, the bias of the variance of intercepts at the school level was only about 0.3%.

Mass and Hox (2004) suggested that the conflict among the studies about the influence of sample size might be due to the different ICC level in different studies. Busing (1993) used an ICC level as high as 0.8, which is not common in educational contexts (Bosker & Witzier, 1995). Furthermore, Kamali (1992) found that if there were biases in estimates of level-2 parameters, increasing number of individuals per group even further increased the biases. In addition, Raudenbush and Bryk (2002) and Goldstein (1995b) pointed out that unequal group sizes might influence the uncertainty of Empirical Bayes (EB) estimates in HLM.

The effect size of the omitted school-level covariate is another factor that influences the effect of ignoring covariates. Weerasinghe and Orsak (1998) found that the instability of school ranks tended to increase as the explained variance decreases. Willms (1988) also found that when school-level had a small amount of explained variance, and the variables that explained the variance was omitted, the rankings of schools changed substantially.

The strength of correlation between the omitted school-level covariate and the specified school-level covariate also influences magnitude of the bias on fixed effect and variance component of the level-1 intercept. Darandari (2004) found that stronger correlation was associated with larger bias. When the level of collinearity was -0.5, ignoring one covariate resulted in a percentage of bias from 15% to 27% on the fixed effect estimates, and 0.4% to 8% on the intercept variance component estimate, depending on the effect size of the omitted covariate.

The ICC of the covariate is another factor that influences the effect of omitted covariate. Rosenbaum (2002) and Rubin (2004) indicated that in observational studies, statistical models can control for the effects of covariates only when groups are similar on the covariates. When groups have small overlap on the covariates, statistical models produce inaccurate estimates even if the covariates are specified correctly. The magnitude of overlap of groups on a covariate can be measured by intraclass correlation of the covariate. The Intraclass correlation is the ratio of between group variance to total variance (Raudenbush & Bryk, 2002). Goldstein (1995b) found a similar phenomenon that intraclass correlation of a covariate could influence the accuracy of parameter estimation in HLM.

In terms of covariate adjustments in Value-Added Models, another interesting question is whether some strategy can reduce or even completely delete the effects of omitting covariates. The LMEM created by Sanders (Sanders & Horn, 1994; Sanders, Saxon, & Horn, 1997) is claimed to have this ability. As noted previously, LMEM does not explicitly specify any covariates in the model, but takes the correlations among repeated measurements nested within a student into account. However, the studies about LMEM found conflicting results regarding its efficiency in isolating the covariates' effects from school effectiveness estimates. Thus, the current study empirically investigated how valid are LMEM school effectiveness estimates for school rankings and classifications.

Influence of Number of Time Points

The definition of value-added school effectiveness implies a focus on modeling achievement growth of individual students, but not modeling achievement outcomes at a single time point. There are two classes of models that model achievement growth. One is the gain score model with data from only two time points, and the other includes different kinds of longitudinal models with data from three or more time points.

Studies that compared the two types of models found that they led to different estimates of school or teacher effectiveness. With the same dataset, Rowen, Correnti, and Miller (2002) estimated teacher effectiveness with both a two-level gain score model and Conditional Cross-Classified Model (CCCM). When using the gain score model, they found that between 6% and 13% of the variance in adjusted gains of mathematics lied among classrooms depending on the grade. On the other hand, with the CCCM model, they found that to which classroom the students were assigned accounted for between 51% and 72% of the reliable variance in students' growth rate in mathematics.

Considering the difference between the gain score model and the longitudinal models, many researchers suggested using longitudinal design with at least three years of data (Hill & Rowe, 1996; Mortimore, Sammons, Stroll, Lewix & Ecob, 1988; Raudenbush 1989; Raudenbush & Bryk, 1989; Teddlie & Reholds, 2000). This is because school effectiveness is most likely to present over a long term. For example, writing ability may not be obviously improved in one year, but may only be detectably improved in a long term of schooling (Teddlie & Reynolds, 2000). Furthermore, Sammons and Goldstain (1995) suggested that beyond adjusting for intake attainment, school effects in previous years should also be controlled in estimating school effectiveness for the current year.

On the other hand, the gain score model has some advantages over the longitudinal models. One is that the test scores of adjacent grades are more likely to measure the same construct than the test scores over a wide span of grades, because subject matters may change qualitatively in their nature across several years of instruction. The change of student achievement test scores over a wide time span is difficult to interpret. This is because the change may be due to the students' development on the same construct measured over time, or may be due to the tests measuring different constructs at different grade levels. For example, the mathematic tests at Grades 3 and 8 may have dramatically different content and require qualitatively different skills. The shift in measured constructs violates the assumption of vertical equating of test scores, which is a requirement of most longitudinal statistical models. Conversely, difference scores on a vertical scale for adjacent grades are easier to interpret because the assessments designed for using in adjacent grades usually measure relatively similar knowledge content and cognitive process.

The second advantage of the gain score model is that the model has fewer requirements on data collection and maintenance than the longitudinal models. The longitudinal models require tracking individual students for at least three years and maintaining the link between the student records and the school records. Although it is possible to establish such a complete database given the modern information technology, cost for such complex databases can be high. Furthermore, many states and school districts around the country have not established such databases. These states or school districts may wonder whether they can use a gain score model with only test scores at two adjacent grades to evaluate schools in terms of value-added effectiveness. The second advantage of the gain score model leads to its third advantage. Because the gain score model has fewer time points of data collection, the data used by the gain score model generally has less missing data than the data used in the longitudinal models.

The identified problems of the annual gain score model and the studies about the differences between the gain score model and the CCCM (Rowan, Correnti & Miller, 2002) were based on a gain score model that adjusted for the test scores collected one year before estimating school effectiveness. Sammons (1996) addressed the use of baseline attainment or achievement data collected after a period of years in the same school and noted that this was likely to lead to a reduction in the estimated school effectiveness. Cuttance (1985) also cautioned against the use of prior achievement as controls when they are approximal to the point at which the school effects are measured. Preece (1989) commented on the potential problem of partialling out school effects in such cases.

Still unknown is whether the estimates of school effectiveness will be more accurate if other prior test scores are used for adjustments. Theoretically, the ideal method should be to adjust test scores at the point of entry to an instruction period. One reason is that by adjusting for test score at the point of entry to a school, the school effectiveness in previous years is not ignored when estimating school effectiveness in the current year. Because test scores collected before entering to a school is not influenced by the quality of the school under concern, adjustment with prior test scores collected before entering a school reduces the correlation between adjusted test scores and the school effectiveness under concern, which reduces the correlation between the predictor and the residual. High correlation between the predictor and the residual leads to underbiased estimates of the school level variance. The current study empirically compared the two methods of adjusting prior achievement: adjusting for test scores at the end of kindergarten which was regarded as the point of entry to an elementary school versus adjusting for test scores at the end of Grade 4 which was one year before the grade in which the school effectiveness was estimated.

CHAPTER III

METHODOLOGY

In the present Monte Carlo study, 200 datasets were generated using the Conditional Cross-Classified Model under each simulation condition. Value-added school effectiveness was estimated with each of the four Value-Added Models. The estimates of value-added school effectiveness were then used to create school rankings and classifications. Pairwise comparisons of school rankings and classifications were conducted. The consistencies of school rankings were quantified by Spearman rho^2; the agreements of school classifications were quantified by unstandardized Kappa coefficient and the Kappa Z coefficient.

This chapter presents the methodology of the study. It is divided into five sections. The first section articulates model specifications, which include a simulation model used to generate the datasets and four estimation models used to estimate value-added school effectiveness. The second section illustrates the simulation conditions. The third section introduces the parameter values used in the current study to generate the datasets. The fourth section presents the simulation procedure. The fifth section addresses the statistical analysis used to compare school rankings and classifications based on different Value-Added Models.

Model Specification

A total of five models were specified in the current study. One model, which is the CCCM model, was used to simulate student test scores from the end of Grade 1 to the end of Grade 5. Two gain score models were used with students' test scores at Grade 4 and Grade 5 to estimate school effectiveness in Grade 5. One gain score model adjusted students' test scores at the end of kindergarten (Gain_kindergarten), the other gain score model adjusted the students' test scores at the end of Grade 4 (Gain_grade4). Two kinds of longitudinal models, the Unconditional Cross-Classified Model (UCCM) and the Layered Mixed Effect Model (LMEM), used student test scores from Grade 3, Grade 4, and Grade 5 to estimate school effectiveness in Grade 5. Next, the specifications of these models are presented.

Simulation Model Used to Generate Simulated Data

A CCCM model was used to generate students' test scores from the end of Grade 1 to the end of Grade 5. The CCCM model assumes a linear growth trajectory for each student given no school effectiveness, which is called the natural growth. The value-added school effectiveness is viewed as "deflection" from the linear growth trajectory if the student encounters a school with effectiveness u_{tj} at time t (Hill & Goldstein, 1998; McCaffrey et al., 2004; Ponisciak & Bryk, 2005; Raudenbush & Bryk, 2002; Rowen, Correnti & Miller, 2002). There are two covariates in the CCCM model. Based on review of the typical covariates adjusted in school effectiveness research, the CCCM included test scores at the end of kindergarten as the student-level covariate that influenced the intercepts of students' individual growth trajectories. For simplicity, no covariate in the current study influenced students' natural growth rates. However, the correlation between intercept and natural growth rate was taken into account. The school-level covariate was student body SES. The school-level effect in the current year (i. e., Grade 5) included the fixed effect of student body SES and the residual school effect which is the value-added school effectiveness. All school effects in the previous years were assumed to persist undiminished in the current year. In the current study, students did not change schools once they entered a given school. The CCCM equation used to simulate data in the current study was:

$$
\begin{aligned}
Y_{tij} = \beta_{00} + \beta_{01} * kinder_{ij} + r_{0i} + u_{0j} \quad & \text{(Intercepts of individual students' growth curves)} \\
+ (\beta_{10} + r_{1i}) * time \quad & \text{(individual students' natural growth rates)} \\
+ \sum_{t=1}^{t} \left(\gamma_{01} * SES_j + u_{tj} \right) \quad & \text{(School level effects)} \\
+ e_{tij} \quad & \text{(residual in each grade)} \qquad (3.1)
\end{aligned}
$$

where

Y_{tij} is the test score of student i in school j at time t (for Grade 1, t = 0),

$kinder_{ij}$ is the test score of student i in school j at the end of kindergarten,

SES_j is the Socio Economic Status of school j,

β_{00} is the grand mean of test scores at the end of Grade 1. Because the Grade 1 test scores and the kindergarten test scores were standardized in the current study, β_{00} was 0 in the current study,

β_{01} is the fixed effect of students' test scores at the end of kindergarten.

r_{0i} is the random effect of student i on the intercept of his growth trajectory, which is his test score at the end of Grade 1,

u_{oj} is the random effect of school j on the intercept,

β_{10} is the overall natural growth rate of all the students,

r_{1i} is the random effect of student on the natural growth rate,

γ_{01} is the fixed effect of student body SES on student development during a year,

u_{tj} is the value-added effectiveness of school j at time t, and

e_{tij} is the residual test score of student i in school j at time t.

Four School Effectiveness Estimation Models

For each simulated data set, four different models were used to estimate value-added school effectiveness in Grade 5. Three models have been used in school accountability systems in various states. These three models were two-level gain score model adjusted for test scores from the previous year, Unconditional Cross-Classified Model (UCCM), and Layered Mixed Effect Model (LMEM). The other model, which is a two-level gain score model adjusted for the test scores collected at the point before entering a school (e. g. the test scores at the end of kindergarten before entering an elementary school), was examined to investigate the problem of adjustment with test scores from the previous year (e. g., the test scores at the end of Grade 4 when estimating school effectiveness in Grade 5).

Model 1. Two-level gain score model adjusted for the test scores at the end of kindergarten (Gain_kindergarten).

Model 1 was a traditional covariate adjusted two-level model. The dependent variable at the student level was the gain score from the end of Grade 4 to the end of Grade 5. The independent variable at the student level was the test score collected at the end of kindergarten, whose effect was fixed or held constant across schools. The intercepts at the student level, which were the adjusted school means of students' gain scores, were randomly varied across schools. At the school level, the adjusted school mean of students' gain scores was the dependent variable, and study body SES was the covariate. The Empirical Bayes estimate of the residual for school j was the value-added effectiveness of school j during the year of Grade 5. The mathematic equation of Model 1 was:

Student level model:

$$d_{ij} = \beta_{0j} + \beta_{1j} * kinder_{ij} + r_{ij} \tag{3.2}$$

School level model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * SES_j + u_{0j} \tag{3.3}$$

$$\beta_{1j} = \gamma_{10} \tag{3.4}$$

When equations 3.3 and 3.4 were used to replace β_{0j} and β_{1j} in equation 3.2, the combined equation is obtained:

$$d_{ij} = \gamma_{00} + \gamma_{01} * SES + \gamma_{10} * Kinder_{ij} + u_{0j} + r_{ij} \tag{3.5}$$

In Model 1, the variances of r_{ij} are constant across schools. The SAS code used to specify and estimate Model1 was (Singer, 1998):

```
proc mixed data=student noclprint;
 class id school;
 model gain = kinder SES/solution ddfm=bw;
 random intercept /sub=school solution;
run;
```

Model 2. Two-level gain score model adjusted for the test scores at the end of Grade 4 (Gain_grade4).

Model 2 was also a two-level gain score model. Model 2 has the same format as Model 1. The only difference was that Model 2 adjusted students' test scores at the end of Grade 4 instead of their test scores at the end of kindergarten. The two-level equation of Model 2 is:

Student level model:

$$d_{ij} = \beta_{0j} + \beta_{1j} * Grade4_{ij} + r_{ij} \qquad (3.6)$$

School level model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * SES_j + u_{0j} \qquad (3.7)$$

$$\beta_{1j} = \gamma_{10} \qquad (3.8)$$

The combined equation of Model 2 is:

$$d_{ij} = \gamma_{00} + \gamma_{01} * SES + \gamma_{10} * Grade4_{ij} + u_{0j} + r_{ij} \qquad (3.9)$$

The SAS code used to specify and estimate Model 2 was:

```
proc mixed data=student noclprint;
 class id school;
 model gain = grade4 SES/solution ddfm=bw;
 random intercept /sub=school solution;
run;
```

Model 3. Unconditional Cross-Classified Model

Model 3 had the same form as the CCCM but did not adjust for any covariates. Specifically, the mathematic equation of Model 3 is:

$$\begin{aligned} Y_{tij} = \beta_{00} + r_{0i} + u_{0j} &\quad \text{(Intercepts of students' growth curves)} \\ + (\beta_{10} + r_{1i}) * time &\quad \text{(students' natural growth rate)} \\ + \sum_{t=1}^{t} u_{tj} &\quad \text{(School level effects)} \\ + e_{tij} &\quad \text{(residual at each grade)} \qquad (3.10) \end{aligned}$$

For each school, the covariance matrix of u_{tj} is diagonal with the variance of u_{tj} at each time point as a diagonal element. For each student, the covariance matrix of r_{0i} and

r_{1i} is unstructured. For each student, the covariance matrix of e_{tij} is diagonal, with constant variances of e_{tij} as the diagonal elements (Doran, 2003). Given five schools, the SAS code used to specify and estimate the UCCM model is:

```
PROC MIXED DATA=t5 method=REML scoring=100 convh=10E-4 noclprint;
CLASS id;
MODEL SCORE= Time / SOLUTION;
random intercept time / type=un sub=id;
random z0_1-z0_5/  type= toep(1) solution;
random z1_1-z1_5/ type= toep(1) solution;
random z2_1-z2_5/ type= toep(1) solution;
parms (0.22) (0.06) (0.325) (0.317) (0.317) (0.317) (0.1)/hold=4 5 6 7;
RUN;
```

This piece of SAS code was transferred from a piece of R code that was used to specify and estimate the same kind of model. The R code was written by Lockwood, Doran and McCaffrey (2003). In order to verify that the SAS code was transferred correctly, the correlation between the SAS and the R estimates of school effectiveness was calculated.

The set of zt _ j variables was a set of dummy variables, with “1” indicating that student i was in school j during the year t (Sanders, Saxon, & Horn, 1997; Tekwe, et al., 2004). Table3.1 illustrates a heuristic dataset with two students in two schools; and each student has three observations.

Table 3.1

Heuristic dataset for fitting the UCCM model

Student ID	School ID	Time	Score	Z0_1	Z0_2	Z1_1	Z1_2	Z2_1	Z2_2
1	1	0	500	1	0	0	0	0	0
1	1	1	580	1	0	1	0	0	0
1	1	2	600	1	0	1	0	1	0
2	2	0	560	0	1	0	0	0	0
2	2	1	590	0	1	0	1	0	0
2	2	2	620	0	1	0	1	0	1

Model 4. Layered Mixed Effect Model (LMEM)

LMEM does not specify any covariates. Furthermore, LMEM does not specify any pattern of student growth trajectory after adjusting for school effectiveness (McCaffrey et al. 2004). Because the current study assumed that students did not change their schools during their elementary schooling, the multiplicative factor C_{tj} *is either 0 or 1. Thus the equation of LMEM in the current study was:

$$Y_{tij} = \mu_t + \sum_{t=1}^{t} S_{tj} + e_{tij} \qquad (3.11)$$

S_{tj} are independently normally distributed variables. Thus, for each school, the covariance matrix of S_{tj} is diagonal with the variance of school effectiveness at time t on the diagonal. For each student, the covariance matrix of e_{tij} is unstructured. It is the unstructured covariance matrix of e_{tij} that is supposed to control the covariates' effects.

This SAS code was adopted from the article of Tekwe et al. (2004). The encoding method of z variables was the same as that in the UCCM model. Given 5 schools, the SAS code for specifying and estimating the LMEM model was:

```
PROC MIXED DATA=valueadd.t5 method=REML scoring=100 convh=10E-4
noclprint ;
CLASS id t1 t2 t3;
```

```
MODEL SCORE= t1 t2 t3/ noint;
random z0_1-z0_5 /type=toep(1) solution;
random z1_1-z1_5 /type=toep(1) solution;
random z2_1-z2_5 /type=toep(1) solution;
repeated /TYPE=UN SUB=ID;
parms
(0.3170)(0.3170)(0.3170)(0.7479)(0.5711)(0.8483)(0.5411)(0.9866)(1.4986)/hold=1 2 3;
RUN;
```

Simulation Conditions

The simulation study had a 2*2 design. Two conditions were varied in the simulation: (1) Number of Schools (NS), which had two levels: 50 and 10; (2) Number of Students per School (NSS), which had two levels; 50 and 10. The levels of the NS condition were partially based on the NCES 2003-2004 Public Elementary/Secondary Universe Survey Data and partially based on the typical number of schools studied in most school effectiveness research. The distribution of the number of elementary schools in a school district is extremely positively skewed in the United States (skewness = 31.78). Among the 13,479 school districts that contain regular elementary schools, four school districts have more than 300 regular elementary schools respectively; 23 school districts have 100-300 regular elementary schools; 61 school districts have 50-100 regular elementary schools; 98 school districts have 30-50 regular elementary schools; 98.65% of the school districts have less than 30 regular elementary schools; 97.3% of the school district have less than 20 regular elementary schools, 92.6% of the school districts have less than 10 regular elementary schools. In school effectiveness research, 50 groups was a frequently occurring number; 30 groups was mentioned as minimum (Mass & Hox, 2004). However, a smaller number (i. e., 10) was used as the lowest level of the NS condition to enlarge the difference between the highest level and the lowest level of the NS condition, so that the influence of the NS condition would be more obvious if it really exists. In addition, NCES 2003-2004 Public Elementary/Secondary Universe Survey Data indicates that 92.6% of school districts have less than 10 elementary schools. In summary, the two levels of the NS condition were 50 and 10.

The levels of the NSS condition were selected based on the literature and the capacity of the computer used in the simulation. A size of 50 was chosen because the literature suggests this number is more than sufficient for such models (Mass & Hox, 2004). Although group size of 30 is normal in educational research (Mass & Hox, 2004), 10 was used as the lowest level of the NSS condition in order to enlarge the difference between the two levels of the NSS condition. Furthermore, in a pilot study, I found that large school size led to both a dramatic increase of computing time and convergence problems. With 50 schools and 100 students per school, the estimation of UCCM and LMEM in SAS and R could not achieve convergence. Thus, this study did not use large school size, such as 100.

Parameter Values

The CCCM model was used to generate the data. The CCCM model included three fixed effect parameters and five variance component parameters. The three fixed effect parameters were: (1) the effect of kindergarten test score on the intercept which was the test score at the end of Grade 1 (β_{01}), (2) the grand mean of students' natural growth rates (β_{10}), and (3) the effect of student body SES (γ_{01}). The five variance component parameters were: (1) the variances of measurement errors in each year , which were equal across years (i.e., $\sigma^2_{e0} = \sigma^2_{e1}= \sigma^2_{e2}= \sigma^2_{e3}= \sigma^2_{e4}= \sigma^2_{e}$), (2) the variance of student random effect on the intercept (i.e., σ^2_{r0i}), (3) the variance of school random effect on the intercept (i.e., σ^2_{u0j}), (4) the variance of student random effect on the natural growth rate (i.e., σ^2_{r1i}), and (5) the variances of school value-added effectiveness, which were constant across years (i.e., $\sigma^2_{u1j}= \sigma^2_{u2j}= \sigma^2_{u3j}= \sigma^2_{u4j}= \sigma^2_{uj}$) . Besides the parameter values in the CCCM model, four kinds of correlations that may influence estimates of value-added school effectiveness were also considered. The four correlations were: (1) correlation between the intercept and the slope of student natural growth trajectory (i.e. r_{r0ir1i}), (2) correlation between the student body SES and the aggregated school kindergarten test score, (3) correlations among school value-added effectiveness over time, and (4) intraclass correlation of the kindergarten test scores. Table 3.2 lists the parameter values of fixed effects, variance components, and the correlations.

The parameters for the fixed effects, variance components, and the correlations were selected to reflect the findings in school effectiveness research. For effect size of the kindergarten test score, Thomas and Mortimore (1996) found that student prior attainment alone accounted for 58% of the total variance of student achievement outcomes at a single time point. Given the within school variance is about 90% when adjusting for student prior attainment (see a meta-analysis by Bosker & Witzier, 1995), student kindergarten test score explains about 64% of the within school variance of student test scores at the end of Grade 1, and 36% of the within school variance is due to student level residual r_{0i} and measurement error e_{0it}. In term of the total variance of test scores at the end of Grade 1, the percentage accounted for by student level residual and measurement error is about 32%. Because the variance of student level residual is usually larger than measurement error variance at each time point, I attributed 22% of total variance of test scores at the end of Grade1 to r_{0i}, and 10% of the total variance to e_{0i}.

Table 3.2

Parameter values of fixed effects, variance components, and correlations

Parameter		Value
	School Level Fixed Effect	
β_{00}	Grand mean of intercept	0.000
γ_{01}	Effect of school SES	-0.414
	School Level Random Effect	
U_{0j}	Variance of School effect on intercept (school selection effect) (τ_{00})	0.100
U_{tj}	Variance of value-added school effectiveness at time t (τ_{tt})	0.317
	School Level Predictor	
SES_j	Mean and variance of School SES	0.000 (1.000)
	Student Level Fixed Effect	
β_{01}	Effect of kindergarten test score on intercept	0.762
β_{10}	Grand mean of natural growth rate	0.630
	Student Level Random Effect	
r_{0i}	Variance of student random effect on intercept	0.220
r_{1i}	Variance of natural growth rate	0.325
	Student Level Predictors and Outcomes	
$Kinder_{ij}$	Mean and variance of student kindergarten test scores	0.000 (1.000)
Y_{0ij}	Mean and variance of student test scores at Grade1 (time=0)	0.000 (1.000)
	Correlations	
r_{01}	Correlation between intercept and natural growth rate	-0.210
$r_{ses.kinder}$	Correlation between school SES and school mean kindergarten test score	-0.930
$r_{tt'}$	Correlation of value-added school effectiveness over time	0.600
r_{kinder}	Intraclass correlation of kindergarten test score	0.210

The school level variance of student test scores at the end Grade1 was determined based on the meta-analysis of Bosker and Witziers (1995). They found that in the United

States, without adjustment, school-level factors explained about 21% of the total variance of student achievement at a single time point; after adjusting for student intake characteristics, school-level factors accounted for 10% of the total student achievement variance at a single time point.

In the current study, students' test scores at the end of kindergarten and at the end of Grade 1 were assumed to be standardized. Given the variance of Grade 1 test scores accounted for by the kindergarten test scores, the regression coefficient for the kindergarten test score can be determined, and the residual variance at each level can also be determined for generating student test scores at the end of Grade 1.

According to the CCCM model, test scores in later years are the sum of student true score in the previous year plus achievement growth plus error. In each year, the overall achievement growth of a student can be divided into two parts. One is the natural growth given no school effectiveness exist. This growth can occur because of natural maturity or other environment influences. Thus, this part of growth is called natural growth. CCCM assumes the natural growth is linear. The other part of annual growth is due to school. By attending a school, a student may gain above and beyond natural growth in a given year. In order to generate test scores at later grades, students' natural growth and growth due to school must be partitioned. Previous research found that most of the variance of student annual growth lied among schools. In a study with a similar model as the CCCM in the current study, Rowan, Currenti and Miller (2002) found that 72% to 73% of the reliable variance in achievement growth in reading lied among schools. In a study with traditional three-level HLM model, Raudenbush (1989) found that 80% of growth variance in math and 43.9% in reading were between schools.

Mortimore et al. (1988) found that 30% of the variance of learning progress from Grade 1 to Grade 3 was between schools. Based on these studies, the simulation posited that 60% of the variance of the overall growth in a year was due to school, and 40% was due to students' natural growth. Thus, the variance of growth due to school was 1.5 times of the variance of growth due to students' natural growth. In the current study, the distribution of students' natural growth rate was adopted from the study of Ponisciak and Bryke (2005). They used an unconditional cross-classified model to estimate school effectiveness from Grade 2 to Grade 8. This study included 388,000 students in 500 Chicago Public elementary schools. In their study, the natural growth rate had a mean of 0.63 and a variance of 0.325. Thus, variance due to school was 0.487 (i.e., 0.325*1.5=0.487), and the total variance of growth was 0.812 (i.e., 0.325+0.487=0.812).

Based on literature (Teddlie & Stringfield, 1993; Willms, 1987), 35% of the between school variance of growth (i.e. 0.487*0.35 = 0.169) can be explained by the difference in school SES. Given school SES had a standardized normal distribution, and the total variance of student annual growth was 0.812, the fixed effect of school SES on student growth was -0.414. The effect was negative because school SES is usually measured by the percentage of students eligible for reduced or free lunch.

The correlation between student residuals on the intercept (r_{0i}) and the natural growth rate (r_{1i}) was set at -0.21. This was done based on the study by Ponisciak and Bryk (2005). Because school SES usually correlates with school aggregated prior attainment, the currently study considered this correlation in data simulation. The correlation between school SES and school mean kindergarten test score was set at -0.93 based on the study by Darandari (2004). The study by Darandari (2004) was conducted on two years of test scores of 3,992 students in 24 schools. The school size ranged from 81 to 276.

The correlations among value-added school effectiveness over time were also taken into account in data simulation. Review of stability of school effectiveness over time suggested that there was a fair degree of stability in secondary schools' effect on the overall measures of academic achievement over time. The same trend was evident for basic skill areas in the primary schools, though correlations were lower (Teddlie & Reynolds, 2000). The study of Willms (1987) found that correlations of school effectiveness over years ranged from 0.6 to 0.8. Bosker and Scheerens (1989) reported the correlations in Netherland were from 0.75 to 0.96. Although Mandevill (1988) reported that the correlations in the USA ranged from 0.34 to 0.66, Bosker and Scheerens (1989) pointed out that these figures might be deflated because of inadequacy of statistical control of intake characteristics. Based on these previous studies of the stability of school effectiveness over time, the correlations among school effectiveness over time were set at 0.60 in the current study.

In addition to the parameters specified in the model, the intraclass correlation of kindergarten test scores was also considered in data generation. According to the meta-analysis of Bosker and Witzier (1995), without any adjustment, the proportion of between school variance to the total variance of student test scores at a single time was about 0.21 in the United States. Thus, the intraclass correlation of kindergarten test scores was set at 0.21 in the current study.

Simulation Procedure

The simulation procedure can be divided into three parts. The first part involved generating student test scores from the end of Grade 1 to the end of Grade 5 with respect to the parameter values and the simulation conditions. The second part involved estimating the value-added school effectiveness in Grade 5 with each of the four Value-Added Models. The third part involved obtaining school rankings and classifications based on the school effectiveness estimates in the four models, and comparing the school rankings and classifications. For each combination of the simulation conditions, the simulation procedure was repeated 200 times. The sequence of the simulation procedure was as follows:

A. Data Generation

A.1. Generate data at the school level:

A.1.1. Sample the values of the value-added school effectiveness in Grades 2, 3, 4, and 5 (i.e., u_1 to u_4) from a multivariate normal distribution, with univariate means of 0, univariate variances of 0.317, and all the bivariate correlations of 0.6.

A.1.2. Sample the values of school effect on the intercept (i.e., u_0) from a normal distribution with a mean of 0 and a variance of 0.1.

A.1.3. Sample school means of kindergarten test scores (i.e. kinder_mean) from a normal distribution with a mean of 0 and a variance of 0.21. This variance was determined according to the intraclass correlation of kindergarten test scores.

A.1.4. Calculate school SES values based on the equation

SES= (-2.02)* kinder_mean + 0.3764 * rannor(-5), given that school SES had a standardized normal distribution, and school SES was correlated with school mean kindergarten test score as -0.93. “Rannor ()” is the SAS random number generation function for the standardized normal distribution.

A.2. Generate data at the student level:

A.2.1. Sample 5 variables (i.e. e_1 to e_5) from 5 univariate normal distributions with means of 0 and variances of 0.1. Each variable represents the measurement error at each grade.

A.2.2. Calculate student kindergarten test scores based on the equation

kinder = kinder_mean + 0.89 * rannor(-5),

given the intraclass correlation of kindergarten test score was 0.21, and student kindergarten test score had a normal distribution with a mean of 0 and a variance of 1.

A.2.3. Sample values of student random effect on the intercept (i.e., r_{0i}) from a normal distribution with a mean of 0 and a variance of 0.22.

A.2.4.Calculate values of students’ natural growth rates (i.e., r_{1i}) based on the equation

r_1=0.63 + (-0.2552)*r_0+0.5573*rannor(-10),

given that r_1 has a normal distribution with a mean of 0.63 and a variance of 0.325, and the correlation between r_1 and r_0 was -0.21.

A.2.5. Calculate student test scores at the end of Grade 1 based on the equation as

Grade 1=0.7616*kinder+u_0+r_0+$e_{1,}$

given that student test score at Grade 1 had a standardized normal distribution.

A.2.6. Generate student test scores at Grade 2, 3, 4, and 5. The test score of a student at each grade is the error free test score at the previous grade plus natural growth rate plus the effect of school SES on achievement growth plus school value-added effectiveness during that year plus measurement error at that grade. For example, the equation for generating student test scores at Grade 2 was:

$$\text{Grade 2=grade1\_True+}r_1\text{-0.414*SES+}u_1\text{+}e_1\text{;}$$

B. Estimation of value-added school effectiveness in different models.

B.1. Create a dataset that includes student test scores from Grade 3 to Grade 5, the time variable, and all the associated covariate variables.

B.2. Transform the dataset in B.1 from wide format to long format. In wide format, each student has only one recode, and test score at each time point takes up one variable. In long form, each student has multiple records, and test score at each time point takes one record (Doran, 2003; Singer, 1998). UCCM and LMEM require long format of data to estimate school effectiveness. The SAS code for transforming the dataset from wide format to long format is:

```
DATA d_long; set d_wide;
  array tvar(3) time3-time5;
  array scorevar(3) grade3 grade4 grade5;
    do i=1 to 3;
```

```
      time=tvar(i);
      score=scorevar(i);
      grade=time+1;
      output;
    end;
  drop i time3-time5 grade3 grade4 grade5;
RUN;
```

B.3. Create the value-added school effectiveness design matrix. In order to estimate value-added school effectiveness with UCCM and LMEM, the design matrix of value-added school effectiveness in the two models must be created by researchers themselves. The design matrix should reflect the accumulation character of value-added school effectiveness in the two models. Given 10 schools, the SAS code for creating the design matrix is:

```
DATA t5; SET d1;
      array z0(*) z0_1-z0_10;
      do s=1 to 10;
       if school=s then z0(s)=1; else z0(s)=0;
      end;
      array z1(*) z1_1-z1_10;
      do s=1 to 10;
       if (school=s and time=3) or (school=s and time=4)then z1(s)=1;
       else z1(s)=0;
      end;
```

```
    array z2(*) z2_1-z2_10;
     do s=1 to 10;
      if school=s and time=4 then z2(s)=1;
      else z2(s)=0;
     end;
    keep id school time score grade z0_1-z0_10 z1_1-z1_10 z2_1-z2_10;
RUN;
```

The resulted design matrix is similar to the z variables in Table 3.1.

B.4. Estimate value-added school effectiveness with the four Value-Added Models, respectively. The SAS code for specifying and estimating the four Value-Added Models were presented in section 1 of this chapter.

C. Compare school rankings and classifications.

C.1. For each estimation model, obtain the estimated school rankings based on the school effectiveness estimates. Larger school effectiveness had higher rank. Furthermore, obtain the known true school rankings based on the generated values of value-added school effectiveness.

C.2. Compare the school rankings.

C.3. For each estimation model, obtain the estimated school classifications based on the school effectiveness estimates and the classification criteria. The schools with effectiveness estimates 1 SE below the mean were classified as ineffective schools; the schools with effectiveness estimates 1 SE above the mean were classified as effective schools; other schools were classified as average. Similarly, obtain the known true school classifications based on the generated values of value-added school effectiveness.

C.4. Compare the school classifications.

Statistical Analysis

The current study focused on two main interests. One was to compare the performance of different Value-Added Models in school rankings. The other was to compare the performance of different Value-Added Models in school classifications.

The pairwise comparisons of school rankings and classifications in the current study can be classified into three series. The first series of comparisons evaluated the consistency of the estimated school rankings and classifications versus the known true school rankings and classifications. Specifically, in the first series, four comparisons were conducted. They were: (1) school rankings and classifications based on the Gain_kindergarten model versus the known true school rankings and classifications, (2) school rankings and classifications based on the Gain_grade4 model versus the known true school rankings and classifications, (3) school rankings and classifications based on the UCCM model versus the known true school rankings and classifications, (4) school rankings and classifications based on the LMEM model versus the known true school rankings and classifications. In the second series of comparisons, the school rankings and classifications based on the Gain_kindergarten model were compared to the school rankings and classifications based on the Gain_grade4 model. In the third series of comparisons, the school rankings and classifications based on the UCCM were compared to the school rankings and classifications based on the LMEM.

The research question answered by each pairwise comparison was:

1. Under the typical situations in elementary school effectiveness research, can a gain score model adjusting for kindergarten test scores validly recover the true school rankings or classifications in Grade 5?

2. Under the typical situations in elementary school effectiveness research, can a gain score model adjusting for Grade 4 test scores validly recover school rankings or classifications in Grade 5?

3. Under the typical situations in elementary school effectiveness research, can we ignore the covariates but still achieve valid school rankings and classifications?

4. Under the typical situations in elementary school effectiveness research, if we model the correlations among test scores over time, can we ignore the covariates but still achieve valid school rankings and classifications?

5. Under the typical situations in elementary school effectiveness research, when a gain score model is used, can adjustment using test scores from the previous year achieve similar school rankings or classifications as adjustment using test scores collected before entering to a school?

6. Under the typical situations in elementary school effectiveness research, can estimation of the correlations among test scores over time alleviate or eliminate the damage caused by ignoring the covariates?

The statistic used to evaluate consistency of school rankings was Spearman rho^2. When investigating the effect of simulation conditions, rho^2 was the dependent variable. Furthermore, descriptive but not inferential ANOVA was used to evaluate the magnitude of the effects of the simulation conditions. The reason is that the null hypothesis test will almost always achieve statistically significance given the large number of replications in simulation studies. In descriptive ANOVA, the η^2 was used to evaluate the effect size of each factor and their interaction.

In order to compare the Value-Added Models for school classifications, the classification criterion is determined first. The current study adopted the classification criterion from the study of Ballou, Sanders and Wright (2004). In their study, the teachers with effectiveness estimates 1.5 SE below the mean were regarded as ineffective; the teachers with effectiveness estimates 1.5 SE above the mean were classified as effective; other teachers were classified as average. This criterion was also adopted by Thomas and Mortimore (1996). The current study used 1 SE instead of 1.5 SE as the cutoff value because the 1.5 SE resulted in zero frequencies in some cells of the contingency tables in the current study, which makes the calculation of Kappa coefficient impossible.

The statistic to evaluate agreement of classifications was the Kappa coefficient. This statistic controls for chance agreement expected from the distribution of the data and employs the table’s row and column totals in determining chance agreement. The general range of Kappa is +1.0 for perfect agreement downward to the point where agreement ratio equals chance agreement (Lang & Tedllier, 1992). The standardized Kappy coefficient is distributed like a z score. Thus, the standardized Kappy coefficient is also called as Kappa z coefficient. When z is larger than 2, the null hypothesis that the agreement of classifications is equal to the agreement by chance is rejected. Because the classifications based on different models are independent, Kappa z can be used for a null hypothesis test, and the Kappa coefficient is used for quantifying the degree of agreement.

Besides using Kappa z and Kappa to evaluate the overall agreement of school classifications, the source of disagreements was explored further. This was done by calculating the frequencies of various kinds of disagreement, so that, people can know what are the most frequent types of misclassifications when using a certain value-added model.

CHAPTER IV

RESULTS

The results of the present study were organized into three sections. The first section reports the results involving checking the simulation program and deciding the replication numbers. In particular, the first section presents the distribution of the simulated data, checking the SAS code for the estimation models, convergence rate and number of replications. The second section reports the results involving the consistencies of school rankings, arranged in the sequence of the six research questions. The third section reports the results involving the agreements of school classifications, also arranged in the sequence of the six research questions.

Program Checking and Replication Numbers

Data Generation Check

The process for generating data was presented in Chapter III. In order to confirm the accuracy of the generated data in representing the desired population distribution and the correlations, descriptive statistics for the variables in a large sample with 500 schools and 500 students per school are presented in Table 4.1. The correlations that were specified in data generation were also estimated in the large sample, and these results are presented in Table 4.2.

Table 4.1

Descriptive statistics of the variables in a large sample (NS = 500, NSS = 500)

		Mean	SD	Skewness	Kurtosis
U_0	School selection effect	-0.010[a] (0.0)[b]	0.312 (0.316)	0.076 (0.0)	-0.515 (0.0)
U_1	School value-added effect during the 2nd Grade	-0.015 (0.0)	0.577 (0.563)	-0.059 (0.0)	-0.224 (0.0)
U_2	School value-added effect during the 3rd Grade	-0.011 (0.0)	0.553 (0.563)	0.074 (0.0)	0.011 (0.0)
U_3	School value-added effect during the 4th Grade	-0.001 (0.0)	0.558 (0.563)	-0.053 (0.0)	-0.204 (0.0)
U_4	School value-added effect during the 5th Grade	0.041 (0.0)	0.589 (0.563)	-0.039 (0.0)	-0.059 (0.0)
$\overline{kinder}_j$	School mean of test scores at the end of kindergarten	-0.001 (0.0)	0.489 (0.46)	0.095 (0.0)	-0.181 (0.0)
SES_j	School SES	0.014 (0.0)	1.036 (1.0)	-0.123 (0.0)	-0.035 (0.0)
r0	Student intercept	0.001 (0.0)	0.468 (0.469)	0.004 (0.0)	0.012 (0.0)
r1	Student pre-exist growth rate	0.631 (0.63)	0.571 (0.57)	-0.002 (0.0)	0.010 (0.0)
kinder	Student test scores at the end of kindergarten	0.001 (0.0)	1.018 (1.0)	0.011 (0.0)	-0.010 (0.0)
Grade1	Student Grade1 test score	-0.010 (0.0)	0.999 (1.0)	0.0174 (0.0)	0.000 (0.0)
e1	Measurement error at Grade1	-0.001 (0.0)	0.316 (0.3162)	0.009 (0.0)	-0.001 (0.0)
e2	Measurement error at Grade2	-0.001 (0.0)	0.316 (0.3162)	0.005 (0.0)	0.008 (0.0)
e3	Measurement error at Grade3	0.000 (0.0)	0.316 (0.3162)	-0.002 (0.0)	0.016 (0.0)
e4	Measurement error at Grade4	0.001 (0.0)	0.316 (0.3162)	-0.002 (0.0)	0.023 (0.0)
e5	Measurement error at Grade5	0.000 (0.0)	0.316 (0.3162)	0.001 (0.0)	0.010 (0.0)

A. sample estimate

B. true value

Table 4.2

Correlations in the large sample (NS = 500, NSS = 500)

Correlation between school mean kindergarten score and SES	Correlations among U_1, U_2, U_3, U_4	Correlation between intercept and slope (R_0 Vs R_1)	Intraclass correlation of kindergarten test scores
-0.932[a] (-0.93)[b]	0.641[c]-0.610[d] (0.6)	-0.211 (-0.21)	0.239 (0.21)

A. sample estimate
B. true value
C. largest correlation
D. smallest correlation

As presented in Table 4.1 and Table 4.2, the distribution of each variable in the large sample closely approximated the desired population distribution. The correlations estimated in the large sample were also very close to the population values specified in data generation.

Estimation Models Check

The SAS codes for a two-level gain score model have been well documented (Littell, Milliken, Stroup, & Wolfinger, 1996; Singer, 1998).The SAS codes for the UCCM model and the LMEM model have not been well documented. To the knowledge of the author, only one published article presented the SAS code for LMEM model (Tekwe et al., 2004), and no published articles have presented the SAS code for UCCM model or CCCM model. However, R codes for both LMEM model and UCCM model were provided in a published article (Lockwood, Doran, & McCaffrey, 2003). The SAS code for the UCCM model in the current study was transformed from the corresponding R code. In order to confirm the accuracy of the transformation, the consistency of school effectiveness estimates between SAS UCCM code and R UCCM code was examined with the largest sample size conditions (i.e., NS=50, NSS=50). With the same estimation function (i.e., Restricted Maximum Likelihood Function) and the same optimization algorithm (i.e., Newton-Raphson iteration), the correlation between the SAS UCCM school effectiveness estimates and the R UCCM school effectiveness estimates was 0.9583. The correlation between the SAS LMEM school effectiveness estimates and the R LMEM school effectiveness estimates was also examined with a smaller dataset (i.e., NS=30, NSS=30). The smaller dataset was used because R had convergence problems when fitting LMEM model with the larger dataset. The correlation between the SAS LMEM school effectiveness estimates and the R LMEM school effectiveness estimates was 0.998. Thus, the SAS codes for specifying UCCM and LMEM were at least as appropriate as the published R codes.

Number of Replications

In order to decide the number of replications in each combination of the sample size conditions, a trial was conducted using 200, 500, 400, and 300 replications. If a larger number of replications can not obviously change the mean and SD of Spearman rho^2 between school rankings, smaller number of replications was used. Because less stable parameter estimates usually result from smaller sample size and a more complex model, the trial was conducted under the smallest sample size conditions (i.e., NS=10, NSS=10) and with the UCCM model. The means and SDs of Spearman rho^2 from 200 replications and 500 replications were very close to each other. Table 4.3 lists the means and SDs of the Spearman rho^2 between the UCCM school rankings and the true school rankings for 200 and 500 replications, respectively.

Table 4.3

Mean and SD of Spearman rho^2 between the UCCM school rankings and the known true school rankings for 200 and 500 replications

Replications	Mean of rho^2	SD of rho^2
200	0.588	0.215
500	0.565	0.212

Table 4.3 indicated that increasing the number of replications from 200 to 500 did not appreciably change the mean and SD of the sampling distribution of rho^2. Thus, the results of 200 replications in each cell were used in subsequent analysis. In order to achieve balanced design, each cell had 200 converged replications. Because of convergence problems encountered when fitting UCCM and LMEM under large sample size conditions, more replications were run under large sample size conditions to achieve the 200 converged replications.

Model Fitting and Convergence Rates

Because the optimization algorithm built in SAS PROC MIXED (i.e., Newton-Raphson iteration) needs to invert the covariance matrix of the random effects, when the number of random effects was large, the computation load may not be handled by a PC with limited RAM. Even when a PC has 2 Gb RAM, the SAS code for the UCCM model and the LMEM model still encountered a huge convergence problem when fitting the datasets that included a large number of students. This was expected because each student has a random intercept and slope in the UCCM model, and the correlations among repeated measures nested within each student are estimated in the LMEM model.

For each level of the school size condition, 5 trials were run. The number of students per school was 50, 60, 80 and 100; and the number of schools 50. The try-out process was stopped whenever none of the 5 trials converged. In order to make the convergence process easier, initial values of the variance components for student and school level random effects were provided; and the measurement error variances were fixed. The maximum number of iterations was set at 100, and the convergence criterion was 10e-4. For both 10 schools and 50 schools, with 50 students per school, 3 out of 5 trials with the LMEM model achieved convergence. When the number of students per school was 50, UCCM took about 30 minutes to converge, and LMEM took about 45 minutes to converge. When the number of students per school was 60, none of the 5 trials converged. Thus, the trial process was stopped, and the largest number of students per school in the current study was determined to be 50.

With 50 schools and 50 students per school, the convergence rates were low for the CCCM and the LMEM. The convergence rates were calculated as:

Convergence Rate = 200 / total number of replication conducted

The convergence rate for different models under different sample size conditions are presented in Table 4.4.

Table 4.4

Convergence rates for different Value-Added Models

NS	NSS	UCCM	LMEM	Gain_Kindergarten	Gain_Grade4
50	50	0.733	0.504	1.000	1.000
50	10	1.000	1.000	1.000	1.000
10	50	1.000	0.691	1.000	1.000
10	10	1.000	1.000	1.000	1.000

School Ranking

Pairwise comparisons of school rankings based on different models were analyzed to answer six research questions: (1) whether a gain score model that adjusted for the test scores collected at the end of kindergarten (Gain_kindergarten) could recover the true school rankings in Grade 5; (2) whether a gain score model that adjusted for the test scores of Grade 4 (Gain_grade4) could recover the true school rankings in Grade 5; (3) whether the cross-classified model that ignores the covariates could recover the true school rankings in Grade 5; (4) whether the LMEM model that estimated correlations among repeated measurements over time could adjust for the covariates without specifying them in the model and recover the true school rankings in Grade 5; (5) when a gain score model was used, whether adjustment with the test scores of Grade 4 could achieve similar school rankings as adjustment with the test scores at the end of kindergarten; (6) whether estimating the correlations among repeated measurements could alleviate the problem caused by omitted covariates in estimating school rankings. It should be noted that the results were obtained under the typical situations found in school effectiveness research. Cautions need to be taken when generalizing the results to other situations.

Question 1 for School Rankings

This question was answered by comparing school rankings based on the Gain_kindergarten model with the known true school rankings in the simulated data. The means and SDs of Spearman rho^2 between the Gain_kindergarten school rankings and the true school rankings across 200 replications are listed in Table 4.5.

Table 4.5

Mean and SD of rho^2 between the Gain_kindergarten school rankings and the known true school rankings across 200 replications

NS[a]	NSS[b]	Mean	SD	Skewness	Kurtosis
50	50	0.947	0.028	-1.582	2.311
50	10	0.835	0.050	-1.035	2.766
10	50	0.854	0.148	-1.577	2.146
10	10	0.749	0.157	-1.138	1.297

A. NS represented Number of Schools
B. NSS represented Number of Students per School

When the number of schools and the number of students per school were both 50, the Gain_kindergarten school rankings achieved high agreement with the known true school rankings. The negative skewness indicates that most of the rho^2 were even larger than 0.947, because for negatively skewed data, the median is greater than the mean. In addition, the small SD indicated that the high consistency was stable across the 200 replications. When the number of schools was 50, with only 10 students per school, the Gain_kindergarten school rankings could still achieved stable and moderate agreement with the known true school rankings. For 10 schools with 50 students per schools, although Gain_kindergarten achieved moderate agreement, the agreement was not stable across the replications (SD = 0.148). For 10 schools with 10 students per schools, both the degree of agreement and its stability were low (Mean = 0.749, SD = 0.157).

In order to examine the effect of sample size conditions on the rho^2 between Gain_kindergarten school rankings and the known true school rankings, eta^2 for the main effects of NS, NSS, and their interaction effect on rho^2 were calculated, respectively. The eta^2 for the NS factor was 13.96%; the eta^2 for the NSS factor was 19.24%; and the eta^2 for the interaction between NS and NSS was 0.03%.

Although Table 4.5 indicated that both small number of schools and small number of students per school could decrease the agreement between Gain_kindergarten school rankings and the known true school rankings, the results might be limited by the sample size conditions. In order to investigate the effect of a wider range of sample size conditions, two small supplementary simulations were conducted. The first small simulation was aimed at answering two questions: (1) whether the moderate agreement between Gain_kindergarten school rankings and the known true school rankings caused by small number of schools can be improved by increasing the number of students per school; (2) whether the moderate agreement caused by small number of students per school can be improved by increasing the number of schools. In order to answer the first question, the number of schools was set at 10, and the number of students per school was set at 60, 80, and 100, respectively. In order to answer the second question, the number of students per school was set at 10; and the number of schools was set at 60, 80, and 100, respectively. The means and SDs of rho^2 across 200 replications are listed in Table 4.6

Table 4.6

Mean and SD of rho^2 between the Gain_kindergarten school rankings and the known true school rankings for other sample size conditions

NS	NSS	Mean	SD
10	60	0.847	0.157
10	80	0.864	0.141
10	100	0.874	0.123
60	10	0.840	0.038
80	10	0.847	0.036
100	10	0.847	0.031

Table 4.6 indicated that when the number of schools was small, it did not help to select more students per school to increase the ability of Gain_kindergarten to recover the

true school rankings. The ability of Gain_kindergarten to recover the true school rankings was also limited by the small number of students per school. When the number of students per school was small, no matter how many schools were ranked, the ability of Gain_kindergarten to recover the true school rankings could not be further improved. Table 4.6 also indicates that SD of the rho^2 was mainly influenced by the number of schools, not by the number of students per school. Other research about the influence of sample size on parameter estimates and the associated SEs in multilevel modeling has also found that a small number of second level units compromises the SEs of sample estimates, but not on the sample estimates themselves, and increasing the number of first level units can not eliminate the bias caused by small number of second level units (Kamali, 1992; Mass & Hox, 2004).

The second small simulation was conducted to investigate how many schools could be ranked validly with 50 students per school. With 50 students per school, 200 replications were conducted for each level of the NS factor, which were 20, 30, 40, 100, and 150, respectively. The means and SDs of rho^2 are listed in Table 4.7.

Table 4.7

Mean and SD of rho^2 between the Gain_kindergarten school rankings and the known true school rankings for different number of schools

NS	NSS	Mean	SD	Skewness	Kurtosis
20	50	0.909	0.073	-1.830	3.685
30	50	0.938	0.044	-2.438	8.584
40	50	0.938	0.038	-1.668	3.345
100	50	0.958	0.013	-1.415	2.747
150	50	0.959	0.013	-2.271	8.025

Table 4.7 indicates that when the number of students per school was 50, Gain_kindergarten model had high ability to recover the true school rankings even when

the number of schools was as small as 20. However, the improvement of the degree of agreement was not linear. There was a plateau phenomenon. When the number of schools was 100 or larger, the degree of agreement did not obviously increase.

Question 2 for School Rankings

This question was answered by comparing school rankings based on the Gain_grade4 model with the known true school rankings in the simulated data. The means and SDs of rho^2 between Gain_grade4 school rankings and the known true school rankings across 200 replications are listed in Table 4.8. Table 4.8 indicates that under all the sample size conditions, the agreements between the Gain_grade4 school rankings and the known true school rankings were low, and the SD was influenced by the number of schools but not by school size.

Table 4.8

Mean and SD of rho^2 between the Gain_grade4 school rankings and the known true school rankings across 200 replications

NS	NSS	Mean	SD	Skewness	Kurtosis
50	50	0.559	0.008	-0.122	-0.420
50	10	0.489	0.096	-0.058	-0.001
10	50	0.522	0.217	-0.430	-0.648
10	10	0.437	0.213	0.043	-0.742

In order to investigate whether the low ability of the Gain_grade4 model to recover school rankings was limited by the sample size conditions in the current study, and whether recovery can be improved by increasing the number of schools and school size, another supplementary simulation with 150 schools and 100 students per school was conducted. The mean across 200 replications was 0.499, and the SD was 0.058. Thus, Gain_grade4 had low ability to recover true school rankings even with larger sample sizes. The results for research question 2 suggested that gain score models that adjust for test scores from the previous year are not good choices for ranking schools with respect to the definition of value-added school effectiveness.

Question 3 for School Rankings

This question was answered by comparing school rankings based on the UCCM model with the known true school rankings under different sample size conditions. The means and SDs of Spearman rho^2 between the UCCM school rankings and the known true school rankings across 200 replications are listed in Table 4.9.

Table 4.9

Mean and SD of rho^2 between the UCCM school rankings and the known true school rankings across 200 replications

NS	NSS	Mean	SD	Skewness	Kurtosis
50	50	0.559	0.157	-1.714	3.093
50	10	0.569	0.096	-0.469	0.699
10	50	0.647	0.190	-0.655	-0.349
10	10	0.588	0.215	-0.682	-0.131

Table 4.9 indicated that under any of the sample size conditions in the current study, the rho^2 between the UCCM school rankings and the known true school rankings was about 0.6, corresponding to rho of 0.775 (i.e., $0.6^{0.5} = 0.775$). This suggests that the UCCM model could not accurately recover the true school rankings under the sample size conditions in the current study. In addition, no changing trend of the agreement level was observed along with the change in the sample size conditions. The results suggested that without other remedial strategies, given the effect size of the covariates in the current study, ignoring the covariates resulted in invalid school rankings with respect to the definition of value-added school effectiveness.

Question 4 for School Rankings

This question was answered by comparing school rankings based on the LMEM model with the known true school rankings under different sample size conditions. The means and SDs of Spearman rho^2 between the LMEM school rankings and the known true school rankings across 200 replications are listed in Table 4.10.

Table 4.10

Mean and SD of rho^2 between the LMEM school rankings and the known true school rankings across 200 replications

NS	NSS	Mean	SD	Skewness	Kurtosis
50	50	0.658	0.077	-0.847	2.232
50	10	0.642	0.082	-0.512	0.389
10	50	0.671	0.181	-0.718	-0.145
10	10	0.638	0.188	-0.842	0.596

Table 4.10 indicates that the LMEM school rankings could only achieve rho^2 of around 0.65 with the known true school rankings under all the sample size conditions in the current study. This suggested that LMEM does not have adequate ability to recover the true school rankings. Furthermore, no changing trend was observed on the degree of consistency along with the changes in the sample size conditions. The results suggest that the strategy of estimating the correlations among repeated measurements could not eliminate the negative effects of ignoring covariates on school rankings. Thus, the claim that each student can serve as his own blocking factor so that no explicit adjustment for covariates is needed (Sanders & Horn, 1994; Sanders, Saxon & Horn, 1997) is not valid for school rankings.

Question 5 for School Rankings

This question was answered by comparing the school rankings based on the Gain_kindergarten model and the Gain_grade4 model under different sample size conditions. The means and SDs of Spearman rho^2 between the Gain_kindergarten school rankings and the Gain_grade4 school rankings across 200 replications are listed in Table 4.11.

Table 4.11

Mean and SD of rho^2 between the school rankings based on the Gain_kindergarten model and the Gain_grade4 model across 200 replications

NS	NSS	Mean	SD	Skewness	Kurtosis
50	50	0.593	0.084	-0.106	-0.294
50	10	0.594	0.083	-0.425	-0.111
10	50	0.599	0.204	-0.66	-0.214
10	10	0.586	0.220	-0.488	-0.486

Table 4.11 indicates that the agreement between school rankings based on the Gain_kindergarten model and the Gain_grade4 model was low under any of the sample size conditions in the current study. This result was consistent with the studies of Sammons (1996), Cuttance (1985), and Preece (1989). They warned against adjustment of test scores that were approximate in time with the current test scores. They recommended adjusting test scores collected at the point of entry to school (e. g. the end of kindergarten for elementary schools). Considering the results of question1 that the Gain_kindergarten model had high ability to recover the true school rankings with enough schools and students per school, the Gain_kindergarten model is recommended for school rankings.

Question 6 for School Rankings

This research question was answered by comparing school rankings based on the UCCM model and the LMEM model. The results for research questions 4 and 5 for school rankings indicated that neither the UCCM model nor the LMEM model could adequately recover the known true school rankings. However, this research question evaluates whether estimating correlations among repeated measurements makes any difference in school rankings. The mean and SD of Spearman rho^2 between the UCCM school rankings and the LMEM school rankings across 200 replications are listed in Table 4.12.

Table 4.12

Mean and SD of rho^2 between the UCCM school rankings and the LMEM school rankings

NS	NSS	Mean	SD	Skewness	Kurtosis
50	50	0.874	0.014	-2.579	5.821
50	10	0.924	0.022	-0.761	0.580
10	50	0.996	0.004	-5.189	40.353
10	10	0.914	0.079	-2.476	8.939

Table 4.12 indicates that the school rankings based on the UCCM and the LMEM were highly consistent under all the sample size conditions in the current study, with rho^2 larger than 0.87 corresponding to rho larger than 0.93. Thus, estimating correlations among repeated measurements did not make big differences in school rankings versus ignoring the covariates.

Conclusions Regarding School Rankings

The Gain_kindergarten model accurately recovered the known true school rankings when the number of schools and the number of students per school were not too small. When the number of schools or the number of students was too small, increasing the sample size of the units at the other level did not improve the accuracy of school rankings. When the number of students per school was sufficiently large, such as 50, the model achieved high accuracy in ranking both large number of schools, such as 150, and relatively small number of schools, such as 20.

The other Value-Added Models (Gain_grade4, UCCM and LMEM) could not accurately recover the known true school rankings under all the sample size conditions in the current study. Furthermore, in the gain score models, adjustment using the grade 4 test scores resulted in different school rankings versus adjustment using the kindergarten test scores. In addition, ignoring covariates consistently invalidated school rankings no matter whether or not correlations among repeated measurements were estimated.

School Classification

Another important usage of value-added estimates of school effectiveness is to identify effective schools to accumulate successful educational strategies, or to identify ineffective schools that require help. The same set of pairwise comparisons was conducted to answer the six questions for school classifications instead of school rankings.

Question1 for School Classifications

Agreement between school classifications based on the Gain_kindergarden model and the known true school classifications was evaluated to answer this question. Table 4.13 reports the means and SDs of simple Kappa coefficients and Kappa Z coefficients across the 200 replications. It should be noted that the Kappa Z coefficient for each replication was calculated with the asymptotic SE but not the empirical SD of the Kappa coefficients.

Table 4.13

Agreement between the Gain_kindergarten school classifications and the known true school classifications across 200 replications

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
50	50	0.822	0.087	7.781	0.815
50	10	0.683	0.102	6.489	0.953
10	50	0.739	0.246	3.157	1.011
10	10	0.603	0.271	2.584	1.115

The mean Kappa Z coefficients in Table 4.13 indicate that for all the sample size conditions, the agreement between school classifications based on the Gain_kindergarten model and the known true school classifications was statistically significantly higher than the chance agreement. However, the simple Kappa coefficients indicate that for small sample size conditions, such as 10 schools with 10 students in each school, the agreement of school classifications was not high. When both the number of schools and the number of students were relatively large, such as 50 in the current study, the agreement of school classifications was larger than 0.80; and the SD was as small as 0.087. Thus, the answer to research question 1 about school classifications was that the Gain_kindergarten model could recover the true school classifications with a high degree of accuracy when the sample size conditions were adequately large, and its performance in school classifications was stable across repeated sampling.

In order to further examine the effect of the NS and the NSS factor on simple Kappa coefficient, eta^2 for the effect of the NS factor, the NSS factor, and their interaction were calculated, respectively. The eta^2 for the NS factor was 4.2%; the eta^2 for the NSS factor was 11%, and the eta^2 for their interaction was 0.001%.

A practical interesting question is whether the low level of agreement of school classifications caused by too small number of schools could be enhanced by increasing the number of students per school. For example, if a small school district has only 10 elementary schools, the district superintendent may wonder whether valid classifications of the schools could be achieved if more students in each school are selected. Another practical interesting question is whether the low level of agreement caused by a too small number of students per school can be improved by including more schools in the analysis. For example, if because of missing data only the data of 10 students per school could be used for analysis, a school accountability analyst may wonder whether school classifications are valid for statewide accountability, even if not valid for districtwide accountability. A small supplementary simulation was conducted to answer the two questions. In order to answer the first question, the number of schools was fixed at 10, and the number of students per school was set at 60, 80, and 100, respectively. In order to answer the second question, the number of students per school was fixed at 10; and the number of schools was set at 60, 80, and 100, respectively. The means and SDs of Kappa coefficients and Kappa Z coefficients across 200 replications are listed in Table 4.14.

Table 4.14

Agreement between the Gain_kindergarten school classifications and the known true school classifications when either NS or NSS was too small

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
10	60	0.759	0.247	3.244	0.997
10	80	0.754	0.242	3.212	0.985
10	100	0.753	0.235	3.197	0.951
60	10	0.675	0.105	7.017	1.076
80	10	0.673	0.087	8.068	1.033
100	10	0.684	0.075	9.172	1.008

Table 4.14 indicates that when the number of schools was too small, such as 10 in the current study, increasing the number of students per school did not increase the validity of school classifications based on the Gain_kindergarten model. Thus, Gain_kindergarten model is not appropriate for school classifications in a small school district. Table 4.14 also suggests that when the number of students per school was too small, such as 10, increasing the number of schools did not increase the validity of school classifications based on the Gain_kindergarten model. Thus, when too few students in each school have the required data, Gain_kindergarten model can not achieve valid school classifications no matter whether the classifications are districtwide or statewide.

Another small supplementary simulation was conducted in order to investigate how many schools could be validly classified with the Gain_kindergarten model given the number of students per school was 50. The number of students per school was fixed at 50, and the levels of the number of schools (NS) factor were set at 20, 30, 40, 100, and 150, respectively. Again, 200 replications were run under each level of the NS factor. The Means and SDs of simple Kappa coefficients and Kappa Z coefficients are reported in Table 4.15.

Table 4.15

Agreement between the Gain_kindergartern school classifications and the known true school classifications for different number of schools across 200 replications

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
20	50	0.784	0.155	4.701	0.884
30	50	0.807	0.113	5.915	0.803
40	50	0.803	0.107	6.799	0.887
100	50	0.836	0.058	11.190	0.771
150	50	0.844	0.046	13.820	0.742

Given each school having data for 50 students, the Gain_kindergarten model could validly recover school classifications for more than 30 schools. Although more schools resulted in more valid classifications, the improvement of validity of the classifications was not obvious when the number of schools was larger than 30.

In order to further explore the source of misclassifications, the maximum frequencies of various kinds of misclassifications in the 200 replications are reported in Table 4.16. Serious misclassifications were defined as misclassifying effective schools as ineffective or ineffective schools as effective. Table 4.16 indicates that most misclassifications were non-serious misclassifications that misclassified schools across two near categories. Among the non-serious misspecifications, no specific pattern was observed. Only one serious misclassification was observed. This misclassification happened when both the number of schools and the number of students per school were as small as 10. This serious misclassification classified an effective school as ineffective.

Table 4.16

Maximum frequency of misclassifications for the Gain_kindergarten model across 200 replications

NS	NSS	True classification	Gain_kindergarten classification	Maximum Frequency
50	50	1[a]	0[b]	4
		1	-1[c]	0
		0	1	4
		0	-1	5
		-1	0	4
		-1	1	0
50	10	1	0	7
		1	-1	0
		0	1	6
		0	-1	5
		-1	0	6
		-1	1	0
10	50	1	0	2
		1	-1	0
		0	1	2
		0	-1	2
		-1	0	2
		-1	1	0
10	10	1	0	2
		1	-1	1
		0	1	3
		0	-1	2
		-1	0	3
		-1	1	0

A.“1” represents effective schools.
B. “0” represents average schools.
C. “-1” represents ineffective schools.

Question 2 for School Classifications

This research question was answered by comparing school classifications based on the Gain_grade4 model and the known true school classifications. The means and SDs of simple Kappa coefficients and Kappa Z coefficients are presented in Table 4.17.

Table 4.17

Agreement between the Gain_grade4 school classifications and the known true school classifications across 200 replications

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
50	50	0.418	0.129	3.988	1.232
50	10	0.394	0.123	3.755	1.172
10	50	0.418	0.293	1.814	1.238
10	10	0.363	0.318	1.572	1.350

Table 4.17 indicates that under all the sample size conditions in the current study, the agreements between the school classifications based on the Gain_grade4 model and the known true school classifications were around 0.4. Although the Kappa Z coefficients indicated statistically significant different from chance agreement, the level of agreement of 0.4 was not high enough to indicate adequate agreement. Thus, the answer to research question 2 about school classifications was that the Gain_grade4 model could not validly recover the known true school classifications.

In order to investigate whether the low degree of agreement was limited by the sample size conditions in the current study, a small supplementary simulation with 150 schools and 100 students per school was conducted. The mean and SD of Kappa coefficient across 200 replications were 0.446 and 0.073, respectively.

Maximum frequencies of various kinds of misclassifications across the 200 replications are reported in Table 4.18. For Gain_grade4 model, most of the misclassifications were non-serious misclassifications. No specific pattern was observed in the non-serious misclassifications. There were seven serious misclassifications across the sample size conditions, which was higher than that for the Gain_kindergarten model. Furthermore, the serious misclassification happened even when the sample sizes were relatively large (i.e., NS=50, NSS=50).The serious misclassification for the Gain_grade4 model had no specific pattern.

Table 4.18

Maximum frequencies of misclassifications for the Gain_grade4 model across 200 replications

NS	NSS	True classification	UCCM classification	Maximum Frequency
50	50	1	0	9
		1	-1	1
		0	1	8
		0	-1	8
		-1	0	7
		-1	1	1
50	10	1	0	9
		1	-1	1
		0	1	9
		0	-1	7
		-1	0	7
		-1	1	1
10	50	1	0	3
		1	-1	0
		0	1	3
		0	-1	3
		-1	0	3
		-1	1	1
10	10	1	0	2
		1	-1	1
		0	1	2
		0	-1	3
		-1	0	3
		-1	1	1

Question 3 for School Classifications

The agreement between school classifications based on the UCCM model and the known true school classifications was examined to answer this research question. The means and SDs of simple Kappa coefficients and Kappa Z coefficients across 200 replications are presented in Table 4.19.

Table 4.19

Agreement between the UCCM school classifications and the known true school classifications across 200 replications

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
50	50	0.448	0.148	4.315	1.337
50	10	0.446	0.123	4.252	1.169
10	50	0.099	0.167	0.431	0.731
10	10	0.433	0.298	1.872	1.276

Table 4.19 indicates that under all the sample size conditions in the current study, the school classifications based on the UCCM model did not achieve high level of agreement with the known true school classifications. Thus, ignoring covariates might result in invalid school classifications.

An interesting result was that when the number of schools was too small, such as 10, increasing the number of students per school resulted in worse classifications. This finding corresponded to the results reported by Kamali (1992) who found that if there were biases in estimates of level-2 parameters, increasing the number of individuals per group even further increased the biases.

Maximum frequencies of various kinds of misclassifications are reported in Table 4.20. Table 4.20 indicates that most of the misclassifications were non-serious. However, the frequency of non-serious misclassification was not negligible. When the number of schools was 50 and the number of students per school was 10, the number of ineffective schools that were misclassified as average was even larger than the number of ineffective schools that were correctly classified. The serious misclassifications could happen even when the sample sizes were relatively large. In addition, when the number of schools was 10 and the number of students per school was 50, there were 3 serious misspecifications in which effective schools were classified as ineffective.

Table 4.20

Maximum frequencies of misclassifications for the UCCM model across 200 replications

NS	NSS	True classification	UCCM classification	Maximum Frequency
50	50	1	0	9
		1	-1	1
		0	1	8
		0	-1	7
		-1	0	10
		-1	1	1
50	10	1	0	8
		1	-1	1
		0	1	8
		0	-1	6
		-1	0	7
		-1	1	1
10	50	1	0	2
		1	-1	3
		0	1	3
		0	-1	2
		-1	0	2
		-1	1	2
10	10	1	0	2
		1	-1	1
		0	1	2
		0	-1	3
		-1	0	3
		-1	1	1

Question 4 for School Classifications

In order to answer this research question, the agreement between school classifications based on the LMEM model and the known true school classifications was examined. The means and SDs of Kappa coefficients and Kappa Z coefficients are reported in Table 4.21

Table 4.21

Agreement between the LMEM school classifications and the known true school classifications across 200 replications

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
50	50	0.503	0.131	4.791	1.235
50	10	0.498	0.122	4.746	1.155
10	50	0.524	0.298	2.289	1.278
10	10	0.490	0.276	2.125	1.165

Although the Kappa Z coefficients indicated that the levels of agreement were statistically significantly different from chance agreement, the Kappa coefficients of about 0.5 indicated that the levels of agreement between the LMEM school classifications and the known true school classifications were low for all the sample size conditions. Thus, the answer to research question 4 about school classifications was that LMEM could not accurately recover the true school classifications. This suggests that even if the correlations among repeated measurements were considered, a model ignoring covariates could not recover the true school classifications.

Further exploration of the source of misclassifications was conducted. Maximum frequencies of various misclassifications across 200 replications are reported in Table 4.22. For LMEM, there were serious misclassifications under all the sample size conditions. There was no specific pattern observed for both non-serious misclassifications and serious misclassifications.

Table 4.22

Maximum frequencies of misclassifications for the LMEM model across 200 replications

NS	NSS	True classification	UCCM classification	Maximum Frequency
50	50	1	0	7
		1	-1	1
		0	1	7
		0	-1	7
		-1	0	7
		-1	1	1
50	10	1	0	8
		1	-1	1
		0	1	7
		0	-1	9
		-1	0	8
		-1	1	1
10	50	1	0	8
		1	-1	1
		0	1	2
		0	-1	2
		-1	0	2
		-1	1	0
10	10	1	0	3
		1	-1	0
		0	1	3
		0	-1	2
		-1	0	3
		-1	1	1

Question 5 for School Classifications

This research question was answered by comparing the school classifications based on the Gain_kindergarten model and the Gain_grade4 model under different sample size conditions. The means and SDs of Kappa coefficients and Kappa Z coefficients across 200 replications are listed in Table 4.23.

Table 4.23

Agreement between the Gain_kindergarten school classifications and the Gain_grade4 school classifications

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
50	50	0.443	0.125	4.224	1.190
50	10	0.467	0.114	4.456	1.073
10	50	0.459	0.280	1.991	1.192
10	10	0.442	0.327	1.912	1.392

The Kappa coefficients in Table 4.23 indicate that the school classifications based on the Gain_grade4 model had low levels of agreement with the school classifications based on the Gain_kindergarten model. Considering the fact that Gain_kindergarten model could validly recover the true school classifications given enough number of schools and number of students per school, the Gain_kindergarten model was preferred over the Gain_grade4 model for school classifications.

Question 6 for School Classifications

This research question was answered by comparing school classifications based on the UCCM model and the LMEM model. The answer to research questions 3 and 4 for school classifications indicated that neither the UCCM model nor the LMEM model could accurately recover the school classifications in the known true values of school effectiveness. However, whether estimating correlations among repeated measurements can make any differences in school classifications remains unknown. The agreements between school classifications based on the UCCM model and the LMEM model are reported in Table 4.24.

Table 4.24

Agreement between the school classifications based on the UCCM model and the LMEM model

NS	NSS	Mean of Kappa	SD of Kappa	Mean of Kappa Z	SD of Kappa Z
50	50	0.864	0.092	8.204	0.690
50	10	0.780	0.098	7.393	0.911
10	50	0.297	0.146	1.216	0.962
10	10	0.782	0.218	3.329	0.880

Table 4.24 indicates that when the number of schools was relatively large, such as 50, the LMEM model consistently achieved moderately high degree of agreement with the UCCM model in school classifications. Thus, given that sample size was adequately large, only estimating correlations among repeated measurements does not make any obvious difference versus omitting covariates in school classifications.

Conclusions Regarding School Classifications

Among the four Value-Added Models, the Gain_kindergarten model is recommended for school classification, because only the Gain_kindergarten model achieved high levels of agreement with the known true values of school effectiveness for school classification. Furthermore, the high levels of agreement were stable across repeated sampling when the sample size conditions were sufficiently large. For Gain_kindergarten model, the frequencies of various kinds of misclassifications were small and most of the misclassifications were non-serious. When sample sizes were large enough, no serious misclassifications were observed.

However, the agreements between school classifications based on other Value-Added Models and the known true values of school effectiveness were low even when the sample sizes were relatively large. The frequencies of misclassifications were not negligible. Even when the sample sizes were relatively large, serious misclassifications were still observed.

CHAPTER V

CONCLUSIONS AND DISCUSSION

This chapter first summarizes the main message of the present study, and addresses the implications of the findings for school accountability practices. Second, this chapter integrates the results with prior studies in the literature. Third, this chapter addresses the limitations of the present study and offers suggestions for future studies.

Conclusions and Implications for Practice

The main message of the present study is that the Gain_kindergarten model can recover the true school rankings and classifications with a high degree of accuracy. However, other Value-Added Models, including the Gain_grade4 model, the UCCM model, and the LMEM model can not accurately recover the true school rankings and classifications. But the good performance of the Gain_kindergarten model in recovering school rankings and classifications was not exhibited when either the number of schools or the number of students per school was too small, such as 10. This problem could not be remedied by increasing the sample size at the other level (i. e., the number of students if schools were few, or the number of schools if students were few). On the other hand, with enough students per school, such as 50, the Gain_kindergarten model could rank more than 20 and classify more than 30 schools with high levels of accuracy. Thus, small school districts that have less than 20 elementary schools should be caution when using the Gain_kindergarten model to rank or classify schools. The small school districts take up 97.3% of the school districts around the States (NCES 2003-2004 Public Elementary/Secondary Universe Survey Data). Thus, in the United States, the

Gain_kindergarten model is more appropriate for state-wide school accountability than for district-wide accountability. Alternative statistical models and estimation methods that can result in valid school rankings and classifications with small number of schools will be valuable for most of the school districts in the United States.

In order to evaluate the effectiveness of elementary schools on students' annual gains, the current study suggests measuring student achievement level at the point of entry to an elementary school (e. g., the end of kindergarten). This measure of prior achievement should be adjusted in evaluating school unique contributions to students' gains in a given subsequent year (e. g., Grade 5). Compared with keeping students' test scores over several successive years, this strategy of using test scores at the point of entry to school for adjustments is more economic and results in less missing data. Another advantage of this strategy is that test scores from two successive years are more likely to measure the same psychological construct than test scores over a wide span of time (Martineau, 2006). The measurement invariance makes vertical equating of test scores meaningful.

The failure of the Gain_grade4 model in recovering school rankings and classifications and its difference with the Gain_kindergarten model implies that adjustment with test scores from the previous year is not appropriate for school rankings and classifications and could not replace the adjustment with test scores at the point of entry to school. This finding warns against the popular practice of adjusting test score from the previous year in gain score modeling.

The failure of the UCCM model in school ranking and classification implies that the student level and the school level covariates with effect sizes similar to those in the current study should be adjusted for; otherwise school rankings and classifications will be invalid. The failure of the LMEM model and the similarity of school rankings and classifications between the LMEM model and the UCCM model imply that estimating the correlations of repeated measurements within a student can not adjust for the covariates without specifying the covariates in the model. Considering these results, explicitly specifying the covariates in the statistical models is strongly recommended.

It should be noted that these conclusions were limited by the constraints in the current study. In particularity, the constraints in the current study included: (1) the effect sizes of the covariates in the current study were moderately large, with student kindergarten test scores explaining 58% of the total variance of student test scores at the end of Grade 1, and the school SES explaining 35% of the between school variance of annual growth; (2) the intake test scores were collected five years before the grade in which the school effectiveness was estimated; (3) the model that was used to generate test scores assumed that each student had a natural growth even if no school effectiveness existed, and the natural growth was linear; (4) the intraclass correlation of the kindergarten test scores was 0.21; (5) the total variance of growth in a year was 0.812; (6) for UCCM and LMEM, the sample size conditions didn't include large number of schools or large number of students per school, such as more than 100; (7) The test scores at different grades represent a single construct or the combined constructs with unchanging proportions across grades. These constraints might limit the results of the current study from generalizing to other situations. However, as previously explained in considerable detail, these simulation parameters were derived from prior research.

Explanation of the Results

The present study found that the accuracy of school rankings and classifications based on value-added estimates of school effectiveness depends upon the choice of prior measurements. The gain score model that adjusted for the test score at the end of kindergarten recovered the true school rankings and classifications in Grade 5, however, the gain score model that adjusted for the test scores at end of Grade 4 did not recover the known true school rankings or classifications in Grade 5.

This difference between the Gain_kindergarten model and the Gain_grade4 model in school rankings and classifications is expected according to the independence assumption of regression. The independence assumption of regression requires that the predictor variables should be independent with the residuals. In gain score models, value-added school effectiveness is the residual. Because students' test scores at the end of kindergarten are collected before they enter an elementary school, so the kindergarten test scores are independent from school effectiveness. On the contrary, because the test scores at the end of Grade 4 are influenced by both the test score at the end of kindergarten and the effectiveness of the schools the students attended in previous years, the Gain_grade4 model actually adjust for both the kindergarten test scores and the effectiveness of the schools in previous years. Because the effectiveness of the schools in previous years is correlated with their effectiveness in the current year, this violates the assumption of independence between predictors and residuals. Dependence between predictors and residuals partitions out some of school effectiveness in the current year and leads to lower estimated variance of school effectiveness in the current year. The reduced variance may distort school rankings and classifications.

This finding was supported by the research of Cuttance (1985). Cuttance noted that in many studies, the measurement of background factors was carried out at around the same time as the outcomes are measured. This predictor could very well be affected by the type of school attended, and it seemed unwise to partial out such a factor in a study of school effects. Peece (1992) also pointed out that the background factors used as predictors should be measured at the beginning of the period of instruction. In the United Kingdom, Sammons (1996) reported that control of GCSE scores at age 16 and ability at age 17 was likely to lead to reduced estimates of departmental differences at A-level testing that took place at age 18. GCSE and arguably ability scores were themselves likely to have been influenced by earlier secondary school (or departmental) effects.

The arguments against the use of the gain score model concerned the reliability and validity of difference scores as measures of growth. In terms of validity, some researchers argued that difference scores can not measure the shape of growth curve. However, deficiency in measuring the shape of growth curve does not influence whether difference scores are valid measurements of the amount of change (Rogosa, 1995). In terms of reliability, a common perception is that difference scores are intrinsically unreliable (Lord, 1956). However, Rogosa (1995) argued that the low reliability of difference scores was the artificial effect of the assumptions in some studies. These studies assumed equal reliabilities $\rho(X_1)= \rho(X_2)$, and equal variances $\sigma^2_{x1}=\sigma^2_{x2}$, for the observed scores at Time 1 and Time 2. These assumptions imply equal true score variances at Time 1 and Time 2 and a negative correlation between true change and true initial score. Furthermore, these studies assumed a high correlation between the test scores at Time 1 and Time 2. All these assumptions imply that growth curves are nearly parallel which translates into almost no individual differences in true change. If there are almost no individual differences in growth, the low reliability of the difference scores should be no surprise. In an empirical study, Rogosa and Willett (1983) showed that when the correlation between scores on Time 1 and Time 2 was moderate, even with other constraints, the reliability of difference scores could be as high as 0.94, which was nearly as reliable as the separate measurements on Time 1 or Time 2.

Another argument against the use of the gain score model is based on the idea of regression toward the mean. Intuitively, regression toward the mean says that on the average you are going to be closer to the mean at Time 2 than you were at Time 1 if you were far from the mean at Time 1. If regression toward the mean is real, the change of student test scores may be an artifact of regression and not due to educational interventions. It is a common belief in the research community that regression toward the mean is unavoidable as long as test scores at Time 1 and Time 2 are not perfectly correlated (Doran, 2003). Rogosa (1995) criticized this belief. He argued that the formal statement of regression toward the mean in the literature defined it in standard deviation units. A more realistic definition of regression toward the mean should use the actual metric of measurement. Regression toward the mean in actual metric pertains only when the correlation between change and initial status is negative or the true score variances at two time points are equal. Rogosa (1995) argued that correlation between change and initial status was not necessarily negative. The correlation could be positive, zero, or negative, depending on the initial time of measurement. The current study also found that the initial time of measurement was important for achieving valid school rankings and classifications.

Although Gain_kindergarten model has the advantages mentioned previously and may avoid the dilemma of low reliability and the problem of regression toward the mean, caution also need to be taken when using the Gain_kindergarten model to rank or classify schools. When using the Gain_kindergarten model, researchers need to check three things. One is to check whether the test scores measure the same thing at different observation occasions. This assumption is likely to be violated when there is a long interval between two measurements, or the two measurements are conducted during a period of rapid development. Second is to check whether the variance of difference scores is large enough to induce high reliability in measurement of growth. Third is to check the correlation between the initial measurement and the growth rate and how serious the problem of regression toward the mean is.

Another finding of the current study was the impact of sample size conditions on the performance of the Gain_kindergarten model in school rankings and classifications. The current study found that the poor performance of Gain_kindergarten model caused by inadequate sample size at student level or school level could not be improved by increasing the sample size at the other level. This was consistent with other studies about the effect of sample size on the parameter estimates in a two-level HLM model (Busing, 1993; Kamali, 1992; Mass & Hox, 2004). In addition, the current study found that with 50 students per schools, 20 schools were enough for the Gain_kindergarten model to achieve valid school rankings and classifications. The study of Mass and Hox (2004) also suggested 20 groups as a rule of thumb for achieving accurate parameter estimates in a two-level HLM model. The study of Brown and Draper (2000) and the study of Snijders and Bosker (1999) suggested using more than 10 second level units. These studies provided similar guidance about the second level sample size as the current study. However, Busing (1993) recommended using more than 100 groups to achieve accurate parameter estimates. Mass and Hox (2004) supposed that different level of intraclass correlation of dependent variable led to different conclusions about the sample size conditions across these studies.

Another important finding of the current study was that estimation of the correlations among repeated measurements could not remedy the damage caused by omitting covariates when creating school rankings and classifications. This finding was opposite to the claim of Sanders and his colleagues that by estimating correlations of repeated measurements, each student could serve as his own blocking factor, so that explicit specification of the covariates was not needed (Sanders, Saxon & Horn, 1997).

As mentioned in Chapter II, prior research results about the accuracy of LMEM in school effectiveness estimates were conflicted. The research team led by Sanders provided some empirical evidence to support their claim. However, some research found that LMEM had no advantage even over the Simplest Fixed Effect Model (Tekwe et al., 2004). The study of Ballou, Sanders, and Wright (2004) and the study of McCaffrey et al. (2004) may be helpful in explaining the difference. Ballou, Sanders and Wright (2004) found that LMEM could overcome the problem of omitting the student level covariates, but could not overcome the problem of omitting the school or class level covariates. The current study and the study of Tekwe et al. (2004) omitted the covariates at both the student level and the school level. This may be a reason why the LMEM model could not remedy the damage caused by omitting the covariates. The study of McCaffrey et al. (2004) found that the value-added school effectiveness estimates in LMEM did not correlate with the covariates when the covariates had small intraclass correlations. However, LMEM could not control for the effect of the covariates when the intraclass correlations of the covariates were large. The intraclass correlation of the kindergarten test scores in the current study was 0.21, which might destroy the ability of LMEM to adjust for the influence of the covariates without explicitly specifying them in the model.

Limitations and Future Studies

As mentioned in the first section of this chapter, the current study was conducted with some constraints, and these constraints influenced the results of the current study. Future studies about the effects of these constraints are needed to investigate to what extent the results of the current study may be generalized.

Constraint 1: The Effect Sizes of the Covariates in the Current Study Were Moderately Large

In the current simulation study, the covariates and the parameter values of the effects of the covariates were decided based on literature review, and aimed at reflecting the typical situation in school effectiveness research. The current study only used one level, instead of a range of levels, of effect sizes of the covariates. The effect sizes of the covariates influenced the estimates of school effectiveness in unconditional models (Dorandari, 2004). Without exploring other levels of effect sizes of the covariates, we do not know whether the poor performance of the UCCM model in school rankings and classifications, as found in the current study, will persist in other levels of effect sizes of the covariates.

Future studies are required to study the effect of omitted covariates on school rankings and classifications under a range of effect size conditions. This may provide guidance about when the covariates could be omitted without hurting the accuracy of school rankings and classifications, and when the covariates must necessarily be specified.

Constraint 2: The Kindergarten Test Scores Were Collected Five Years Before the Grade in Which School Effectiveness Was Estimated

In the current study, school effectiveness in Grade 5 was estimated and used for school ranking and classification. The test scores adjusted in the Gain_kindergarten model were assumed to be measured at the end of kindergarten. Thus, there are five years between the initial measurement and the grade of concern.

Although some researchers suggested that initial measurement at the point of entry to an education period is the ideal measure of prior attainment and should be used in adjustment (Cuttance, 1985; Peece, 1992; Sammons, 1996), Rogosa (1995) suggested that it is the interval between initial measurement and the time point of concern that really matters. As noted previously, an appropriate interval can create positive correlation between initial measurement and growth, which can avoid the problem of regression toward the mean. Thus, the author of the current study suspected that the long interval between the initial measurement and the grade of concern in the Gain-kindergarten model might be the key for the Gain_kindergarten model to perform well in ranking and classifying schools. The author questions whether a shorter time interval between initial measurement and the time point of concern might also facilitate the better performance of the Gain_kindergarten model in school rankings and classifications, no matter whether or not the initial measurement is made at the end of kindergarten. Future studies may explore other initial measurements and find out which time intervals are most appropriate for using the gain score model to estimate school effectiveness, and applying the estimates in school ranking and classification. Future studies may also evaluate whether the results of the present study generalize to all the cases in which the time period between initial measurement and the grade of assessing school rankings or classifications is not five years.

Constraint 3: The Simulation Model That Was Used to Generate Test Scores Assumed That Each Student Had a Natural Growth even if No School Effectiveness Existed, and the Natural Growth Was Linear

The model used to generate student test scores was the CCCM model. This model assumes that for each student a linear growth curve exists even without school effectiveness. In addition, this model assumes that the school effectiveness in the previous years can persist without diminishment in the following years. Furthermore, this model does not consider teacher effectiveness when estimating school effectiveness.

McCaffrey et al. (2004) proposed a general longitudinal model, which was supposed to be the most general form subsuming all the models involved in the current study. We still do not know what will happen if this more general model is used to generate the longitudinal data. Although the general longitudinal model is not used in school accountability practice, because of the model’s complexity, more research about this model should be conducted.

Because the computation capability of the computer used here prohibited fitting the CCCM model to a even small dataset, such as 20 schools with 10 students per school and three measurements per students, the performance of the CCCM model in school ranking and classification was not examined. Although the data were generated with the CCCM model, the CCCM model may be not able to completely recover the true school rankings and classifications under all sample size conditions and its performance may be even worse than the Gain_kindergarten model. When a highly capable computer is available, the performance of the CCCM model in school effectiveness estimates, school rankings and classifications, should be examined and compared with other models.

Constraint 4: The Intraclass Correlation of the Kindergarten Test Scores Was 0.21 in the Current Study

The study of McCaffrey et al. (2004) found that the ability of LMEM to control for the effects of covariates was low when the intraclass correlation of the omitted covariates was high. Considering their study, I wondered whether the poor performance of the LMEM model in school ranking and classification was due to the specific level of the intraclass correlation of the kindergarten test scores.

The intraclass correlation of the kindergarten test score was based on a meta-analysis of school effectiveness research and represented the typical level of intraclass correlation of test scores at a single time point (Bosker & Witzier, 1995). The current study did not investigate the effect of varying the intraclass correlation of the covariates on the ability of the LMEM to control for the covariates. A systematic exploration of a range of intraclass correlations of the covariates would be valuable, and would provide guidance about when LMEM can control for the covariates without specifying them in the model.

Another valuable research topic would be to investigate other strategies that can control for the covariates. Although specifying covariates in a statistical model is a commonly used strategy in covariates adjustment, using covariates has many deficiencies, especially in educational research. This is because which covariates should be used in adjustments is not clear yet, and the measurement of the covariates in education, such as the school environment, can present difficult challenges. Furthermore, because of the multicollinearity among the covariates in education, specification of all the associated covariates may also result in biased estimates even if the covariates can be perfectly measured. Other strategies of covariates adjustment are needed to overcome or avoid the problem associated with the strategy that specifies covariates in the statistical model.

The studies of Rausenbaum (2002) and Rubin (2004) about how to infer casual effects in observational studies are extremely important for exploring other strategies of covariates adjustment. They proposed a strategy that matches the participants based on their propensity scores at first, and further statistical analysis is conducted within each group of matched participants. Calculation of the propensity score is based on the associated covariates and is the central piece of this strategy. Further study of this strategy and comparing it with other covariates adjustment strategies in the typical situation of school effectiveness research would be very valuable for value-added assessment of school effectiveness.

Constraint 5: The Total Variance of Growth Was 0.812 and Was Constant across Years

According to the study of Rogosa (1983, 1995), the variance of growth influenced the reliability of gain scores as measurements of growth. The moderate variance of growth in the current study might have facilitated the Gain_kindergarten model to perform well in school ranking and classification. In addition, the constant variance of growth across years specified in the simulation might also have benefited the Gain_kindergarten model.

One topic for future research is to systematically investigate how the reliability of gain scores changes along with the change of gain score variance. This would provide guidance for people to decide whether gain scores are sufficiently reliable so that estimation of school effectiveness can base on it.

Constraint 6: The Current Study Did Not Investigate UCCM and LMEM Model under Large Sample Size Conditions

Because of the limitation of the RAM of the computer, the largest sample size conditions used when studied the UCCM model and the LMEM model was 50 schools with 50 students per schools and 3 measurements per students. The performance of the two models for school rankings and classifications might be better with larger sample size conditions and/or with more measurements per students.

In literature review, I did not find any research about the influence of sample size in more complex multilevel models, such as the cross-classified model, the three-level HLM model, or the LMEM model. All the available studies about the influence of sample size in multilevel models were conducted with the classical two-level HLM model. When the research hardware is sufficiently capable, research about influence of sample size in more complex multilevel models will be highly appreciated in school effectiveness research, and in many other areas.

Constraint 7: The Test Scores at Different Grades Represent a Single Construct or the Combined Constructs with Unchanging Proportions across Grades.

Because the current study explored the model specification issues but not the measurement issues in value-added assessment of schools, all the measurement assumptions associated with value-added assessment of schools were assumed to be satisfied. One of the measurement assumptions is that the test scores used in Value-Added Models represent the same single construct or the same proportional mix of several constructs at each grade. For example, if a Value-Added Model employs student mathematic test scores from Grade 1 to Grade 5, all the mathematic tests from Grade 1 to Grade 5 should measure the same construct (e. g. calculation ability), or the combination of multiple constructs (e. g. calculation ability and reasoning ability). If the test scores measure the combination of multiple constructs, the proportions of the constructs in the combined test scores should be static across the grades (e. g. for tests at all the grades, 20% of the items measure calculation ability and 80% measure reasoning ability). We know that this assumption is not likely to be satisfied either by the statewide achievement tests (e. g. Texas Assessment of Knowledge and Skill) or by the tests published by testing organizations (e. g. Stanford 9 achievement test published by Harcourt Assessment). The proportions of the constructs measured at each grade by the currently used achievement tests are not static across grades. The grade specific statewide achievement tests are created to align with the state curriculum of each grade. The difference of the curriculum in each grade results in different test contents. The achievement tests published by testing organizations are also grade specific which weight different constructs in different ways depending on the knowledge contents and cognitive processes typically obtained by a certain grade of students. Martineau (2006) found that this constructs shift problem caused value added estimates of school or teacher effectiveness in the current year contaminated by the effectiveness of the units the student attended in previous years. For example, the true value of a school's value added effectiveness in the current year is 0.1, because the average schools the students attended before are effective in reasoning instruction but ineffectiveness in calculation instruction, and the proportion of calculation decreases but the proportion of reasoning test increases from the previous grade level, the estimate of the school's value added effectiveness will be spuriously inflated to 0.8. This contamination decreases the reliability of value-added estimates of school effectiveness to a level that is unacceptable for high stake accountability usage (i. e. reliability < 0.9).

Martinuea (2006) suggested a solution to reduce constructs shift. This solution is to embed a large majority of upper- and lower-grade items at each grade level tests, and create a separate vertical scale for each pair of adjacent grades rather than a uniform vertical scale for all the grade levels. However, how well this strategy works is unknown. Furthermore, new vertical equating methods that may reduce or resolve the constructs shift problem are of great value to value-added assessment and other longitudinal studies.

REFERENCES

Aitkin, M. A. & Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149,* 1-43.

Alwin, D. F. (1976). Assessing school effects: Some identities. *Sociology of Education, 49,* 294-303.

Ballou, D., Sanders, W. & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29,* 37-65.

Bosker, R. J. & Scheerens, J. (1989). Issues in the interpretation of the results of school effectiveness research. *International Journal of Educational Research, 13,* 741-751.

Bosker, R. J. & Witzier, B. (1995). *A meta analytical approach regarding school effectiveness: The true size of school effects and the effect size of educational leadership* (Tech. Rep.). The Netherlands: University of Twente, Department of Education, Division of Education Administration.

Braun, H. (2005). Value-added modeling: What does due diligence require? In R.W. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 19-40). Maple Grove, MN: JAM Press

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

Busing, F.M. T. A. (1993). *Distribution characteristics of variance estimates in two-level models: A Monte Carlo study* (Tech. Rep.). The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Carter, R. L. (2004). Rejoinder. *Journal of Educational and Behavioral Statistics, 29,* 135-137.

Casteel, D. (1994). *Principal and teacher perceptions of school climate related to value-added assessment and selected school contextual effects in the first Tennessee District.* Unpublished doctoral dissertation. East Tennessee State University, TN.

Cook, D. L. (1985). *An objective component for the evaluation of teaching.* Unpublished doctoral dissertation. The University of Tennessee, Knoxville, TN.

Crocker, L. M & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Cuttance, P. (1985). Frameworks for research on the effects of schooling. In D. Reynolds (Ed.), *Studying School Effectiveness*. (pp. 34-68) Lewe: Falmer Press.

Cuttance, P. (1992). Evaluating the effectiveness of schools. In D. Reynolds and P. Cuttance (Ed.), *School effectiveness: Research, policy and practice,* (pp. 71-95). Londen: Cassell.

Darandari, E. Z. M. (2004). *Robustness of hierarchical linear model parameter estimates under violations of second-level residual homoskedasticity and independence assumptions.* Unpublished doctor dissertation, Florida State University. Tallahassee, FL.

De Leeuw, J. & Kreft, I. G. G. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics, 20*, 171-190.

Donoghue. J. R., & Jenkins, F. (1992). *A Monte Carlo study of the effects of model misspecification on HLM estimates* (Tech. Rep.). Princeton, NJ: Educational Testing Services.

Doran, H. C. (2003). *Value-added analysis: A review of related issues.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ferron, J. (1997). Moving between hierarchical modeling notations. *Journal of Educational and Behavioral Statistics, 22,* 119-123.

Fitz-Gibbon, C. T. (1997). *The value-added national project final report* (Tech. Rep.) London: SCAA.

Goldstein, H. (1984). The methodology of school comparison. *Oxford Review of Education, 10,* 69-74.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics, 16,* 89-91.

Goldstein, H. (1995a). *Multilevel models in educational and social research: A revised edition.* London: Edward Arnold.

Goldstein, H. (1995b). *Multilevel statistical models.* New York, NY: Halsted Press.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8,* 369-395.

Goldstein, H., Rasbach. J., Yang, M., Woodhouse, G, Pan, H., et al. (1993). A multilevel analysis of school examination results, *Oxford Review of Education*, *19*, 425-433.

Gray, J, Jesson, D., & Sime, N. (1990). Estimating difference in the examination performance of secondary schools in six LEAS-A multilevel approach to school effectiveness. *Oxford Review of Education, 16*, 137-158.

Haville, D. A. (1995). A review of Tennessee value-added assessment system (TVAAS). Retrieved from the website http://www.cgp.upenn.edu/ope_techreports.html, 2005, Nov 14th.

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics*, *23*, 117-128.

Hill, P.W. & Rowe, K. J. (1996). Multilevel modeling in school effectiveness research. *School Effectiveness and School Improvement*, *7*, 1-34.

Kamali, M. A. (1992). *A study of single and multi-level logistic regression models using real and computer simulated data.* Unpublished doctor dissertation. Michigan State University, East Lansing, MI.

Kim, K. S. (1990). *Multilevel data analysis: A comparison of analytical alternatives.* Unpublished doctoral dissertation. University of California, Los Angeles, CA.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies* (Tech. Rep.). Los Angeles, CA: University of California, Department of Statistics.

Krull, J. (1997). *The effect of misspecification resulting from analysis decisions in multilevel models.* Unpublished doctoral dissertation. Arizona State University, Phoenix, AZ.

Linn, R. (2005). *Issues in design of accountability systems* (CSE technical report 650). Los Angeles: University of California at Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute Inc.

Lockwood, J. R., Doran, H. D., & McCaffrey, D. F. (2003). Using R for estimating longitudinal student achievement models. *R News, 3,* 17-23.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 18,* 437-454.

Martineau, J. A. (2006). Distorting value-added: The use of longitudinal vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31,* 35-62.

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

McCaffrey, D. F., Lockwood, J. R. Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability* (Tech. Rep.). Santa Monica, CA: Rand Corporation.

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

McLean, R. A., & Sanders, W. L. (1984) *Objective component of teacher evaluation: A feasibility study* (working paper No. 199). Knoxville: University of Tennessee, College of Business Administration.

McLean, R. A., Sanders, W. L., & Stroup, W. W. (1991). A unified approach to mixed linear models. *American Statistician, 45,* 54-64.

Mendro, R. L. (1998). Student achievement and school and teacher accountability. *Journal of Personal Evaluation in Education, 12,* 257-267.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Educational Review, 16,* 283-301.

Mortimore, P., Sammons, P., Stoll, L., Lewix, D. and Ecob, R. (1988). *School matters: The junior years.* Somerset: Open Books.

Ponisciak, S. M. & Bryk, A. S. (2005). Value-added analysis of the Chicago public schools: An application of hierarchical models. In R.W. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 19-40). Maple Grove, MN: JAM Press.

Preece, P. (1989). Pitfalls in research on school and teacher effectiveness. *Research Papers in Education*, *4*, 47-69.

Raudenbush, S. W. (1989). The analysis of longitudinal, multilevel data. In B. P. M. Creemers & J. Scheerens. (Eds.). Developments in school effectiveness research, Special issue of *International Journal of Educational Research*, *13*, 721-739.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18,* 321-349.

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,* 121-129.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*, 25-31.

Raudenbush , S. W. & Bryk, A. S. (1989). Methodological advances in analyzing the effects of schools and classrooms on student learning. *Review of Research in Education, 15*, 423-475.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publication.

Raudenbush, S. W. & Willms, J. D. (1995). The estimation of school effect. *Journal of Educational and Behavioral Statistics, 20,* 307-335.

Rosenbaum, P. R. (2002). *Observational studies (*2nd ed.). New York, NY: Springer-Verlag Inc.

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,* 1525-1567.

Rubin, D. B. (2004). Teaching statistical inference for causal effect in experiments and observational studies. *Journal of Educational and Behavioral Statistics, 29,* 343-367.

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,* 103-116.

Sammons, P. (1996). Complexity in the judgment of school effectiveness. *Educational Research and Evaluation, 2*, 113-149.

Sammons. P., Nuttall, D., Cuttance, P. & Thomas, S. (1995). Continuity of school effects: A longitudinal analysis of primary and secondary school effects on GCES performance, *School Effectiveness and School Improvement, 6,* 285-307.

Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*, 329-339.

Sanders, W. L., & Horn, S. D. (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., Saxon, A. M. & 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?* (pp. 137-163). Thousand Oaks, CA: Corwin Press.

Scheerens, J. (1992). *Effective schooling: Research, theory, and practice*. London: Cassell.

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24,* 323-355.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage Publications.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research.* New York, NY: Falmer Press.

Teddlie, C. & Stringfield, S. (1993). *Schools do make a different: Lessons learned from a 10-year study of school effects*. New York: Teachers College Press.

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

Thomas, S. & Mortimore, P. (1996). Comparison of value-added models for secondary-school effectiveness. *Research Paper in Education, 11,* 5-33.

Thomas, S., Sammons, P., & Mortimore, P. (1997). Stability and consistency in secondary schools effects on students GCSE outcomes over 3 years. *School Effectiveness and Improvement, 8*, 169-198.

Webster, W. J. (2005). The Dallas school level accountability model: The marriage of status and value-added approaches. In R.W. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 19-40). Maple Grove, MN: JAM Press.

Webster, W. J. & Mendro, R. T. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure? (*pp. 81-99). Thousand Oaks, CA: Corwin Press.

Weerasinghe, D. & Orsak, T. (1998). *Can hierarchical linear modeling be used to rank school: A simulation study with conditions under which hierarchical modeling is applicable.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Willms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review, 51,* 224-241.

Willms, J. D. (1987). Differences between Scottish educational authorities in their educational attainment. *Oxford Review of Education, 13*, 211-232.

Willms, J. D. (1988). *Estimating the stability of school effects with a longitudinal hierarchical linear model.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. London: Falmer Press.

VITA

Contact Information

Zhongmiao Wang
2525 Preston Rd. #322
Plano, TX 75093

Education

Texas A&M University, College Station, TX
December 2006
Ph.D., Educational Psychology – with emphasis in research methods, measurement, and statistics

Beijing Normal University, Beijing, China
Master of Science in Education – Curriculum and Instruction
2002
Thesis: "Measurement Model of Positive Emotional Experience in Elementary Students' Learning"
Bachelor of Science in Psychology
1999

Publications

Wang, Z. & Thompson, B. (in press). Is the Pearson r^2 biased, and if so, what is the best correction formula? *Journal of Experimental Education.*

Stellefson, M & **Wang, Z.** (under review). Effects of cognitive dissonance on diet and physical activity behaviors in college students. *American Journal of College Health.*

Presentations

Wang, Z., & Wen, Luo (2006). *Can we use gain score model to rank or classify schools, and if so, for which previous attainment we should adjust.* Paper presented at the American Educational Research Association annual meeting. San Francisco, CA.

Wang, Z., & Willson, V. (2005). *Indifference region and confidence region in Structural Equation Modeling.* Paper presented at the American Educational Research Association annual meeting. Montreal, CA.

Willson, V., & **Wang, Z.** (2004). *Indifference region in Structural Equation Modeling.* Paper presented at the Psychometric Society annual meeting. Monterey, CA.

Academic Awards

Department Research Award
Department of Educational Psychology, Texas A&M University, 2005, 2006
Graduate Student Research and Presentation Grant
Texas A&M University, 2003