**The Dissertation Committee for Judith Ann Jennings Certifies that this is the approved version of the following dissertation:**

# A Comparison of Statistical Models Used to Rank Schools for Accountability Purposes

**Committee:**

Susan N. Beretvas, Co-Supervisor

Keenan A. Pituch, Co-Supervisor

Barbara G. Dodd

Gary D. Borich

Charles T. Clark

# A Comparison of Statistical Models Used to Rank Schools for Accountability Purposes

**by**

**Judith Ann Jennings, B.A.**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

**May, 2006**

# Dedication

For Hal

## Acknowledgements

Never has a document required so much of so many. I would like to thank:

Barbara Dodd, the first person to encourage me to pursue graduate school. You have stayed the course over the many years, and I appreciate all you've done for me.

Tasha Beretvas and Keenan Pituch, co-chairs extraordinaire. Thanks for your support, patience, and help.

Spiritcrafters, AWOL, and all the other incredible women in my life. You have been there for me through so many life experiences, and you always know when to hug and when to push.

The women of RFL, who seem to have an infinite supply of support, laughter, intellectual challenge and Mistos. You are the best group of first-tiers a girl could ever have.

My children, Kari Maurer and Kevin Jennings, for being ok with a mom who isn't quite like the others. The two of you have taught me so much about what is really important in life. Thanks for being YOU!

Most importantly, I want to thank my husband, Hal Jennings, for believing in me before I did. None of this would have happened without you.

# A Comparison of Statistical Models Used to Rank Schools for Accountability Purposes

Publication No._____

Judith Ann Jennings, Ph.D.

The University of Texas at Austin, 2006

Supervisors: Susan N. Beretvas, Keenan A. Pituch

Public school accountability has become an important part of national educational policy. Schools are ranked for accountability purposes using a variety of statistical models. These models include performance, or status models, and productivity, or student change models. Performance models do not separate the influence of school effects from background effects such as socioeconomic status on student achievement, while productivity models can isolate school and background influences on student achievement.

The current study investigated the differences between school rankings calculated using a performance model and three types of productivity models in terms of consistency of rankings and relation between ranking and school percent low socioeconomic status students. In the first study, using real data, school rankings were calculated using the percent passing, cohort difference, unadjusted and adjusted single-level regression, and unadjusted and adjusted multilevel models. A simulation study was also conducted which simulated 250 schools with varying percentages of low

socioeconomic status students. Student achievement for 30 students was simulated with varying degrees of relation between school percent low socioeconomic status students and student achievement, student socioeconomic status and test score, and the amount of variation in student test scores between schools.

School rankings were remarkably different when calculated using the different models; especially the percent passing model. The magnitude of differences is especially important when policy makers consider rewards for top-performing schools or sanctions for low-performing schools. Correlation of rank calculated using the percent passing model with school percent low socioeconomic status students was as high as 0.41 The simulation study showed that ranking calculated using each of the models was most highly correlated with school percent low socioeconomic status students when there was a strong (-0.10) simulated relation between student socioeconomic status and individual student test score as well as a relation (-0.10) between percent low socioeconomic status students and individual student test score. In all conditions the correlation between rank and school percent low socioeconomic status students was weaker when a larger proportion of variance in student test scores was within school.

# Table of Contents

List of Tables ..... x

**CHAPTER 1: INTRODUCTION** **1**

**CHAPTER 2: LITERATURE REVIEW** **4**

Inclusion of Student Background Characteristics ..... 7

Statistical Models ..... 10

Performance/Cross-Sectional Models ..... 10

Productivity/Student Change Models ..... 16

Multilevel modeling ..... 23

Previous Comparisons ..... 34

Statement of Problem ..... 39

**CHAPTER 3: METHODS** **42**

Real Data Study ..... 42

Sample ..... 42

Assessment ..... 43

Models ..... 44

Comparisons. ..... 50

Simulation Study ..... 52

**CHAPTER 4: RESULTS** **58**

Real Data ..... 58

*Descriptive Statistics* ..... 58

*Research Question One* ..... 66

*Research Question Two* ..... 69

*Research Question Three* ..... 71

Simulated Data Sets .....88

**CHAPTER 5: DISCUSSION** **101**

Conclusions .....113

**GLOSSARY** **116**

**REFERENCES** **117**

**VITA**

# List of Tables

Table 1: Models used in Analysis: Percent Passing and Cohort Difference.....45

Table 2: Models used in Analysis: Unadjusted and Adjusted School-Level Regression.....46

Table 3: Detail of School Level Regression Configurations .....46

Table 4: Models used in Analysis: Dallas-Type, Unadjusted and Adjusted Multilevel Regression.....48

Table 5: Detail of Multilevel Configurations.....49

Table 6: Summary of Simulation Design.....56

Table 7: Descriptive Statistics of Grade Levels Included in the Study .....58

Table 8: Descriptive Statistics of Students Included in the Study.....58

Table 9: Descriptive Statistics of Schools Included in the Study.....59

Table 10: Descriptive Statistics for the School Status Model.....60

Table 11: Descriptive Statistics for the Cohort Difference Model .....62

Table 12: Descriptive Statistics for School Regression Model .....63

Table 13: Descriptive Statistics for Multilevel Models.....65

Table 14: Comparison of Unadjusted (USLR) and Adjusted (ASLR) School Level Regressions .....66

Table 15: Comparison of Unadjusted Multilevel Model (UNINT) with other Multilevel Models having Random Intercept (DLINT & AINT).....68

Table 16: Comparison of Unadjusted Multilevel Model (UNINTSLP ) with other Multilevel Models having Random Intercept and Slope (DLINTSLP & AINTSLP).....68

Table 17: Correlation of Ranking Results from Each Model with Percent of Students Receiving Free or Reduced Price Lunch .....71
Table 18: Correlations of Ranking across Models, Grade 4 Mathematics .....73
Table 19: Correlations of Ranking across Models, Grade 4 Reading.....73
Table 20: Correlations of Ranking across Models, Grade 5 Mathematics .....74
Table 21: Correlations of Ranking across Models, Grade 5 Reading.....74
Table 22: Average Correlation of School Rankings across Grades and Subjects 75
Table 23: Average Absolute Value (AAV) of Difference Between Percent Passing (PCTPASS) Model and Cohort Difference (COHDIFF) Model .....76
Table 24: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Single-Level Regression Models (UNSLR & ASLR) .....77
Table 25: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept Multilevel Models (DLINT, UNINT & AINT) .....78
Table 26: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept/Random Slope Multilevel Models (DLINTSLP, UNINTSLP & AINTSLP) .....79
Table 27: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Cohort Difference Model (COHDIFF) for Schools in the Top 20% with 30 or More Students .....80
Table 28: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Single-Level Regression Models (UNSLR & ASLR) for Schools in the Top 20% with 30 or More Students .....81

Table 29: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept Multilevel Models (DLINT, UNINT & AINT) for Schools in the Top 20% with 30 or More Students .....81

Table 30: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept/Random Slope Models (DLINTSLP, UNINTSLP & AINTSLP) for Schools in the Top 20% with 30 or More Students.....82

Table 31: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Cohort Difference Model (COHDIFF) for Schools in the Bottom 20% with 30 or More Students.....84

Table 32: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Single-Level Regression Models (UNSLR & ASLR) for Schools in the Bottom 20% with 30 or More Students ..85

Table 33: Average Absolute Value (AAV) of Difference Between Models Percent Passing Model (PCTPASS) and Random Intercept Multilevel Models (DLINT, UNINT & AINT) for Schools in the Bottom 20% with 30 or More Students .....85

Table 34: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept/Random Slope Models (DLINTSLP, UNINTSLP & AINTSLP) for Schools in the Bottom 20% with 30 or More Students.....86

Table 35: Simulation Conditions .....89

Table 36. Correlation Between Ranking and Percent Low SES Students When No Relation Between SES and Test Score is Simulated .....91

Table 37. Simulated Conditions Under Which Ranking was Most Highly Correlated with Percent of Low SES Students .....93
Table 38: Correlation Between School Rank and Percent Low SES Students using Simulated Data.....95
Table 39. Simulated Conditions Under Which Unadjusted and Adjusted Regression Models Differ Most.....97

# CHAPTER 1: INTRODUCTION

In 2001, the federal government reauthorized the Elementary and Secondary Education Act of 1965 with a revised education law commonly referred to as *No Child Left Behind* (NCLB) (U. S. Department of Education, 2001). One of the requirements of this reauthorization is that each state must have a method of evaluating schools for the purpose of federal accountability. Some states had no previous state accountability system and were faced with developing an entire system in a relatively short period of time. Other states had a well-developed system of educational accountability for state purposes, and were required to ensure that the state system complied with federal requirements.

In some states with already-established systems, the conversion of the state accountability system for federal accountability purposes has been relatively easy, while the conversion has been complex and difficult in other states. Some states had created systems which evaluated only districts or campuses for state accountability purposes. The federal system requires that both be judged on how well they are educating students. Other states were testing students only upon graduation. The federal system demands annual testing in elementary, middle and high school grades. Still other states were reporting the results of evaluation components at the aggregate level, perhaps grade, school, or district, but the federal system requires that results be reported before aggregating, by ethnicity, student SES, and for students with disabilities.

In some states with state accountability systems already in place, aspects of the state system which conflict with federal requirements are being discarded, and the federal accountability system is being used to evaluate schools (Consortium for Policy Research in Education, 2000). In other states, it has been decided to comply with the federal

accountability system as required by law, but to continue evaluation of schools and/or districts for state accountability using the already-established state evaluation system (Texas Education Agency, 2006).

While the federal law has been prescriptive in the methods used for evaluation, the state systems are based on a variety of models. Accountability decisions based on such models can have tremendous impact. In some states, schools are issued a grade depending upon students' performance on statewide assessments. Other consequences can include teacher or principal bonuses or school closure (Rouk, 2000). School and district accountability decisions have numerous impacts on the lives of the students, parents, teachers, and taxpayers who are represented by the schools or the school districts. It is therefore imperative that decisions be well thought out and make practical sense to all parties concerned, in addition to making statistical sense.

One of the differences between accountability systems in different states is the statistical method used to evaluate whether the campus or district has in fact educated the students who attended during the year. Most states use results of a statewide examination to determine how many of the attending students have actually learned the elements of the curriculum being tested. In some states, this method is virtually identical to the mandated federal system, and consists of determining the percentage of students who demonstrated proficiency on the statewide assessment (Consortium for Policy Research in Education, 2000). . If that percentage is equal to or higher than a state standard, the campus or district has "passed." In other states, the method involves determining the difference between the percent of students demonstrating proficiency from one year to the next. If that difference exhibits a sufficiently large improvement, the campus or district has "passed." Still other states match individual student records from one year to

the next, aggregate the increase or decrease in individual scores on a statewide assessment, and rate the campus or district accordingly.

An important issue in construction of state evaluation systems is the influence of home background, and other factors outside of the control of the school which impact student test scores. Research has shown differences in student test scores based upon such factors as SES. Systems which do not take SES into account are accused of ignoring the bigger challenges faced by some campuses and districts in educating their students. Systems which do take such factors into account are accused of setting different standards for different groups of students, based on the students' SES.

This dissertation will examine different models in use in various state accountability systems. Results from various models will be compared and similarity in results reported, as well as the relationship between results from each model and percent of low socioeconomic students in schools.

# CHAPTER 2: LITERATURE REVIEW

Beginning in the 1970s, states began to mandate minimum competency testing to ensure that students had acquired minimal basic skills before being awarded a high school diploma. Since that time, the mandate has grown to include assessment of school effectiveness in preparing students for such testing (Alban, 2002). A common method of classifying effective schools is using student achievement scores to categorize a school as "acceptable" or "unacceptable" or to rank schools from highest to lowest based upon student achievement scores. There is concern, however, with judging school effectiveness based upon student achievement scores alone, because such scores are partially dependent on factors beyond the control of teachers and school staff.

Hanushek and Raymond (2002) describe student achievement as being a function of two primary factors, *school* and *other*. *School* effects are those programs and policies enacted by a school and controlled by the school, which have an effect on student achievement. *Other* effects refer to student ability, family influences, history, and measurement error. Separating *school* effects from *other* effects is a complex task.

Further, the purpose of the determination of *school* effects confuses the issue. If the purpose of the determination of school effectiveness is accountability by district administration or state or federal government, then *school* effects should be isolated from *other* effects, over which school personnel have no control. Effects due to the influences of *school* are referred to by Raudenbush and Willms (1995) as Type B effects. Type B effects isolate the influence of school practice, including curricular content, classroom instruction, administrative leadership and utilization of resources from other influences on student achievement, such as student background and school context (Raudenbush and Willms, 1995). Alternatively, if the purpose is to assist parents in selecting the best of all

possible schools which their children might attend, then both *school* and *other* effects can be important. These are called Type A effects by Raudenbush and Willms (1995). Such effects do not separate different influences on student achievement. Type A effects include student background and school context (such as average socioeconomic background) as well as school practice.

As state and federal education agencies struggle with the complexities of implementing school accountability systems, they are facing the decision of whether to differentiate between Type A and Type B school effects. Only in the past decade or two have school accountability systems been put into use, and the details of such systems are still being worked out. The purpose of determination of school effects is often not clear, and the variety of state accountability models in use across the United States bears witness to the lack of clarity.

Consequences of the impact of decisions made by policy makers can be severe. A report by the Southwest Educational Development Laboratory describes rewards and sanctions in five states ranging from cash awards for top performing schools to school takeover (Rouk, 2000).

Alban (2002) classifies school accountability models as being either performance or productivity models. Performance models are those that base school performance classifications on whether or not schools are meeting performance standards such as percent of students showing a certain level of proficiency on a statewide test. These models make no attempt to break down achievement into *school* and *other* influences on student achievement, and can thus be considered Type A Models.

Productivity models classify schools based on the relative impact of teachers or school programs. These models are sometimes called "value-added" models. Productivity models control for student background effects such as prior year test score

and SES, as well as contextual effects such as socioeconomic makeup of the school, to determine how much of the student achievement can be attributed to the influence of the school. These models can be considered Type B Models.

Fletcher and Raymond (2002) also view the accountability models used by states as two types, each of which can further be broken down into two types of models. In their classification system, performance, or Type A, models are called cross-sectional models, and include school status and grade level change models. School status models classify schools depending on whether there is a sufficiently high percentage of students at each grade level passing a statewide achievement test. Grade level change models compare the percent of students passing a statewide achievement test at each grade level to the percent of students passing the test at the same grade level in the previous year. Productivity, or Type B, models, which Fletcher and Raymond (2002) call student change models, include cohort difference and individual gain models. Cohort gain models classify schools based on the comparison of percent of students passing a statewide achievement test at each grade level compared to the percent of students in that same cohort passing the test at the previous grade level in the previous year. Individual gain models use single-level regression or hierarchical linear regression to classify school effectiveness. The classification systems currently used in most states are performance, or cross-sectional models (Consortium for Policy Research in Education, 2000).

The following sections will address choices to be made in selection of a school accountability model. First, the issue of whether or not to include student background characteristics will be discussed. That will be followed by a sequential consideration of performance (cross-sectional) and productivity (student change) models, including model descriptions, examples of states that use each type of model, strengths and weaknesses of each, and which components of student achievement are included in the model.

## Inclusion of Student Background Characteristics

Student background characteristics outside the control of the school, such as socioeconomic status (SES) or limited English proficiency have been shown to be strongly related to school achievement (Clotfelter & Ladd, 1996; Coleman 1966). The choice of whether to assess schools using Type A effects which do not isolate various influences on student achievement, or Type B effects which isolate school influence from other effects, is difficult. As stated earlier, much of the decision should depend on the purpose of the school classifications being created. If the purpose is to determine which schools are performing best overall, then Type A effects should be used to classify schools. If, on the other hand, the purpose is to evaluate the performance of teachers and other school personnel based on the progress made by students in their school, Type B effects should be used. The reasons behind this decision will be addressed in this section.

The debate about the influence of SES on student achievement began in 1966 with the publication of the report entitled Equality of Educational Opportunity, otherwise known as the Coleman Report. That work showed that the strongest predictor of student achievement was student SES (Coleman, et al., 1966). Debate quickly surfaced over the amount of impact schools and teachers could have on the achievement of students given the strong influence of family background, and whether it is even appropriate to evaluate schools using the test scores of students (Clotfelter and Ladd, 1996).

Student background, consisting of many factors outside of the influence of the school, is a complex issue. If, as Coleman (1966) claims, SES is the best predictor of achievement, and if schools are held accountable for educating all students to a certain standard, then schools with large numbers of relatively low SES students have a more difficult job than schools with predominantly higher level SES students. If schools are

rated without consideration of student SES, schools with a high proportion of students with relatively low SES may consistently fall to the bottom of the ratings (Betts and Danenberg, 2002; Clotfelter and Ladd, 1996). Accountability systems which do not take such factors into account are accused of ignoring the difference in difficulty of educating students at schools with relatively low wealth, compared to students at schools with relatively high wealth. Systems which do take such factors into account are accused of setting different expectations for the achievement level of different groups of students, based on factors such as the students' poverty. Thum and Bryk (1997) state that the purpose of the analysis (productivity of schools or showing progress in reaching standards) determines whether or not such adjustments should be made. It can be argued that state accountability systems attempt to both assess productivity and assess progress against standards.

Clotfelter and Ladd (1996) found that whether or not family variables are included in the school classification makes a difference in the results of the school assessments. They compared methods of establishing school indices, and provided correlations of results among alternate methods of obtaining the indices. Results from the index including consideration of socioeconomic differences were very different from the results of other methods which did not include adjustments for SES. Overall, the correlations between different methods ranged from .22 to .58.

Performance models as a rule do not isolate different influences on student achievement. In order to acknowledge the impact of the effects of student background in states which use performance models, some states, such as California and Pennsylvania, form groups of "like schools" based upon factors such as the percent of students with low SES, which are known to be related to student test scores (Linn, 2001b). Schools are then classified based upon how they did in relation to the schools in their group. This

method has the effect of setting standards for some schools lower than the standards set for other schools, whose contextual characteristics such as average student SES might be correlated with test scores (Clotfelter and Ladd, 1996).

Hill and Lake (2002) studied two statewide samples of elementary schools from 1997 and 1998. Their study compared two groups of schools with similar demographics (majority low income) in which students had scored below average on the fourth grade test in 1997. They defined as "rapidly improving schools" those that had test scores increase at more than twice the statewide average. After extensive interviews with principals, teachers, parents and students, they state that, "…whether a school improved or not depended on what the adults in it did in response to the new standards and tests" (203). Successful rapidly-improving schools had personnel who were focused on figuring out just why scores were low, used the data to decide what to do to improve, and then followed through with changes. In the end, Hill and Lake (2002) conclude that "Family income is an advantage for some schools and a problem for others, but in itself it does not cause student learning." (204)

A problem with including student background variables in school accountability models is possible multicollinearity. Some demographic variables, such as SES and minority status, are often highly correlated. Inclusion of both can then provide inaccurate estimation of the influence of each of the variables. When limited English Proficiency and student mobility are added as variables, multicollinearity becomes an even bigger possibility, because in many large southern states especially, the groups of low SES, highly mobile, Hispanic, limited English proficient students are comprised of the same set of young people

All of the performance or productivity models described in the next sections have the potential for controlling for student background variables as part of the classification

of schools. While performance models are in general Type A models, it is possible for schools to be classified in ways that control for student background variables. Productivity models are in general Type B models, which attempt to isolate school effects from student background and school context effects. States which use performance models have been accused of ignoring the difficulties associated with hard-to-educate students, and school accountability systems in those states could be biased against schools with high numbers of such students. Addressing the potential for bias against schools with high percentages of students who are considered difficult to educate is an important part of any accountability system which attempts to isolate the impact of schools and teachers on student achievement.

Another decision that must be made in school accountability systems is the choice of models used for the classification process. The following section describes in detail each of the four types of models listed earlier.

## Statistical Models

### PERFORMANCE/CROSS-SECTIONAL MODELS

Performance, or cross-sectional models, such as those required by the federal *No Child Left Behind* law, set the same academic achievement expectations for all schools, regardless of school context or student background. Schools are held accountable for educating all children to the same standard. These models include the school status model, in which all students are held to one standard, and the grade-level change model, in which performance of students in one grade is compared to the performance of other students in the same grade the previous year.

***School Status Model***

The federal *No Child Left Behind* law requires that schools and districts be evaluated by each state according to a school status model (U.S. Department of Education, 2001). This model was already commonly used in state accountability systems (Consortium for Policy Research in Education, 2000). Absolute standards are set prior to the school year by the state board of education or other governing body at the state level for a required average test score or percent of students passing a statewide test. The average test score or percent of students passing the test at the end of each school year for each school is compared to the statewide standard. In this situation, schools either "make" or "do not make" the absolute standard established for acceptability. School status models can utilize average scores or percents passing on a grade-by-grade basis or aggregated to the school level. As Tekwe, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, and Resnick, (2004) state, "The distinguishing characteristic of status-based methods is the absence of adjustment for students' incoming knowledge level" (p. 12).

In fourteen states, schools are evaluated for state accountability purposes based on whether their students meet an absolute target in terms of percent showing proficiency (Consortium for Policy Research in Education, 2000). That number could potentially increase because the federal legislation requires the use of status models for federal accountability, and it may be that states will prefer not to have the added complexity of two separate systems of accountability. The status models used by Texas and West Virginia will be described next as examples.

The West Virginia accountability system uses the Adequate Yearly Progress model, which is the model used by *No Child Left Behind*, to evaluate schools. For state accountability, schools are designated as either meeting or not meeting the standards required by the federal system. Schools are not compared to each other, rated or ranked

in any way other than "met" or "did not meet" the standard set for the school. Under this federal system, the same performance standards are used separately for all racial/ethnic subgroups, as well as low SES students, students with disabilities, and limited English proficient students (Consortium for Policy Research in Education, 2000)

The state accountability system used in Texas through 2002 had four categories into which a school could be classified. These included *unacceptable, acceptable, recognized* and *exemplary*. Percent of students passing the statewide assessment, the Texas Assessment of Academic Skills (TAAS), determined into which category a school fell. The percent of students passing was determined and reported separately for economically disadvantaged students, African-American, Hispanic, white students, and aggregated to the all-students level. Reporting results separately for different demographic groups was intended to introduce a more complete level of accountability for all students. The standards for school classification were increased throughout the duration of the accountability system, and in 2002 the percent proficient required for each categorization included: 90% of each group passing the assessment for a school to be rated *exemplary*, 80% for *recognized*, and 70% for *acceptable*. The performance of any school with less than 70% of students passing the TAAS test was rated *unacceptable* (Texas Education Agency, 2002).

Texas dealt with the issue of whether or not to include other factors in evaluations by creating comparable improvement as part of a special acknowledgement system called the Gold Performance Acknowledgement system. Comparable improvement groups of 40 campuses were created for each campus using the 39 campuses statewide with the most similar student demographic composition. These groups were not consistent groups. Comparable groups were defined for each school and thus the 40 schools were not partitioned into mutually exclusive groups of schools. Schools that had been rated

acceptable or higher in the regular system, and that performed well in comparison to this group of like schools, were rewarded with special acknowledgement (Texas Education Agency, 2002). The Texas model first evaluates schools using Type A effects, and then attempts to control for the impact of school context through comparable groups used for the Gold Performance Acknowledgement system.

The most important advantage of school status models is their simplicity. It is easy for school administrators to calculate the average test score or the percent of students passing a test. In addition, both of those numbers are relatively easy for the general public to understand.

School status models do not discriminate between variation in achievement due to *school* or *other* factors. A problem with a school status model is that it is not equally difficult for all schools to attain the performance standard. If student achievement varies according to SES, then making accountability decisions based on Type A school effects means that schools with high numbers of low SES students will require educators to exert more time and effort to bring sufficient numbers of students to the performance standard than schools with few such students.

A related problem with a school status model is that it may be tempting for school administrators or teachers to move students who enter school at a low level of achievement into groups whose performance is not included on the test, such as special education classes, or encourage such students to be absent on the day of testing. By excluding these students, schools look better in an accountability system which evaluates them based on an absolute level of performance (Hanushek and Raymond, 2002). States have attempted to deal with this exclusion by including all groups, including special education students, in the accountability system, and following the example of the federal

system by requiring a certain level of participation in the test. The federal standard for participation is inclusion of 95% of students (U.S. Department of Education, 2001).

A further problem with school status models involves school rankings. If schools are evaluated without the inclusion of student demographic characteristics in the model, then rankings are fairly straightforward. However, in this case it is not possible to isolate Type B from Type A effects, and if the purpose of the determination of school effectiveness is accountability, then *school* effects should be isolated from *other* effects.

***Grade Level Change Model***

A second type of performance/cross-sectional model compares performance of students in one grade with the performance of other students in the same grade the previous year. The grade level change model is a version of the status model, one that includes stratification by grade (Fletcher and Raymond, 2002). Such methods are called grade level change systems (Fletcher and Raymond, 2002), status change models (Hanushek and Raymond, 2002) or successive groups approaches (Linn and Haug, 2002). These systems are an attempt to measure the strength of a program, rather than the strength of the students in the program. In grade-level change models, the performance of students in a certain grade, for instance 3$^{rd}$ grade, in one year is compared to the performance of other students in the same grade (here, 3$^{rd}$) the previous year. Usually the state accountability system requires improvement from year to year, requiring that students in the grade(s) evaluated show a higher level of absolute performance than other students in the same grade(s) the previous year. This higher level of performance is often defined as a higher percentage of students achieving a certain performance level, or a higher average test score for the students being evaluated.

Wisconsin used a grade level change model prior to 2002-2003, in which schools were required to meet a "Continuous Progress Indicator." Students were administered

the Wisconsin Knowledge and Concepts Examination in grades 4, 8 and 10. The Continuous Progress Indicator credited schools for improvement in the percent of students scoring at or above the proficient category, and for movement of the percent of students from "Not Tested" or "Minimal Performance" into "Basic" or above (Consortium for Policy Research in Education, 2000).

A problem with the grade level change models is that different students comprise the groups being compared from year to year, and there is no assurance that student characteristics that affect student achievement are relatively stable from year to year (Linn and Haug, 2002). Two possible sources of error arising from comparison of different groups of students over two or more years are student mobility and heterogeneity of student body. It could be that in some schools student mobility is low enough and the background of students attending the school similar enough, that this is not an issue. In this instance, the other influences on student achievement are relatively consistent for consecutive groups of students. Studying grade level change in student achievement in this case might be a valid comparison. However, Hanushek and Raymond (2002) report that as few as half of students attending a school have lived in the same house for three years in a row, which indicates high mobility in school populations.

A grade level change model can be useful in a state in which not all consecutive grades are tested. In this situation, it is not possible to follow a cohort or individual students from year to year. If the grades tested are not close (for instance testing occurs in grades 3 and 8 only) then there may be little continuity of students tested in the two grades, which is a problem for the student change models described later.

Models assessing grade-level change yield data that do not distinguish between change in the quality of school programs and change in student characteristics. Thus there is still confusion between the two components which predict achievement, *school*

and *other*. The next section discusses productivity, or student change models, which are designed to separate the influence of schools from student background and school contextual characteristics.

## PRODUCTIVITY/STUDENT CHANGE MODELS

Productivity, or student change models, can be classified into two types. The first is a cohort difference model, which examines "pseudo cohorts" of students as they move from grade to grade. The second model is an individual change model, in which student records are matched from year to year and individual students are followed as they move from grade to grade. The latter models have begun to emerge in the past decade as student identifier systems have become more sophisticated, and will be described after cohort difference models

### *Cohort Difference Model*

In some states, a cohort method of evaluation compares how students in each grade performed relative to how the same students did the previous year. This is not a "true" cohort because students are not matched year-to-year, and so some students move in or out of the cohort between testing periods. Meyer (2000) calls this a gain indicator.

A cohort difference model can eliminate the problems in a grade level change model due to comparing different groups of students in each grade from one year to the next. For example, with a cohort difference model it would not be a problem if the particular group of students enrolled in a school in one year happened to be a particularly bright group of students (Fletcher and Raymond, 2002). With a cohort difference model, this group of students will be compared to itself each year, as it progresses through the school. This eliminates one problem of using a performance model, which compares different groups of students in the same grade in subsequent years.

A cohort difference model can be preferable to a grade level change model in that it takes student background characteristics into account to the extent that it follows the same group of students (except for mobility) from grade to grade. An advantage of a cohort difference model over individual gain models is that all tested students influence the results of a school evaluation in a given year (Linn, 2001b). In an individual gain model, only students tested in two consecutive years influence the results of a school evaluation for either of those years. A cohort difference model is more simple than an individual gain model, in that individual students do not have to be tracked, and can be more inclusive in that students can be included if they were in a school only the second year, not the first (Linn, 2001a).

A problem with a cohort difference model would be encountered in the case of a state in which only a few grades (for example 3rd and 8th) are tested. The composition of the cohort may change significantly from 3rd to 8th grade. In a case such as this, it may be advisable to use a grade level change model. However, *No Child Left Behind* requires that states move toward testing all students in every grade, which would eliminate this particular problem.

There are other problems with longitudinal approaches such as cohort difference models. Because such models examine the progress of a cohort of students from one grade to another, it might benefit schools to exclude students as in the school status models, by putting them in special education programs or encouraging absences on test day. However, such exclusions are less likely because such models measure change rather than absolute performance (Hanushek and Raymond, 2002).

In addition, when using a cohort difference model, progress should be measured using scores on a common, or vertically equated, scale. A single scale is important so that scores from year to year are comparable, and growth can then be validly measured

(Linn, 2001a, 2001b). In other words, evaluation systems which compare performance of a group of students one year with the performance of the same group of students the next year must be able to make a clear distinction about whether the two sets of performances are equal or not. If two different tests are used, or a test which does not have a vertical scale covering the two grades, then it is impossible to tell how an average score one year relates to an average score the next year. In this situation, a scale score of, for instance, 1500 one year may not be at all comparable to a scale score of 1500 the next year on a different test. A cohort change model which compared average examination score of students in one year to the average examination score the previous year would require the examinations to be on a common scale in order for the average score to be truly comparable over the two years.

Cohort difference models can be seen as preferable to performance models in that they attempt to distinguish between variables which can be influenced by schools and those which cannot. The issue that they do not track exactly the same students can be overcome by individual gain models, which match students from year to year. Those models will be explained in the next section.

### *Individual Gain Models*

Individual gain models aggregate matched student gains to the classroom, school, or district level. Such models are being used in North Carolina, Tennessee, and in the Dallas Independent School District in Texas (Public Schools of North Carolina, 2003a; Webster, 1998).

Individual gain models can be referred to as value-added models. Value-added models use assessment of individual student scores matched across two or more years to evaluate schools. The average change in test score from year to year is considered to be appropriate growth for the average student, class, campus, or district. Growth above this

expectation is termed "value added." Teachers, schools and districts are rated according to the progress of their students. Campuses or teachers whose aggregated student scores are higher than average are considered to be adding value to the education of their students.

There is disagreement about the specific meaning of value-added models. It is generally accepted that value-added models are longitudinal models (Baker & Xu, 1995; Drury & Doran, 2003; Kupermintz, 2002; Meyer, 2002; Thum, 2003). This definition includes the assumption that changes in test scores from year to year are an accurate reflection of student progress in learning, and that tying this progress to particular schools allows for evaluation of the educational effects of such schools (Drury & Doran, 2003; Kupermintz, 2002; Tekwe, et al., 2004; Thum, 2003). Some value-added models use linear regression, and use non-school related variables in the regression equation to represent such non-school related influences (Webster, 1998) while others state that measuring individual gains isolates the influence of variables not controlled by the school (Sanders and Saxton, 1997).

There are various methods of calculating individual student assessment scores in order to evaluate schools. Some states, such as North Carolina, use single level multiple regression. Single-level regression models use a series of predictors to estimate the average test score of a student, class or school. The relevant predictors include previous test score and may also include other variables that can influence test scores, including SES, gender, and ethnicity. Actual test scores are then compared to predicted test scores, and ratings are made based on the difference between the actual and predicted scores for a teacher or school. Because important influences on scores are controlled for, there is an assumption in this situation that the differences between predicted and actual test scores are due to the effects of the influence of the teacher or school.

Both Linn and Haug (2002) and Fletcher and Raymond (2002) classify the system used in North Carolina as a cohort gain model. It really is an individual gain model, however, for the following reason. Both Linn and Haug (2002) and Fletcher and Raymond (2002) define cohort gain models as models that follow groups of students from year to year without matching individual student records. However, the system used in North Carolina matches individual student records in order to assess schools for accountability purposes. The Division of Accountability Services Reporting Section of the Public Schools of North Carolina publishes a document entitled "Determining Composite Scores in the ABCs Model." In this document it is stated that the first step in determining expected growth for end-of-grade tests is to "Determine the actual growth in reading and mathematics at each grade level in the school, using data on matched groups of students (i.e., students with both reading and mathematics 'pretest' and 'posttest' scores)" (Public Schools of North Carolina, 2003b, page 2, emphasis added). Expected growth is determined using average scores for cohorts from the first years the test was administered (1992-1993 minus 1993-1994). These average rates of growth are constants in the formula used to compute expected growth unless new values are approved by the State Board of Education (Public Schools of North Carolina, 2003a). Despite the fact that the statewide averages were set using cohort growth, schools are evaluated on individual matched students.

North Carolina has six categories for schools, based upon a performance composite and a growth/gain composite, otherwise known as the weighted expected growth composite. The first component, the performance composite, summarizes the percent of student test scores at the level of proficiency required by the state. The second component, the growth composite, is calculated using a school-level regression equation based upon the average rate of growth for the state. This single-level regression equation

models expected growth as an outcome of average growth by grade by subject (mathematics or reading) from the first year the test was administered. North Carolina does not include any student background variables in the consideration of school evaluations. The weighted expected growth composite for each school is computed using actual minus expected growth, which is standardized by dividing it by the standard deviation, then weighted by the number of students taking each test. Weighted standard expected growth is summed by subject and grade, and each school is rated on the weighted expected growth composite along with the performance composite (Public Schools of North Carolina Division of Accountability Services Reporting Section, 2003a, 2003B).

There are some strong advantages of the multiple regression model such as that used in North Carolina including the fact that it is easily understandable, relatively robust, and can accommodate both continuous and categorical predictor variables (Webster, Mendro, Orsak, and Weerasingle, 1998). Single level multiple regression models can be an attractive choice for states. There are, however, problems with the use of single level regression models.

One problem with using single-level regression is the multilevel structure of educational data. Students are clustered in classrooms, which are clustered in schools, which are clustered in districts. Multiple dependencies naturally occur under such nesting of levels of data. Use of single-level regression is based on the assumption that the students whose background variables and scores are used in the evaluation are statistically independent from each other. This means that the achievement scores of students in one classroom, or school, are assumed to be no more alike than the scores of students chosen randomly from throughout the state. It is not hard to see that the scores of students in a classroom who are taught by one teacher might well be more alike than

the scores of students chosen randomly from throughout the state, because of the shared experiences of the students in a classroom. When taken to the school level, the scores of students in a school are more alike than the scores of students chosen randomly, because they share the same principal, school rules, they probably come from the same or similar neighborhoods, etc. From a statistical point of view, the results of these evaluations that ignore the dependencies in test scores can artificially deflate standard errors if the standard errors are estimated using equations that assume statistical independence, with the result that they are likely to show differences between classrooms and schools more often than is true (Hox, 2002).

A problem with school-level regression also lies in the aggregation of student level variables to the school or district level. This practice assumes little within school variance, so that all students are impacted similarly by the aggregated variables. Raudenbush and Bryk, (2002) estimate that 80 to 90 percent of the individual variability on the outcome variable is lost in this case, which can lead to dramatic under-or over-estimation of observed relationships between variables. The greater the within-school variance, the poorer the estimates produced by such models (Mendro and Webster, 1993, as cited in Webster, et al., 1998).

Alternatively, if data are analyzed at the student level, the school variables are repeated exactly for each student in the school, which also gives a false impression of variability because the degrees of freedom used in the analysis are incorrect (Stevens, Estrada and Parkes, 2000). In this case, standard errors will again be artificially deflated, and significant results found more often than appropriate in significance testing.

A solution to the problems of single level multiple regression models is the use of multilevel models. The use of multilevel models avoids the dependency problems encountered using single level regression with nested observations. There is no need to

aggregate student-level effects to the school level, or to repeat school-level effects at the student level. Such models can separate the measured influences on scores due to student, teacher, school, or district characteristics.

## MULTILEVEL MODELING

There are a number of advantages to using a multilevel model when examining school or district achievement. These advantages address the problems of dependence between students within a school, as well as issues regarding aggregation or disaggregation, and cross-level interactions.

In order to examine the relation between student level achievement and SES within a school, a researcher might ignore the clustering within schools, and therefore might begin with the following equation:

$$Y_i=\beta_0+\beta_1 X+r_i \qquad (1)$$

In this equation, achievement for student $i$ is the dependent variable, $Y_i$. The intercept, $\beta_0$ is the mean achievement when SES for the school ($X$) is zero. $\beta_1$ is the slope, or expected change in achievement when SES changes by one unit. The error term $r_i$ is the deviation from this general equation, associated with student $i$. The error terms in this simple regression equation are assumed to be independent and normally distributed with a mean of zero and a variance of $\sigma^2$. Thus, $r_i \sim N(0,\sigma^2)$ (Raudenbush & Bryk, 2002). This assumption of independence is violated when analyzing school data, however, because students within a school are not independent from each other. Students from one school have more in common than students from different schools because the schools are run by the same principal, the same administration and obtain resources (from teachers to lunches) from the same sources.

Using a multilevel model in this situation allows for partitioning the variation in student achievement to the student level and the school level. Using this multilevel method effectively accounts for the lack of independence between students in the same school.

***Unconditional Model***

The simplest step in constructing a multilevel linear model is to estimate a fully unconditional (or "null") model. In this case, there are no predictors at any level. Thus, the model is written at level one:

$$Y_{ij}=\beta_{0j}+r_{ij} \qquad (2)$$

where $Y_{ij}$ is the outcome variable, achievement, for student $i$ in school $j$, $\beta_{0j}$ is the average achievement in the school $j$, and $r_{ij}$ is the error associated with student $i$ in school $j$. A second (school) level equation would then be written as:

$$\beta_{0j}=\gamma_{00}+u_{0j} \qquad (3)$$

where $\beta_{0j}$, the average achievement in school $j$, is predicted by $\gamma_{00}$, the average overall achievement, and $u_{0j}$ is the random effect associated with school $j$, with a mean of 0 and variance $\tau_{00}$.

Substituting the level 2 equation into the level 1 equation yields

$$Y_{ij}= \gamma_{00}+u_{0j} +r_{ij} \qquad (4)$$

This multilevel model is equivalent to a one-way random effects ANOVA, with a grand mean of $\gamma_{00}$, a group, or school effect $u_{0j}$ and an individual, or student effect $r_{ij}$. In ANOVA terminology, $u_{0j}$ represents between-schools error, or deviation of schools from the grand mean, and $r_{ij}$ represents within-schools error, or deviation of individual students from the group, or school, mean (Raudenbush & Bryk, 2002).

It is important in multilevel modeling to distinguish between random and fixed effects. In single level regression, the intercept and slope parameters are considered to be

fixed coefficients. In multilevel regression, the first, or lower, level intercept and slope can be considered as random coefficients. That is, they can be modeled to vary across the second level (Kreft & DeLeeuw, 1998). In this case, values of the intercept and slope at the student level are modeled to vary across schools.

At this point, as part of estimating the fully unconditional model, it is useful to estimate the intraclass correlation. The intraclass correlation measures the portion of total variation in the dependent variable (student-level achievement) that is explained by the second level groups (schools). If the intraclass correlation is zero, then there is no need to use a hierarchical linear model because only a small proportion of the total variation is due to differences between groups. In other words, the average correlation between outcome variables of students within a school is no higher than the correlation between students from different schools. The hierarchical structure of the data can be ignored, the assumption of independence of student scores holds, and a single level regression model will describe the data just as well (Hox, 2002). The formula for the intraclass correlation is:

$$\rho = \tau_{00} / (\tau_{00} + \sigma^2) \qquad (5)$$

where $\rho$ represents the intraclass correlation consisting of the ratio of between-school variance ($\tau_{00}$) to the total variance ($\tau_{00} + \sigma^2$). If the intraclass correlation is not zero, then a portion of total variance is due to between group variance, the assumption of independent observations is violated, and thus it is appropriate to use a multilevel model (Kreft & De Leeuw, 1998).

***Adding a Level I Predictor***

If it is found that the intraclass correlation is not zero, indicating that there are between-group effects and multilevel modeling is appropriate, it is then useful to add a level-1 (student level) predictor to explain variability in the dependent variable. Using

student level SES, X, as a predictor, the level 1 model (equation 2) can be expanded and written as

$$Y_{ij}=\beta_{0j}+\beta_{1j}(X_{ij}-\overline{X}\,..)+ r_{ij} \qquad (6)$$

where $Y_{ij}$ is the student level achievement, $\beta_{0j}$ is the adjusted mean for group $j$, or the expected outcome for student $i$ in school $j$ with an SES value equal to the grand mean, $X...$ $\beta_{1j}$ is the change in achievement in school $j$ per unit change in SES, $X_{ij}$ is the SES level of student $i$ in school $j$, and $r_{ij}$ is the unique effect associated with student $i$ in school $j$. The effect of the term $(X_{ij}-X..)$ is to center the variable $X_{ij}$ around the grand mean $X..$, for ease of interpretation. When $X_{ij}$ is not centered, the intercept $\beta_{0j}$ represents the expected outcome for student $i$ in school $j$ with a value of 0 on $X_{ij}$. When $X_{ij}$ is centered around $X..$, $\beta_{0j}$ is the expected outcome for student $i$ in school $j$ with an SES value equal to the grand mean, $X$ (Randenbush and Bryk, 2002).

The difference between equation 6 and equation 1 is the assumption that each school has a different intercept $\beta_{0j}$, and a different slope $\beta_{1j}$ (Hox, 2002). Differences in the intercept, $\beta_{0j}$, indicate differences between schools in the mean student achievement level when SES is equal to zero. A difference in the slopes ($\beta_{1j}$) indicates differences in the relationship between SES ($X_{ij}$) and achievement ($Y_{ij}$) across schools. Schools with a high $\beta_{1j}$ have a stronger relation between SES and achievement. Schools with a value for $\beta_{1j}$ closer to zero are schools in which SES has less impact on achievement.

When adding a level 1 predictor, the level 2 equation for the intercept (equation 3) stays the same, written below as equation 7. The coefficient for the predictor X can be modeled to vary across schools. Thus, the level 2 model, with no level 2 predictor, becomes

$$\beta_{0j}=\gamma_{00}+u_{0j} \qquad (7)$$

$$\beta_{1j}= \gamma_{10}+u_{1j} \qquad (8)$$

where $\gamma_{00}$ is the average achievement level across schools, given $X$ (SES) is zero, and $u_{0j}$ is the residual unique to the intercept, or average achievement, for school $j$. Similarly, $\gamma_{10}$ is the slope, or average change in achievement across schools, and $u_{1j}$ is the residual unique to the slope, or change in achievement due to $X$ (SES), for school $j$. Note that it is also possible to model either $\beta_{0j}$ or $\beta_{1j}$ as fixed coefficients, in which case they would be modeled such that they would not vary across schools.

Substitution of equations 7 and 8 into equation 6 yields the following equation:

$$Y_{ij}=\gamma_{00}+\gamma_{10}(X_{ij}-\overline{X}\ldots)+u_{0j}+u_{1j}(X_{ij}-\overline{X}\ldots)+r_{ij} \quad (9)$$

where $Y_{ij}$, achievement at the student level, is predicted by average achievement across schools ($\gamma_{00}$) and the average change in achievement due to student SES ($\gamma_{10}X_{ij}$) plus the random effects of school on the mean ($u_{0j}$) the random effects of school ($u_{1j}$) and random effects of students ($r_{ij}$).

***Adding both Level 1 and Level 2 predictors***

It is also possible to add a level 2, school level, predictor W to equation 9. Equations 7 and 8 change with the introduction of a predictor at level 2, the school level. These equations become:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \quad (10)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \quad (11)$$

where, in Equation 10, $W_j$ is the value of, for example, school $j$'s percent minority students, $\gamma_{00}$ is the mean overall achievement level when percent minority, $W$, is 0, $\gamma_{01}$ is the change in achievement with each unit change in school percent minority, $u_{0j}$ is the unique effect of school $j$ on the average achievement. Thus, Equation 10 predicts the average achievement in a school ($\beta_{0j}$) given $X$=0 using school percent minority ($W_j$) as a predictor (Hox, 2002). In Equation 11, $\gamma_{10}$ is the average achievement-SES slope when school percent minority is zero, $\gamma_{11}$ is the average student SES-achievement slope

difference given a change of one unit in a school's percent of minority students, and $u_{1j}$ is the unique effect of school $j$ on the SES-achievement slope. Equation 11 states that the relation ($\beta_{1j}$) between achievement ($Y$) and SES of the student ($X$) depends upon the percent of minority students in the school ($W$). Given positive within school slopes, if $\gamma_{11}$ is positive, the relationship between student SES and achievement is stronger in schools with higher percent minority students. If $\gamma_{11}$ is negative, there is a weaker relationship between student SES and achievement in schools for schools with a higher percent of minority students (Hox, 2002). The random $u$ coefficients, with $j$ subscripts, represent the between-school variation remaining in the $\beta$ coefficients after they are explained using the $W$ school-level variables (Hox, 2002).

Equation 11 indicates the assumption that the relation ($\beta_{1j}$) between school achievement and the SES ($X$) of the student, depends upon school percent minority students ($W$). The hypothesis would be that whether the school has a high or low value for $\beta_{1j}$ depends to some extent upon the percent minority students in the school. Thus school percent minority acts as a moderator variable for the relationship between school achievement and SES (Hox, 2002). The error terms $u_{0j}$ and $u_{1j}$ are at the school level, with mean of zero, and are assumed to be independent from the student-level error, $r_{ij}$. However, the covariance between the errors at the school level ($u_{0j}$ and $u_{1j}$) is not assumed to be zero (Hox, 2002).

When Equations 10 and 11 are substituted into Equation 6 to predict student achievement, the single equation becomes:

$$Y_{ij}=\gamma_{00}+\gamma_{01}W_j+\gamma_{10}(X_{ij}-\overline{X}\ ..)+\gamma_{11}W_j(X_{ij}-\overline{X}\ ..)+u_{0j}+u_{1j}(X_{ij}-\overline{X}\ ..)+r_{ij} \quad (12)$$

In this equation, $W_jX_{ij}$ is an interaction between a student-level predictor ($X$) and a school-level predictor ($W$). This is termed a cross-level interaction (Hox, 2002). After

controlling for $W$, the variance of $u_{0j}$ is the variance of the intercepts between schools, and the variance of $u_{1j}$ is the variance of the slopes between the schools.

An additional important note with regard to Equation 12 is that the term $u_{1j}X_{ij}$ means that the school-level error ($u_1$) interacts with the student level explanatory variable $X$, which means that total error will vary according to the value of that explanatory variable, student-level SES. This situation is referred to as heteroscedasticity, meaning that error and explanatory variables are not independent of each other, which violates the assumptions of ordinary single-level regression. This is another reason that multilevel modeling can be used to model data such as student achievement, where students are clustered in schools, better than single level multiple regression (Snijders & Bosker, 1999).

***Multilevel school effects models***

The evaluation system used in the Dallas Independent School District of Texas incorporates both a single-level regression and a hierarchical linear regression. The method used in Dallas first uses a single level regression to remove the effects of student ethnicity, gender, and SES, which they call "fairness" variables, by regressing student assessment scores for the current and previous years on the student contextual variables over which the schools had no control. This is done through the use of the single-level linear regression model described below:

$$Y_i=\beta_0+\beta_1X_i+...+\beta_jX_j+r_i$$

In this equation, $Y_i$ represents the criterion or outcome variable, such as achievement in the prior or current year; $X_j$ represents "fairness" predictor variable $j$ such as SES.

A single-level regression equation is created and the dependent variable for the single-level regression is the variable that will be used as a predictor in the second stage

regression, such as pretest (test from the previous year) or current year test score. A residual for each outcome of these first stage regressions is calculated for each student (Webster, 1998).

In the second stage of the analysis the residuals from the first stage regression outcome variables, such as prior year test score, are used as the first (student) level predictors of the residuals from other first stage regression outcome variables, such as the current year test score. In this Dallas system, students represent the first level and schools the second level of the multilevel model. Equations for this model are:

Level 1: $$Y_{ij}=\beta_{0j} + \beta_{1j}X_{1j} + ...\beta_{kj}X_{kij} + r_{ij}$$

Level 2: $$\beta_{0j}=\gamma_{00} + \gamma_{01}W_j + u_{kj}$$

$$\beta_{1j}=\gamma_{10} + \gamma_{11}W_j + u_{kj}$$

$$\beta_{kj}=\gamma_{k0} + \gamma_{k1}W_j + u_{kj}$$

In this equation, $Y_{ij}$ is the residual student outcome variable, such as current year test score, for student *i* in school *j* from the first stage single-level regression; $X_{ij}$ are residuals of prior achievement or attendance variables predicted from the "fairness variables" in the first stage single-level regression; $W_j$ represent school-level predictors such as percent minority students.

It should be noted that the relation between the level one predictors (first year achievement test residual) and the outcome (second year achievement test residual) is modeled to vary randomly across schools. Tekwe, et al., (2004) found that there was little impact on model results of empirical data when school effects were considered to be random rather than fixed effects.

The school level predictor variables including mobility, crowdedness, percent minority, percent black, percent Hispanic, percent free-lunch, and average census

variables are used to predict the intercept and slopes of the student level model (Clotfelter and Ladd, 1996).

The results of the multilevel model for each school are weighted using weights derived by an Accountability Task Force, and combined to give a measure of school effectiveness for each school (Webster, 1998).

Thum and Bryk (1997) reviewed the Dallas model and expressed two concerns. The first is the use of the two stage process, utilizing a single-level multiple regression followed by a multilevel model. They suggest it might have been better to include the "fairness variables" simultaneously in the first level of the two-level multilevel model. There is then no need to run the single-level regression model. In addition, the authors express concern about the use of residuals, stating,

> In general, residuals are notoriously unreliable. Residuals from a set of residuals analysis would only further compound this problem, leading to serious questions concerning the reliability of the productivity indicators. (p. 106)

They suggest that a one-stage multilevel model would be just as effective, and more efficient.

Webster, Mendro, et al (1997) respond to the above criticisms stating that analysis shows one-stage and two-stage models produce effect outcomes with correlations of those outcomes above .978. There is no indication, however, as to which schools and which groups of students are affected by the imperfect correlation of outcomes.

It is possible to apply a one-stage multilevel model in school evaluation. Tennessee uses a multilevel model to evaluate teachers, schools and school districts. The Tennessee Value-Added Assessment System (TVAAS) is a student gain model which does not include demographic variables at the student level, instead the assumption is that individual student gains should be similar for all students, regardless of the student's

achievement level when starting school (Sanders, et al., 1997). This assumption has not been corroborated.

The TVAAS process uses a general form of Henderson's Mixed-Model (Henderson, 1973) equation, which is

$$Y=XB + ZU + e$$

where $Y$ is an $m \times 1$ observation vector representing all of the scale scores for individual students in all tested subjects over all grades; $X$ is a known $m \times p$ incidence matrix that allows for the inclusion of fixed effects; $B$ is an unknown $p \times 1$ fixed vector that is to be estimated from the data; $Z$ is an $m \times q$ incidence matrix, $U$ is an unobservable $q \times 1$ random vector representing the realized values of the random effects to be estimated from the data; and $e$ is an unobservable $m \times 1$ random vector variable representing unaccountable random variation (Sanders, et al., 1997).

The school model used in the TVAAS process is:

$$Y_{ioklmn}=u_{ioklm} + e_{ioklmn}$$

Where $Y_{ioklmn}$ represents a test score in the $m^{th}$ subject for the $n^{th}$ student, and the student was in the $i^{th}$ school system, $o^{th}$ school, $k^{th}$ year, and $l^{th}$ grade;

$u_{ioklm}$ is the fixed school mean score for all students in the $i^{th}$ school system, $o^{th}$ school, $k^{th}$ year, $l^{th}$ grade, and $^{th}$ subject; and

$e_{ioklmn}$ represents the random deviation of the test score for the $n^{th}$student from the school mean (Sanders, et al., 1997).

The Tennessee system uses multiple test scores for multiple years for each student, and does not incorporate demographic variables into the multilevel equations. Sanders, et al., (1997) claim that incorporating at least 3 years of data in each analysis means that "Therefore, the personal fortunes of an individual student have little influence on the estimation of educational effects" (180). Tekwe, et al., (2004) conducted research

using two years of data and found that the results of school classifications using the TVAAS procedure correlated highly (0.91 to 0.98) with the results of a simple fixed effects model, concluding, "It is possible that greater discrepancy of results would be found in studies of three or more years of data" (26).

By modeling the intercept and slope as random coefficients, allowing them to vary across the second level of the model, multilevel modeling allows for the dependency of observations in clustered data, and also allows for the use of first- and second-level data, thus eliminating the issues of level of aggregation of data and cross-level effects.

Problems with all value-added school accountability models include the question of the extent to which the change from year to year in the student's score is influenced by more than just school, and the difficulty of separating Type B effects from Type A effects. In addition, models that follow students across years exclude from analysis the mobile students who change schools or move out of the state between tests, or who for some reason are not tested both years. The school or district being evaluated is not held responsible for these students (Linn, 2001b). A problem shared with other models of school accountability is that students tested in the first year who are known not to be gaining quickly can be encouraged to be absent the second year, thus excluding them from the analysis (Hanushek and Raymond, 2002).

Another potential problem with all value-added models is the belief that in order to most accurately measure the value added by schools and teachers, there must be annual testing in every grade and it must be possible to equate these scores to the same scale each year (Lee & Weimer, 2002). Ballou, et al, (2004) point out that measuring progress can be done either with a difference score, created by subtracting previous year test score from current year score if the two tests are on the same scale, or by using current year test score as the outcome and incorporating the earlier test score as a covariate.

Kingston and Reidy (1997) raise a problem with the use of student gain models, in that the improved reliability of a matched student system comes at a great cost in validity, because so many students are lost when a system matches on students who have been in a school or district for two years. As few as 66 percent of student records can be consistent from year to year (Kane and Staiger, 2001). In Kentucky, Kingston and Reidy (1997) suggest within-year transient rates of 20 percent or more, and losing such numbers of students has the potential for greatly affecting the accountability system. They suggest a cohort difference model as a preferred compromise between the problems of performance models and individual gain models.

The next section will examine results of comparisons of school evaluations determined using different performance and productivity models.

## Previous Comparisons

Comparisons of statistical models used for school evaluations have illuminated two problems associated with choice of statistical models. One is the comparability of the results from different models of school accountability. When the results of accountability decisions made using different models do not correlate highly, and serious consequences result from the evaluation results, it is especially important that the choice between models be considered extremely carefully.

The other dilemma is whether the results of each model are closely correlated to variables beyond the control of the campuses themselves. In a system which emphasizes adequate education for all children, if schools with high percentages of low socioeconomic students fall consistently to the bottom of rating systems, then the model might be considered biased. Webster et al. (1998) have laid out clear criteria for the evaluation of the level of bias in accountability models. They note that the final criterion

must be the degree to which these models control for variables known to be related to the outcome variables but not under control of the school or teacher (SES, for example). The degree to which correlation between school ranking and student background characteristics are nonzero is the degree to which a model is biased. Furthermore, correlations between ranking and such characteristics must be controlled at both the student and school or classroom level (Mendro, 1998).

Raudenbush and Willms (1995) address this dilemma differently. They suggest that differences in the purpose of the school evaluation determines the extent to which inclusion of student background and school context variables should be considered. For parents intending to pick the best of all possible schools for their child, Type A effects, which do not isolate different influences on achievement scores, are most important. For state or federal governments attempting to isolate the Type B effects, which are under control of the school, then student background and school contexts should be removed from influence on the school classification.

Using data from the Dallas Independent School District in Texas, a series of comparisons were summarized by Webster, et al. (1998). The results of a variety of statistical models applied to student data from the Dallas Independent School District were compared to a two-stage, two-level student-school or student-teacher hierarchical linear model and to student background and school context variables. Models were compared through the correlation of the school or teacher effectiveness results of each model with results from the two-stage, two-level student-school multilevel model, along with the correlation of results of models with student background variables such as SES.

The results of a performance model using unadjusted mean school test scores correlated 0.508 with the results from the multilevel model, which incorporated student background variables at both the student and classroom levels. Results from this

performance model correlated as high as 0.648 with a measure of SES (parental education level) (Webster, et al. 1995, as cited in Webster, et al. 1998).

Webster, et al. (1998) also compared a single-level school regression model to the two-stage, two-level student-teacher multilevel model with student background variables included and found correlations between results from the two models equal to 0.8637 (Webster, et al. 1997, as cited in Webster, et al. 1998). Correlations between the results of one-stage and two-stage multilevel models were around 0.95.

A final study by Webster, et al. (1998) compared the results from multilevel model analyses using fixed versus random slopes. Results correlated around 0.98, but they report that the solution for Thum and Bryk's (1997) suggested one-stage multilevel models assuming random slopes with "a full array" of contextual variables did not converge. Webster et al. (1998) included the following as student-level predictors: Black and Hispanic indicator codes, limited English proficient status, gender indicator, free and reduced lunch status, census-block average family income, census-block average family education level, and census-block average family poverty level. At the school level, they included mobility, overcrowdedness, average family education, average family poverty, percentage of students on free and reduced price lunch, percentage minority, percentage black, and percentage Hispanic, percent limited English proficient, and percent instructional days lost to unfilled vacancies. The multicollinearity among the variables in the "full array" may be so high that many of the variables are not necessary as separate predictors, and dropping some of those variables could have resulted in a solution that converged.

Alban (2002) compared teacher and school rankings from a single-level multiple regression, a multilevel regression, and Tennessee Value-Added Assessment System (TVAAS) results for teachers and schools from two school districts in Maryland. Student

level achievement, measured by scale score on the Maryland School Performance Assessment Program, was used as the dependent variable. Predictors included student gender, race, a prior achievement score on a different test, English as a second language status, special education status, and a measure of SES. School-level predictors included percent of low SES students, percent of student receiving special services (English as a second language or special education) school enrollment, mobility, and ethnicity of the population of students. Scores on tests in up to five subject areas (language, writing, mathematics, science and social studies) for different grade levels were analyzed. In total, up to 105 statistical comparisons (21 tests times five subject areas) were conducted using single-level multiple regression (one full multiple regression, ten analyses using random samples of 20 students per teacher, and ten analyses using random samples of 10 students per teacher).

Alban (2002) examined the relative significance of each variable as a function of the model being used separately for District A and District B, and found that the most consistently significant predictor of achievement (105 times in 105 tests for District A, 63 times in 63 tests for District B) in the single-level multiple regression analyses for both districts was previous test score. Minority status was fairly often (56 times in 63 tests) found to be significant at District B, but rarely (22 times in 105 tests) at District A.

Using the two-level multilevel model with student and school effects, Alban (2002) found almost no significant school level effects, and only student's gender and previous achievement in every content area were statistically significant for District A, while all effects were statistically significant for District B.

Schools were ranked according to the difference between their own actual mean performance and the mean performance as predicted from the regression equations. Specifically, Alban compared school rankings using empirical Bayesian residuals for the

intercept for the multilevel modeling analyses to the rankings using the mean of all standardized residuals in a school from the multiple regression analyses. Rankings were then divided into quintiles for comparison, in order to model the process used in TVAAS. Rankings for school system A were consistent (being ranked in the same or an adjacent quintile under two different models) for 64 to 80 percent of schools, and school system B had between 75 to 88 percent of schools ranked consistently.

Clotfelter and Ladd (1996) examined data from schools in South Carolina using different models. Results calculated from ranking schools using a performance measure (average score) correlated 0.36 with rankings from a cohort difference model, and 0.58 with results from a model using single-level regression with SES included as a predictor (Clotfelter and Ladd, 1996). The single-level regression model was the only model examined which included any student contextual variables.

Clotfelter and Ladd (1996) also report the correlation of the school performance evaluation results with the demographic characteristics of students. The correlations of rankings with SES for the performance model for which demographic characteristics were not included averaged -0.71. For the cohort difference model the correlation averaged -0.09, and zero for the single-level regression with SES included as a predictor. These studies indicate that the rankings of schools and their relation with percent of low socioeconomic students in schools are heavily dependent on the type of model being used.

Tekwe et al. (2004) compared various types of value-added models using empirical data, including Sanders' layered mixed effects (or TVAAS) model, a simple change score fixed effects model, and two multilevel models, one of which was adjusted for student demographic characteristics, and the other an unadjusted multilevel model. Results of school grade (A-F) assignment based on standardized value-added measures

from each model were correlated to determine consistency of results. Grades based on the simple change score fixed effects model were found to be highly (0.91 to 0.98) correlated with grades based on the layered mixed effects model. It was noted that school grades based on these two models might be less highly correlated if more than two years of data were examined. The layered mixed effects (TVAAS) model relies on covariance of scores between subjects and years to "replace" the effects of inclusion of demographic variables in prediction equations. Comparison of the grades based on the multilevel models with the difference being inclusion of demographic characteristics showed very little agreement between model results, with results correlating between 0.61 and 0.95, with only two correlations being higher than 0.80.

Tekwe et al. (2004) concluded that the simple fixed effects model is to be preferred over the complex mixed effects model, because school grades resulting from the two models were highly correlated, and there is little or no benefit from using the more complex model. Whether or not to use the simple fixed effects model or an adjusted multilevel model depends on the purpose of the school evaluation.

## Statement of Problem

State education agencies using statistical models to rank schools for accountability purposes may wish to evaluate schools on Type A effects, in which case any of the above models can be appropriate. If, however, the goal is to evaluate schools on Type B effects, the school effects should not be sensitive to the SES of the students. In this case, choice of a statistical model may be more important to the results.

Clotfelter and Ladd (1996) report a low average correlation (-0.09) of the results of a cohort difference model to percent of low SES students in a school. Given Kingston and Reidy's (1997) suggestion of cohort difference models as a preferred compromise

between performance models and individual gains models, the results of cohort difference models should be further investigated. Cohort difference models have not been compared to single-level regression models or multilevel regression models to determine the differences between models which use cohorts of unmatched students and models which match individual students across years.

As reported earlier, Webster et al. (1998) compared a one-stage multilevel model to a two-stage model, and reports possible suppressor effects in the one-stage model which have not been corroborated. Studies have not otherwise compared the two-stage Dallas method as currently used with the one-stage method suggested by Thum and Bryk (1997).

This dissertation is designed to address the following questions:

1. Are the different statistical approaches equally sensitive to the inclusion or exclusion of student SES? How similar are results of school rankings based on each model without this variable compared to results of the same model with student SES included?

2. Are the different statistical approaches equally unbiased against schools with a high percentage of low SES students? That is, how do the rankings of schools calculated using the different models including or excluding SES correlate with the percent of low SES students in the school?

3. Are the different statistical approaches comparable? How do rankings calculated using each model compare to rankings calculated using each of the other models?

4. How do rankings calculated using the various statistical models differ as the relation between student SES and student achievement varies?

5. How do rankings calculated using the various statistical models differ as the relation between school percent low SES students and average school achievement varies?

6. How do rankings calculated using the multilevel models differ from rankings calculated using performance, cohort difference, and single-level regression models as the portion of total variation in student achievement explained by schools varies?

# CHAPTER 3: METHODS

Two studies were conducted to address the research questions in this dissertation. In the first, school rankings from four types of accountability models were compared using real test score data. Relationship between SES composition and school ranking was examined to assess the potential for bias in rankings based on each model against schools with a high percentage of low SES students. Similarity between rankings calculated using the different models was also assessed. In the second study, the relationship between SES composition and school ranking based on different statistical models was further examined with a simulation study. The statistical models and the outcomes using real test score data will be described next.

## Real Data Study

**SAMPLE.** Data for approximately 1,000,000 3rd, 4th and 5th grade students (approximately 250,000 3rd grade and 250,000 5th grade students, and 500,000 4th grade students) from the state of Texas who were administered the mathematics or reading portion of the Texas Assessment of Academic Skills (TAAS) in the 2000-2001 school year and the 2001-2002 school year were used for this study. For models requiring matched sets of students, students who participated both years were included. For all other models, all students who attended and were tested one or both years were included.

Information on each student included student ID number, score on mathematics and reading tests, SES (as measured by free and reduced price lunch participation), and school ID number. Special education or limited English proficient students taking alternative tests were not included in this study.

**ASSESSMENT.** Scores on the criterion-referenced TAAS used for this study exhibit internal consistency reliabilities in the high .80s to low .90s range. Kuder-Richardson Formula 20 has been utilized for the dichotomously scored items, and stratified coefficient alpha was utilized for tests containing a mix of dichotomous and polytomous items. Stratified coefficient alpha refers to an extension of coefficient alpha to accommodate a mixture of item types on a test. In this case, items of the same type, such as multiple choice or essay, are combined and internal-consistency reliability computed for all of the items of one type, as if those items were a separate subtest. Estimates of reliability for each item type component are then combined and reported as the stratified coefficient alpha (Texas Education Agency, 2003). Student assessment scores on TAAS are reported on one scale.

The scale used for reporting assessment scores is the Texas Learning Index (TLI). The TLI is not a vertical scale. Rather, it is a scale measuring student performance compared to the student population tested in the spring of 1994. A student attaining a reading TLI of 70 in one year and a reading TLI of 75 in the next year is described to have demonstrated higher than typical progress between grades, and to have changed his or her relative rank-ordered position inside each grade distribution (Texas Education Agency, 2003).

Procedures used to assess and support test scores' content and construct validity are reported in the technical digest (see Texas Student Assessment Program Technical Digest, 2003). The test is aligned with the state curriculum through the inclusion of numerous educator committees. No convergent validity coefficients were reported (Texas Education Agency, 2003).

Student and school level achievement measures from the criterion-referenced TAAS were used as the dependent variables in a series of analyses using four different

types of accountability models. The school rankings resulting from application of each model were compared to the school rankings resulting from application of each of the other models, and results were also compared to the socioeconomic makeup of the students in the school. Each school was ranked once for each subject (reading, mathematics) and once for each grade ($4^{th}$, $5^{th}$) for each model.

**MODELS.** The following models were used for comparison. All models are summarized in Tables 1 through 3. Model A, the percent passing (PCTPASS) model, is a performance model of the type used in federal accountability (NCLB) requirements. This model ranks schools based on the percent of students in fourth and fifth grades for the second year of testing who passed the reading and mathematics tests. SES is not included in this NCLB model, because of difficulties ranking all schools on one scale when schools are evaluated in separate groups depending on SES. Schools were given four rankings for this model; percent of fourth grade students passing the mathematics and reading assessments, and percent of fifth grade students passing the mathematics and reading assessments. The PCTPASS model is described in Table 1.

Model B is a cohort difference (COHDIFF) model and is summarized in Table 1. This model evaluates the change in test scores from 2001 to 2002. First, the group of students in third grade in the first year, and fourth grade in the second year, and second the group of students in fourth grade in the first year and in fifth grade in the second year were included. Students who had valid test scores for at least one of the two years were included in the analysis under this model. For reasons similar to those involved in the school ranking using results of the percent passing model, this model does not include SES in the ranking of schools. Schools were ranked by average amount of difference in the reading test score and the mathematics test score the second year for the two groups (third to fourth grade, and fourth to fifth grade). Average amount of difference for fourth

grade was calculated by subtracting the school average Texas Learning Index (TLI) for math for third grade in the first year of testing from the school average TLI for math for fourth grade in the second year of testing. The same calculation and ranking method was performed for reading. Average amount of difference for fifth grade was calculated by subtracting the school average TLI for math for fourth grade in the first year of testing from the school average TLI for math for fifth grade in the second year of testing. The same calculation and ranking method was performed for reading. The COHDIFF model is described in Table 1.

Table 1: Models used in Analysis: Percent Passing and Cohort Difference

| **Model** | **Model Type** | **Analysis Strategy** | **Ranking Outcome** | **Predictors** |
|---|---|---|---|---|
| **PCTPASS** | Performance | Descriptive | Percent passing | N/A |
| **COHDIFF** | Cohort difference | Descriptive | School average gain | N/A |

Model C is a single-level school regression model using school average second year test score as the outcome and school first year average test score as the predictor. Cohorts of students are those who were in third grade the first year and fourth grade the second year, and fourth grade the first year and fifth grade the second year. Separate regression models were created for mathematics and reading for the two cohorts of students. Model C1 (UNSLR) includes first year school average test score as a predictor. Model C2 (ASLR) includes first year school average test score and school percent low SES students as predictors. The single-level regression models used the same cohorts of students as the cohort difference model. Schools were ranked for each test (reading, mathematics) for each grade (4, 5) by the residuals resulting from application of the regression equations. Models UNSLR and ASLR are described in Table 2. .

Table 2: Models used in Analysis: Unadjusted and Adjusted School-Level Regression

| **Model** | **Model Type** | **Analysis Strategy** | **Ranking Outcome** | **Predictors** |
|---|---|---|---|---|
| | School-level change | Single-level regression | | |
| **UNSLR** | One predictor | | School residuals | School mean first year test score |
| **ASLR** | Two predictors | | School residuals | School mean first year test score; school percent free and reduced price lunch students |

The regression equations used for the school-level regression models are presented in Table 3.

Table 3: Detail of School Level Regression Configurations

| **Model** | **Model Outcome (Y)** | **Level 1 Equation** |
|---|---|---|
| **UNSLR** | School average second year test score | $Y_j = \beta_0 + \beta_{1j}X_j + r_j$ |
| **ASLR** | School average second year test score | $Y_j = \beta_0 + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + r_j$ |

$X_{1j}$ = School average first year score
$X_{2j}$ = School percent free and reduced price lunch students

Model D is a multilevel model, with all equations summarized in Table 3. Models D1 (DLINT) and D4 (DLINTSLP) are two-stage models similar to that used in the Dallas, Texas, school district accountability system. DLINT is the Dallas model with random intercepts, and DLINTSLP is the Dallas model with random intercept and slopes. Models D2 (UNINT), D3 (AINT), D5 (UNINTSLP) and D6 (AINTSLP) are the one-stage models that Thum and Bryk (1997) suggested would be just as effective as, but more efficient than, the Dallas model. UNINT refers to the unadjusted random intercept model, AINT to the adjusted random intercept model, UNINTSLP to the unadjusted random intercept/slope model, and AINTSLP to the adjusted random intercept/slope model.

Model D1 (DLINT) is the first of the two-stage Dallas-type models. The first stage is a series of single-level regression equations for mathematics and another series of single-level regression equations for reading. The first regression equation in the first stage uses student SES as measured by participation in the free and reduced price lunch program to predict student test score in the first year of testing. The second regression equation uses student free and reduced price lunch participation to predict student test score in the second year of testing. Free and reduced price lunch is a dichotomous variable, with a value of 1 indicating that the student does not participate, and a value of 0 indicating that the student does participate. The equations for the first stage of this model are:

First year test score = student SES + residual

Second year test score=student SES + residual

The second stage of the Dallas DLINT model is a multilevel regression using residuals from the first stage second year test score regression as the outcome and residuals from the first stage first year test score regression as the predictor, with students as the first level and schools as the second level of the multilevel model. Average school-level percent free and reduced price lunch students is included as a level two predictor.

Second year residuals = first year residuals + school percent low SES

Because this model uses SES (as measured by free and reduced price lunch) to predict first year and second year test score, and residuals from those first stage regressions are used in the second stage of the model, it is not possible to include a version of this model without SES included.

Model D2 (UNINT) is a one-stage two-level model which uses only previous test score as a predictor at the student level. Model D3 (AINT) is a one stage two-level

model using previous test score and student free and reduced price lunch status as predictors at the first level, and school level percent free and reduced price lunch students as a predictor at the second level. The configuration of these three multilevel models includes random intercepts with fixed slopes.

Model D4 (DLINTSLP) is a Dallas-type two-stage model identical to DLINT except that in addition to random intercepts, slopes were modeled randomly across schools. Model D5 (UNINTSLP) is a one-stage model with random intercepts and slopes, using only previous test score as a predictor at the student level. Model D6 (AINTSLP) is a one stage two-level model with random intercepts and slopes using previous test score and student free and reduced price lunch status as predictors at the first level, and school level percent free and reduced price lunch students as a predictor at the second level. Models DLINT, UNINT, AINT, DLINTSLP, UNINTSLP and AINTSLP are described in Table 4.

Table 4: Models used in Analysis: Dallas-Type, Unadjusted and Adjusted Multilevel Regression

| **Model** | **Model Type** | **Analysis Strategy** | **Ranking Outcome** | **Predictors** |
|---|---|---|---|---|
| | Individual change | Hierarchical linear regression | | |
| **DLINT** | Two stage model | Random intercept Fixed slopes | School residuals | 1st stage: student free and reduced price lunch status; 2nd stage: residuals from first year regression; school percent free and reduced price lunch students |
| **UNINT** | One stage model | Random intercept Fixed slopes | School residuals | Student first year test score |

| | | | | |
|---|---|---|---|---|
| **AINT** | One stage model | Random intercept Fixed slopes | School residuals | Student first year test score; student free and reduced price lunch status; school percent free and reduced -price lunch students |
| **DLINTSLP** | Two stage model | Random intercept Random slopes | School residuals | 1st stage: student free and reduced price lunch status; 2nd stage: residuals from first year regression; school percent free and reduced price lunch students |
| **UNINTSLP** | One stage model | Random intercept Random slopes | School residuals | Student first year test score |
| **AINTSLP** | One stage model | Random intercept Random slopes | School residuals | Student first year test score; student free and reduced price lunch status; school percent free and reduced price lunch students |

Predictors used in all multilevel models were grand mean centered, and schools were ranked using empirical Bayesian residuals for the intercept. The equations for all multilevel models are presented in Table 5.

Table 5: Detail of Multilevel Configurations

| **Model** | **Model Outcome (Y)** | **Level 1 Equation** | **Level 2 Equation** |
|---|---|---|---|
| **DLINT** | First stage residuals | $Y_{ij}=\beta_{0j} + \beta_{1j}X_{1j} + r_{ij}$ | $\beta_{0j}=\gamma_{00} + \gamma_{01}W_j + u_{0j}$<br>$\beta_{1j}=\gamma_{10}$ |
| **UNINT** | Second year test score | $Y_{ij}=\beta_{0j} + \beta_{1j}(X_{1j}- \overline{X} ..) + r_{ij}$ | $\beta_{0j}=\gamma_{00} + u_{0j}$<br>$\beta_{1j}=\gamma_{10}$ |
| **AINT** | Second year test score | $Y_{ij}=\beta_{0j} + \beta_{1j}(X_{1j}- \overline{X} ..) +$<br>$\beta_{2j}(X_{2j}-X..) + r_{ij}$ | $\beta_{0j}=\gamma_{00} + \gamma_{01}W_j + u_{0j}$<br>$\beta_{1j}=\gamma_{10}$<br>$\beta_{2i}=\gamma_{20}$ |
| **DLINTSLP** | First stage residuals | $Y_{ij}=\beta_{0j} + \beta_{1j}X_{1j} + r_{ij}$ | $\beta_{0j}=\gamma_{00} + \gamma_{01}W_j + u_{0j}$<br>$\beta_{1j}=\gamma_{10} + u_{1j}$ |
| **UNINTSLP** | Second year test score | $Y_{ij}=\beta_{0j} + \beta_{1j}(X_{1j}- \overline{X} ..) + r_{ij}$ | $\beta_{0j}=\gamma_{00} + u_{0j}$<br>$\beta_{1j}=\gamma_{10} + u_{1j}$ |

| | | | |
|---|---|---|---|
| **AINTSLP** | Second year test score | $Y_{ij}=\beta_{0j} + \beta_{1j}(X_{1j}- \overline{X} ..) + \beta_{2j}(X_{2j}-X..) + r_{ij}$ | $\beta_{0j}=\gamma_{00} + \gamma_{01}W_j + u_{0j}$<br>$\beta_{1j}=\gamma_{10} + u_{1j}$<br>$\beta_{2j}=\gamma_{20} + u_{2j}$ |

$X_{1j}$ = Student first year test score (or residual)
$X_{2j}$ = Student free and reduced price lunch status
$W_j$ = School percent free and reduced price lunch students

**COMPARISONS.**

The following comparisons were used to investigate the research questions of this dissertation.

The first question addressed was question 1: *Are the different statistical approaches equally sensitive to the inclusion or exclusion of student SES? How similar are results of school rankings based on each model without this variable compared to results of the same model with student SES included?* For each model in which free and reduced price lunch status can be included as part of the rankings, the schools were ranked by the outcome of the model separately by grade for mathematics and reading. For campuses with rankings under the same model with and without the contextual variable included, correlations between rankings were calculated. Correlations of the rankings produced by the unadjusted single-level regression model with the adjusted model, and the unadjusted and adjusted multilevel models, were examined in order to evaluate the similarity of results of each model without the contextual variable (student free and reduced price lunch status) compared to the results of the same model with the contextual variable included.

The second question addressed was: *Are the different statistical approaches equally unbiased against schools with a high percentage of low SES students? That is, how do the rankings of schools calculated using the different models including vs. excluding SES correlate with the percent of low SES students in the school?* Results of

the rankings for each model by subject and grade were also correlated with the percent low SES students in each school in order to assess the extent to which rankings are influenced by percent of low SES students in the school.

Next to be addressed was Question 3: *Are the different statistical approaches comparable? How do rankings calculated using each model compare to rankings calculated using each of the other models?* For campuses ranked under every model, results of the rankings under each model were correlated with rankings for other models within test subjects and grades, using Pearson product moment correlation coefficients in order to assess the similarity of rankings based on each model compared to rankings based on each of the other models.

To further investigate the third question, results of rankings using each model were compared to the results of the rankings determined using the percent passing model. Comparisons are made to the results of the percent passing model because that is the simplest model and is the model required by federal *No Child Left Behind* legislation. An average absolute change statistic was created for comparison by calculating the absolute value of the difference in rankings from each of the other models with the rankings from this model.

After comparison of rankings calculated using the different models, one possible impact of differences was explored through the examination of the difference to schools which were ranked in the top 20 percent and bottom 20 percent of schools using the percent passing model. These are the schools that would be impacted by reward to top-performing schools and/or sanctions against the lowest performing schools. Average Absolute Value statistics were examined in more detail for these schools, as well as the percent of schools remaining in the top or bottom 20 percent when ranked by other models, and the characteristics of these schools.

## Simulation Study

Further investigation of the relation between achievement, free and reduced price lunch and model type was conducted using a small simulation study to investigate the other questions. *How do rankings calculated using the various statistical models differ as the relation between student SES and student achievement varies? How do rankings calculated using the various statistical models differ as the relation between school percent low SES students and average school achievement varies? How do rankings calculated using the multilevel models differ from rankings calculated using performance, cohort difference, and single-level regression models as the portion of total variation in student achievement explained by schools varies?*

A simple two-level model with data from 250 schools, each containing 30 individuals per school, was simulated. The data were generated by specifying models, separately, for the pretest and posttest. For the pretest, two explanatory variables were used (you can list and describe them here). For the posttest, these same two explanatory variables were used as well as the pretest scores obtained from the first equation.

The following models were used to generate pretest and posttest scores:

Pretest= $\gamma_{10}(SES)$ + $\gamma_{01}$(FRLPCT) + scherrpre + stuerrpre

Postest= $\gamma_{20}(SES)$ + $\gamma_{01}$(FRLPCT) + pretest +scherrpost + stuerrpost

In these equations, SES represented student socioeconomic status. FRLPCT was the percent of low SES students in the school, and student error on the pretest (stuerrpre) and posttest (stuerrpost) were generated separately from a normal distribution with a mean of zero and standard deviation of one. School-level error (scherrpre and scherrpost) was simulated separately from a normal distribution with a mean of zero and standard

deviation of one, so that the proportion of variation between schools fit the ICC values of .05 or .15, depending on condition.

There were 3 (relation between school percent low SES students and student achievement) x 2 (relation between student SES and student achievement on first year test score) x 2 (relation between student SES and student achievement on second year test score) x 2 (proportion of variance between groups) = 24 conditions. For each condition, 30 student records in each of 250 schools were generated. In the real datasets the proportion of students receiving free and reduced price lunch was approximately 50 percent. Therefore, the proportion of free and reduced price lunch students was generated randomly for schools from a uniform distribution with mean of 0.5 and standard deviation of 0.1. The value generated was then used as the cutoff for determining the lower level variable *SES*, individual student free and reduced price lunch status. The lower level variable *SES* was randomly drawn from a uniform distribution with a mean of 0 and a variance of 1. Because in the real data sets *SES*, representing free and reduced price lunch, is a dichotomous variable, if the randomly drawn value was equal to or lower than the cut point determined by the percent of free and reduced price lunch students in the school, *SES* was assigned the value 1, representing a student receiving free and reduced price lunch, and if the randomly drawn value was higher than the cut point, *SES* was assigned the value 0, representing a student not receiving free and reduced price lunch.

The coefficient for the intercept was specified as 0.0 in this simulation. The residual variance of the intercept at the lowest level was set at 1.0. The residual variance of the intercept at the second level followed from the specification of the conditional ICC (.05 and .15) and the residual variance of the intercept at the first level, given the formula $ICC = \tau_{00} / \tau_{00} + \sigma 2$. The values of the conditional ICC were chosen to represent actual

proportions of 0.10 to 0.20 found in educational research (Kreft & Yoon, 1994) and also representative of the Texas data used in the previous analysis (0.02 to 0.05). School level error $\tau_{00}$ was thus generated from a normal distribution for two conditions with a mean of 0 and variance 0.053 and 0.176. It was expected that as the portion of total variation in student achievement explained by schools increased that multilevel models would produce rankings more different from the performance, cohort difference, and single-level regression models.

The relation between school percent free and reduced price lunch students and student test score in the real data sets varies between 0.01 and -0.10. The first value, being positive, seems not only unrealistic, but opposite to the investigation of the research questions studied in this dissertation. If, however, schools with high percentages of low SES students are being given more funding as part of a more intense focus on student performance, then it would make sense for such schools to have a more positive relation between percent low SES students and an individual student's test score. It was therefore decided to include such a condition for purposes of investigation. The values of first and second year test score and SES were standardized and three conditions were modeled to simulate the relation between test score and school-level SES, using values of 0, 0.01 and -0.10. It was expected that a weaker relation between school percent low SES students and average student achievement would produce more similar rankings across models.

The standardized relation between student free and reduced price lunch status and student test score in the real data sets varies between -0.05 and -0.12. Therefore, the values of first and second year test score and SES were standardized and two conditions were modeled to simulate the values of 0 and -0.10. It was expected that a weaker

relation between student SES and student achievement would produce more similar rankings across models.

Student pretest score was generated using disadvantaged status, school percent free and reduced price lunch students, school error and student error. Student error was generated randomly from a normal distribution with a mean of 0 and standard deviation of 1. Student posttest score was then generated using the same elements as pretest score with the addition of pretest score. Student error was generated separately for pretest and posttest. Percent of students passing the posttest was then used as the outcome variable for the percent passing model. Average posttest score minus average pretest score for each campus was used as the outcome variable for the cohort difference model. The unadjusted school-level regression model used school average pretest score to predict school average posttest score, and the adjusted school-level regression model used school average pretest score and percent of free and reduced price lunch students to predict school average posttest score.

The first multilevel model to be estimated, the two-stage two-level model similar to the Dallas model, used student free and reduced price lunch status to predict pretest score and posttest score in two single-level linear equations, and then residuals from the first model and school percent free and reduced price lunch students to predict residuals from the second equation. The second multilevel model was a one-stage two-level model which predicted posttest score using centered pretest score. The third multilevel model was also a one-stage two-level model, but used centered pretest score as well as free and reduced price lunch status as well as school percent free and reduced price lunch students to predict posttest score.

Table 6: Summary of Simulation Design

| *pct* a | *pre* b | *post* c | ICC d |
|---|---|---|---|
| 0.00 | 0.00 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| | -0.10 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| 0.01 | 0.00 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| | -0.10 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| -0.10 | 0.00 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| | -0.10 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| 0.00 | 0.00 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| | -0.10 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| 0.01 | 0.00 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| | -0.10 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| -0.10 | 0.00 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| | -0.10 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |

a *pct* =coefficient for relation between percent low SES students and $Y_{ij}$
b *pre* =coefficient for relation between student first year test score and *SES*
c *post* =coefficient for relation between student second year test score and SES
d ICC = intraclass correlation

Data were generated and models estimated using the SAS programming language (SAS). In order to investigate the amount of variability in rankings based on generation of data under each condition, ten replications were conducted for each of the 24

conditions, producing 240 simulated data sets. Data sets were then used to rank schools using each of the models used earlier with empirical data as follows.

The 250 schools used in this simulation were ranked separately based on the requirements of each of the statistical models. A raw score of -0.50, at which approximately 70 percent of students were designated as having passed the assessment under conditions of no relation between SES and test score, was set as the passing standard. For the percent passing model, percent of students passing was calculated. Schools were then ranked using the cohort difference model. The single-level school regression model ranked schools with and without SES included. The hierarchical linear model was simulated using the random intercept model for purposes of this study. The equation for the random intercept model is described for the AINT and UNINT models in Table 3.

Rankings based on each of the statistical models were compared across the simulated conditions. The rankings produced by each of the models for each of the conditions were correlated with the percent of free and reduced price lunch students in the school. Results of the rankings under each model were correlated with rankings for other models and results of rankings under each model were compared to the results of the rankings determined using the percent passing model.

The median rank for each school across each condition was compared to each of the other models. Range and number of places changed in the rankings was reported.

Results of the analysis using real data will be explained next, followed by results of the simulation study.

# CHAPTER 4: RESULTS

First to be discussed will be the analysis of real student data, followed by analysis of the simulated data.

## Real Data

The data for this study come from public school districts in the state of Texas. Schools with students enrolled in the studied grades (3, 4 or 5) and taking the mathematics or reading Texas Assessment of Academic Skills (TAAS) in the 2000-2001 or the 2001-2002 school year were included in this study. A description of the data follows.

### *DESCRIPTIVE STATISTICS*

Table 7 provides summary information for the grade levels and number of test takers included in the study.

Table 7: Descriptive Statistics of Grade Levels Included in the Study

| Year | Grade | Students | TAAS Takers | % TAAS Takers |
|---|---|---|---|---|
| 2000-2001 | 3 | 316,535 | 275,807 | 87.13 |
| 2000-2001 | 4 | 313,731 | 281,996 | 89.88 |
| 2001-2002 | 4 | 318,674 | 284,690 | 89.34 |
| 2001-2002 | 5 | 317,137 | 287,875 | 90.77 |

Only students who were tested on either the mathematics or reading TAAS test were included in the study. Table 8 describes the students included in the study.

Table 8: Descriptive Statistics of Students Included in the Study

| Year | Grade | F&R | Black | Asian | Hispanic | Native | White |
|---|---|---|---|---|---|---|---|

| | | Lunch | | | | Am. | |
|---|---|---|---|---|---|---|---|
| 2000-2001 | 3 | 48.92% | 15.27% | 2.90% | 37.67% | 0.35% | 43.81% |
| 2000-2001 | 4 | 49.03% | 14.81% | 2.83% | 38.34% | 0.33% | 43.69% |
| 2001-2002 | 4 | 49.17% | 14.64% | 2.98% | 39.69% | 0.34% | 42.35% |
| 2001-2002 | 5 | 49.19% | 14.33% | 2.85% | 39.98% | 0.30% | 42.54% |

F&R Lunch = Percent Free or Reduced Price Lunch
Black = Percent African American
Asian = Percent Asian/Pacific Islander
Hispanic = Percent Hispanic
Native Am. = Percent Native American
White = Percent White

The population of students in grades 3, 4 and 5 used in the study is representative of all students in the state. The overall state percent of students receiving free or reduced price lunch was 49.3 in 2000-2001 and 50.5 in 2001-2002. The overall state percent of African American students was 14.4 in both school years. The overall state percent Asian/Pacific Islander students was 2.7 in 200-2001 and 2.8 in 2001-2002. Hispanic students in those years represented 40.6 (2000-2001) and 41.7 (2001-2002) of all students. White students were 42.0 of the total state population in 2000-2001 and 40.9 in 2001-2002.

Only schools with students tested on either the mathematics or reading TAAS test were included in the study. Table 9 describes the schools included in the study.

Table 9: Descriptive Statistics of Schools Included in the Study

| Year | Grade | Schools | Enrollment | F&R Lunch | Male |
|---|---|---|---|---|---|
| 2000-2001 | 3 | 3625 | 76.08 | 50.05% | 50.33% |
| 2000-2001 | 4 | 3613 | 78.05 | 50.35% | 50.08% |
| 2001-2002 | 4 | 3685 | 77.26 | 49.80% | 50.10% |
| 2001-2002 | 5 | 3451 | 83.42 | 50.26% | 50.06% |

Enrollment = Average school enrollment in grade
F&R Lunch = Average School Percent Free or Reduced Price Lunch
Male = Average school percent males students

Data were analyzed using each of the statistical models described in Chapter 3. The first model used was the percent passing model. This model ranks schools by grade and subject based on percent of students in each grade in the school who passed each subject test (reading or mathematics). Schools which had one or more students taking a subject test were ranked on the percent of students passing that test. In the event of tied rankings, all schools were assigned the same rank. The school earning the rank below the tie was assigned the rank common to the tied schools plus the number of schools with a tied rank. In other words, if two schools were ranked number 1, both of those schools were assigned a rank of 1, while the third school was ranked number 3.

For purposes of this study, the percent passing model was first used to assess schools on 4$^{th}$ and 5$^{th}$ grade mathematics and reading results in the second year of testing. Of the 3,685 schools with fourth grade students in the 2001-2002 school year, 3,599 schools tested those fourth grade students on the mathematics and reading TAAS. Eighty-six schools were dropped from the analysis for fourth grade because they had no students tested on TAAS. In the 2001-2002 school year, 3,451 schools had fifth grade students, with 3,362 of those schools testing students on the mathematics and reading TAAS. Eighty-nine schools were dropped from the analysis for fifth grade because they had no students tested on TAAS.

Schools were assigned four rankings for this model; percent of fourth grade students passing the mathematics and reading assessments, and percent of fifth grade students passing the mathematics and reading assessments. Descriptive statistics for the school status model are presented in Table 10.

Table 10: Descriptive Statistics for the School Status Model

| Grade | Subject | Schools | Passing | F&R Lunch |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| 4 | Math | 3599 | 88% | 50% |
| 4 | Reading | 3599 | 86% | 50% |
| 5 | Math | 3362 | 90% | 51% |
| 5 | Reading | 3362 | 86% | 51% |

Schools = Number of schools included in rankings
Passing = Average proportion of students passing
F&R Lunch = Percent Free or Reduced Price Lunch

The next statistical model used to assign school rankings was the cohort difference model. For this model schools were ranked based on the difference in school average subject test score for a cohort of students across two years. Average score on the Texas Learning Index (TLI) for 3rd grade students in the first year of testing was subtracted from average TLI score for 4th grade students in the second year of testing, and average TLI score for 4th grade students in the first year of testing was subtracted from average TLI score for 5th grade students in the second year of testing. This process was completed separately for reading and math.

Schools were only ranked using the cohort difference model if they had student scores for a grade/subject combination in both years of testing. There were 3,625 schools with third grade students in 2000-2001, and 3,685 schools with fourth grade students in 2001-2002. 3,362 of those schools had students taking the third grade TAAS test in the first year of testing and students taking the fourth grade TAAS test in the second year of testing. 3,613 schools had fourth grade students in 2000-2001, and 3,451 schools had fifth grade students in 2001-2002. 3,061 schools had students taking the fourth grade TAAS test in the first year of testing and students taking the fifth grade TAAS test in the second year of testing. Students were not matched across years of testing, so that the

same students were not necessarily included in a cohort across years. Descriptive statistics for the cohort difference model are presented in Table 11.

Table 11: Descriptive Statistics for the Cohort Difference Model

| Grades | Subject | Schools | Difference Mean (SD) | F&R Lunch |
|---|---|---|---|---|
| 3-4 | Math | 3362 | 3.31 (8.12) | 50% |
| 3-4 | Reading | 3362 | 4.77 (8.43) | 50% |
| 4-5 | Math | 3061 | 4.18 (7.61) | 51% |
| 4-5 | Reading | 3061 | 2.54 (8.28) | 51% |

Schools = Number of schools included in rankings
Difference Mean (SD) = Mean (Standard Devation) of Difference Score
F&R Lunch = Percent Free or Reduced Price Lunch

Next, schools were ranked using a single-level regression model conducted at the school level. There were 3,625 schools with third grade students in 2000-2001, and 3,685 schools with fourth grade students in 2001-2002. 3,362 of those schools had students taking the third grade TAAS test in the first year of testing and students taking the fourth grade TAAS test in the second year of testing. There were 3,613 schools with fourth grade students in 2000-2001, and 3,451 schools had fifth grade students in 2001-2002. There were 3,061 schools with students taking the fourth grade TAAS test in the first year of testing and students taking the fifth grade TAAS test in the second year of testing. Students were not matched across years of testing, so that the same students were not necessarily included in a cohort across years.

For the unadjusted school regression model, school average TLI scale score for each subject in the second year of testing was predicted by school average TLI scale score in the same subject in the first year of testing. For the adjusted school regression model, school average TLI scale score for each subject in the second year of testing was

predicted by school average TLI scale score in the same subject in the first year of testing, along with school percent of students receiving free and reduced price lunch. Schools were then ranked for each grade and subject by school-level residual. Descriptive statistics for both models are presented in Table 12. The unadjusted school-level regression model uses school average first year test score to predict second year test score. The adjusted school-level regression model uses school average first year test score and school average percent free and reduced price lunch students to predict second year test score.

Table 12: Descriptive Statistics for School Regression Model

| Model | Grade | Subject | Schools | TLI - Mean (SD) | | F&R Lunch |
|---|---|---|---|---|---|---|
| | | | | 1st Year | 2nd Year | |
| USLR | 3-4 | Math | 3362 | 75.76 (10.28) | 79.07 (7.48) | 50% |
| USLR | 3-4 | Reading | 3362 | 77.82 (10.56) | 82.58 (8.32) | 50% |
| USLR | 4-5 | Math | 3061 | 77.59 (8.83) | 81.77 (8.05) | 51% |
| USLR | 4-5 | Reading | 3061 | 81.26 (9.77) | 83.80 (9.44) | 51% |
| ASLR | 3-4 | Math | 3362 | 75.76 (10.28) | 79.07 (7.48) | 50% |
| ASLR | 3-4 | Reading | 3362 | 77.82 (10.56) | 82.58 (8.32) | 50% |
| ASLR | 4-5 | Math | 3061 | 77.59 (8.83) | 81.77 (8.05) | 51% |
| ASLR | 4-5 | Reading | 3061 | 81.26 (9.77) | 83.80 (9.44) | 51% |

Schools = Number of schools included in rankings
TLI - Mean (SD) = Mean (Standard Deviation) of Texas Learning Index
F&R Lunch = Percent Free or Reduced Price Lunch

The last set of models used to rank schools was a series of multilevel models. The first and fourth were two-stage models similar to that used in the Dallas, Texas, school district accountability system. The other four multilevel models were the one-stage models that Thum and Bryk (1997) suggested would be just as effective as, but more efficient than, the Dallas model.

There were five fewer schools in the grade 3-4 comparison for this dataset than in the cohort difference and school-level regression models, and one fewer school in the grade 4-5 comparison than in the cohort difference and school-level regression models. After thorough examination of all datasets, this discrepancy appears to result from the difference in level of analysis from school level in the cohort difference and school level regression models to individual level for the multilevel models. The multilevel models required merging at the student level, and the order in which requirements for inclusion in the data set were made could very well be the cause of the discrepancies. As shown in tables 11 12 and 13, the differences do not result in differences in the averages of the predictor or outcome variables.

Table 13: Descriptive Statistics for Multilevel Models

| Model | Grade | Subject | Schools | TLI - Mean (SD) | | F&R Lunch |
|---|---|---|---|---|---|---|
| | | | | 1st Year | 2nd Year | |
| DLINT | 3-4 | Math | 3357 | 75.85 (9.95) | 79.14 (7.15) | 50% |
| | 3-4 | Reading | 3357 | 77.92 (10.22) | 82.67 (8.02) | 50% |
| | 4-5 | Math | 3060 | 77.62 (8.72) | 81.78 (8.03) | 51% |
| | 4-5 | Reading | 3060 | 81.29 (9.66) | 83.83 (9.34) | 51% |
| UNINT | 3-4 | Math | 3357 | 75.85 (9.95) | 79.14 (7.15) | 50% |
| | 3-4 | Reading | 3357 | 77.92 (10.22) | 82.67 (8.02) | 50% |
| | 4-5 | Math | 3060 | 77.62 (8.72) | 81.78 (8.03) | 51% |
| | 4-5 | Reading | 3060 | 81.29 (9.66) | 83.83 (9.34) | 51% |
| AINT | 3-4 | Math | 3357 | 75.85 (9.95) | 79.14 (7.15) | 50% |
| | 3-4 | Reading | 3357 | 77.92 (10.22) | 82.67 (8.02) | 50% |
| | 4-5 | Math | 3060 | 77.62 (8.72) | 81.78 (8.03) | 51% |
| | 4-5 | Reading | 3060 | 81.29 (9.66) | 83.83 (9.34) | 51% |
| DLINTSLP | 3-4 | Math | 3357 | 75.85 (9.95) | 79.14 (7.15) | 50% |
| | 3-4 | Reading | 3357 | 77.92 (10.22) | 82.67 (8.02) | 50% |
| | 4-5 | Math | 3060 | 77.62 (8.72) | 81.78 (8.03) | 51% |
| | 4-5 | Reading | 3060 | 81.29 (9.66) | 83.83 (9.34) | 51% |
| UNINTSLP | 3-4 | Math | 3357 | 75.85 (9.95) | 79.14 (7.15) | 50% |
| | 3-4 | Reading | 3357 | 77.92 (10.22) | 82.67 (8.02) | 50% |
| | 4-5 | Math | 3060 | 77.62 (8.72) | 81.78 (8.03) | 51% |
| | 4-5 | Reading | 3060 | 81.29 (9.66) | 83.83 (9.34) | 51% |
| AINTSLP | 3-4 | Math | 3357 | 75.85 (9.95) | 79.14 (7.15) | 50% |
| | 3-4 | Reading | 3357 | 77.92 (10.22) | 82.67 (8.02) | 50% |
| | 4-5 | Math | 3060 | 77.62 (8.72) | 81.78 (8.03) | 51% |
| | 4-5 | Reading | 3060 | 81.29 (9.66) | 83.83 (9.34) | 51% |

Schools = Number of schools included in rankings
TLI - Mean (SD) = Mean (Standard Deviation) of Texas Learning Index
F&R Lunch = Percent Free or Reduced Price Lunch

Datasets containing school rankings under each of the models were then combined and only campuses ranked under all models were included in the final combined dataset. Rankings for each model were recomputed using only campuses included in the final combined dataset. Analysis of the real dataset will be described next.

***RESEARCH QUESTION ONE***

The first analysis compared the results of adjusted models with free and reduced price lunch as a component of the model with the same unadjusted models which did not include free and reduced price lunch as a component of the model. School rankings produced by the unadjusted model were compared with rankings produced by the adjusted model. As shown in Table 12, the correlation of rankings between the unadjusted and adjusted single-level regression models for mathematics in grade 4 and grade 5 were .98 and .99. Correlation of rankings based on results of the reading assessment were also very high, at .92 for grade 4 and .96 for grade 5. There appeared to be very little difference using a school-level regression model with or without including school percent free and reduced price lunch students. Table 14 presents the results of this comparison.

Table 14: Comparison of Unadjusted (USLR) and Adjusted (ASLR) School Level Regressions

| Grade | Subject | Correlation of rankings |
|---|---|---|
| 4 | Math | 0.98 |
| 4 | Reading | 0.92 |
| 5 | Math | 0.99 |
| 5 | Reading | 0.96 |

Correlation of rankings = Correlation of school rankings from models USLR with ASLR

Models UNINT and AINT are one-stage two-level models which predict student second year test score using first year test score. In Table 13, the rankings resulting from UNINT are compared to those from Models DLINT and AINT, which are similar random intercept models but which include student free and reduced price lunch status as a predictor as well as school average free and reduced price lunch. Likewise, Model

UNINTSLP, which is the same model as Model UNINT but with random intercept and slopes, is compared to Models DLINTSLP and AINTSLP in Table 14.

The correlation of results from the unadjusted multilevel model with the Dallas and adjusted multilevel models is consistent across models for each grade and subject. Model DLINT and Model AINT are essentially the same model, with posttest score being predicted by pretest score, student free and reduced price lunch status and school percent free and reduced price lunch students. The difference between Models DLINT and AINT is that Model DLINT is the Dallas-type two-stage model, while Model AINT is a one-stage model. It is not unexpected that the results of the unadjusted model UNINT would be correlated at the same level with the results of the two-stage Dallas model and the adjusted one-stage model. For instance, the results from the unadjusted model correlate 0.95 with the adjusted two-stage (Dallas) and one-stage adjusted models in Grade 4 mathematics. Correlations are above 0.90 for all but grade 4 reading, which is 0.88.

Results of the models with random intercepts and slopes were then compared for correlation across models as shown in Table 14. The results of Model UNINTSLP were correlated with the results from Models DLINTSLP and AINTSLP. Again, the results from Model UNINTSLP were correlated at the same level across models with DLINTSLP and AINTSLP, for each grade/subject combination. Correlations range from 0.87 to 0.97, with the lowest correlation occurring for the grade 4 reading results.

Recall that the correlation between results for the unadjusted and adjusted school-level regression models were the lowest (at 0.92) for grade 4 reading also. It appears that the inclusion of SES in accountability models for this grade/subject combination creates a bigger difference than at the other grade/subject combinations. This finding will be further investigated in the next section in which rankings based on each of the models is correlated with school percent free and reduced price lunch students.

Comparison of the random intercept multilevel models showed that results of rankings resulting from use of the unadjusted one-stage model correlated at the same level with results of rankings resulting from use of the two-stage Dallas model and the one-stage adjusted model. These results are summarized in Table 15.

Table 15: Comparison of Unadjusted Multilevel Model (UNINT) with other Multilevel Models having Random Intercept (DLINT & AINT)

| Grade | Subject | DLINT | AINT |
|---|---|---|---|
| 4 | Math | 0.95 | 0.95 |
| 4 | Reading | 0.88 | 0.88 |
| 5 | Math | 0.97 | 0.97 |
| 5 | Reading | 0.93 | 0.93 |

DLINT = Correlation of rankings produced by models UNINT and DLINT
AINT = Correlation of rankings produced by models UNINT and AINT

Further comparison of the random intercept multilevel models showed that the random intercept and random slope model results from use of the unadjusted one-stage model correlated slightly more highly with results of the two stage model that with results of the adjusted one-stage model.

Table 16: Comparison of Unadjusted Multilevel Model (UNINTSLP ) with other Multilevel Models having Random Intercept and Slope (DLINTSLP & AINTSLP)

| Grade | Subject | DLINTSLP | AINTSLP |
|---|---|---|---|
| 4 | Math | 0.94 | 0.95 |
| 4 | Reading | 0.87 | 0.88 |
| 5 | Math | 0.94 | 0.97 |
| 5 | Reading | 0.89 | 0.92 |

DLINTSLP = Correlation of rankings produced by models UNINTSLP and DLINTSLP
AINTSLP = Correlation of rankings produced by models UNINTSLP and AINTSLP

Results of unadjusted and adjusted regression models were generally highly correlated, with some exceptions. The school-level regression models and the multilevel models, both those with random intercepts and those with random intercepts and slopes, all showed relatively high correlations between rankings produced by models with and without free and reduced price lunch status included as a predictor. The lowest correlation between these models was 0.87. Whether this difference is important in school accountability systems depends on the extent to which the difference between rankings is biased against schools with different types of students.

### *RESEARCH QUESTION TWO*

A second question addressed in this study was the extent to which each model was biased against schools with a high percentage of low SES students. If the results of school rankings with a particular model are positively correlated with the percent of free and reduced price lunch students in the school, the model may be considered biased against schools with high percentages of low SES students. If the results of the rankings are negatively correlated with the percent of free and reduced price lunch students in the school, the model can be considered to be biased against schools with low percentages of low SES students.

As expected, as shown in Table 15, in general models which do not include SES of students show higher positive correlations of school ranking with the percent of free and reduced price lunch students in the school. Results from the percent passing model has the highest correlations, ranging from 0.24 to 0.41. Of this group of models which do not include SES, the unadjusted single-level regression model has the lowest correlations of school ranking and percent of free and reduced price lunch students, ranging from 0.06 to 0.33.

Models which do include SES have results that are hardly correlated with percent of free and reduced price lunch students in a school. These results are generally around 0.00, ranging from -0.07 for the adjusted single-level regression model to 0.03 for the adjusted multilevel model.

The cohort difference model is a strong exception to this pattern. This model produces school rankings which are somewhat negatively correlated with the school percent of free and reduced price lunch students. These rankings range from -0.01 to -0.18. Rankings using the cohort difference model are thus shown to be somewhat biased against schools with low percentages of free and reduced price lunch students, especially in $4^{th}$ and $5^{th}$ grade mathematics.

As mentioned in the previous section, there was a lower correlation between the same models with and without free and reduced price lunch included in the model for rankings based on grade 4 reading results than there was for other grade/subject combinations. It should be noted from Table 15 that the grade 4 reading results reinforce this finding. In models which do not include free and reduced price lunch status, the results for this grade/subject combination are more biased than results for other grades. In the other model that does not include SES, the cohort difference model, the results for this grade/subject combination are less biased against schools with low percentages of free and reduced price lunch students. It again appears that the inclusion of SES in accountability models for this grade/subject combination creates a bigger difference than at the other grade/subject combinations.

Examination of the relation between test score and SES in the original data sets reveals that the two are correlated on average .04 for mathematics, but .09 for reading in fifth grade and .10 for reading in fourth grade. There is apparently more of a relation

between SES and achievement, as measured by test score, for reading than for mathematics.

Table 17 presents the correlation of school ranking calculated using each model with the percent of students receiving free or reduced price lunch.

Table 17: Correlation of Ranking Results from Each Model with Percent of Students Receiving Free or Reduced Price Lunch

| Grade/ Subject | Model | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PCTPASS | COHDIFF | USLR | ASLR | DLINT | UNINT | AINT | DLINTSLP | UNINTSLP | AINTSLP |
| 4 Math | .34 | -.18 | .15 | -.04 | -.01 | .19 | -.01 | .00 | .16 | .01 |
| 4 Reading | .41 | -.01 | .33 | -.01 | .00 | .35 | .00 | .01 | .32 | .01 |
| 5 Math | .24 | -.16 | .06 | -.07 | -.01 | .09 | -.01 | .01 | .09 | .02 |
| 5 Reading | .37 | -.06 | .21 | -.04 | .00 | .24 | .00 | .02 | .26 | .03 |

***RESEARCH QUESTION THREE***

Having investigated the differences occurring as a result of including SES as a predictor in an accountability model and the correlation between rankings and percent low SES students in a school, the next question addressed was how the rankings calculated using each of the different statistical approaches compared to the rankings calculated with each of the other models.

In order to compare all models on the same schools, only schools which had a ranking using each of the models were included in the dataset. Rankings under each of the models were then re-calculated based on the complete list of schools.

As shown in Tables 18-21, rankings based on the percent passing model had the lowest correlation with the cohort difference and school-level regression models.

Similarity of rankings between the percent passing and cohort difference models were the lowest overall, ranging between 0.19 (grade 4 mathematics) and 0.37 (grade 5 mathematics), with an average of 0.31. Rankings based on the cohort difference model were most highly correlated with rankings based on the school-level regression models, at 0.79 and 0.80. Rankings based on the cohort difference model were less highly correlated with results of the multilevel models than the single-level regression models, ranging from 0.33 to 0.47. The rankings calculated using the cohort difference model produced slightly higher correlations with the random intercept multilevel models than the random intercept/random slope multilevel models.

Average rankings among the multilevel models were quite highly correlated, ranging from 0.74 (the Dallas models with the adjusted model) to 1.0 (the Dallas model with the adjusted random slope model). Tables 18-21 present correlations calculated between each model for each grade/subject combination.

Table 22 presents the average of the four grade and subject combination correlations between models. An unexpected result was the relatively low correlation between the random intercept multilevel models (DLINT, UNINT and AINT) and the corresponding random intercept and random slope multilevel models (DLINTSLP, UNINTSLP and AINTSLP). These correlations were 0.80 for the unadjusted models as well as the adjusted models, and 0.83 for the one-stage Dallas models. Because the only difference between the corresponding models is the random slope, it was anticipated that the results would be more highly correlated. There was unanticipated variation in the slopes that was not modeled in the random intercept models.

Table 18: Correlations of Ranking across Models, Grade 4 Mathematics

| | PCTPASS | COHDIFF | USLR | ASLR | DLINT | UNINT | AINT | DLINTSLP | UNINTSLP | AINTSLP |
|---|---|---|---|---|---|---|---|---|---|---|
| PCTPASS | — | .19 | .71 | .67 | .50 | .57 | .50 | .43 | .47 | .42 |
| COHDIFF | | — | .73 | .75 | .48 | .44 | .47 | .42 | .37 | .40 |
| USLR | | | — | .98 | .67 | .70 | .67 | .56 | .57 | .55 |
| ASLR | | | | — | .68 | .67 | .68 | .57 | .54 | .55 |
| DLINT | | | | | — | .95 | 1.0 | .85 | .78 | .82 |
| UNINT | | | | | | — | .95 | .81 | .83 | .79 |
| AINT | | | | | | | — | .85 | .78 | .82 |
| DLINTSLP | | | | | | | | — | .94 | .97 |
| UNINTSLP | | | | | | | | | — | .95 |
| AINTSLP | | | | | | | | | | — |

Table 19: Correlations of Ranking across Models, Grade 4 Reading

| | PCTPASS | COHDIFF | USLR | ASLR | DLINT | UNINT | AINT | DLINTSLP | UNINTSLP | AINTSLP |
|---|---|---|---|---|---|---|---|---|---|---|
| PCTPASS | — | .31 | .80 | .71 | .47 | .62 | .47 | .41 | .53 | .40 |
| COHDIFF | | — | .72 | .74 | .46 | .42 | .46 | .39 | .34 | .38 |
| USLR | | | — | .92 | .58 | .70 | .58 | .50 | .59 | .49 |
| ASLR | | | | — | .63 | .61 | .63 | .54 | .50 | .53 |
| DLINT | | | | | — | .88 | 1.0 | .85 | .72 | .82 |
| UNINT | | | | | | — | .88 | .75 | .83 | .73 |
| AINT | | | | | | | — | .85 | .72 | .82 |
| DLINTSLP | | | | | | | | — | .87 | .97 |
| UNINTSLP | | | | | | | | | — | .88 |
| AINTSLP | | | | | | | | | | — |

Table 20: Correlations of Ranking across Models, Grade 5 Mathematics

| | PCTPASS | COHDIFF | USLR | ASLR | DLINT | UNINT | AINT | DLINTSLP | UNINTSLP | AINTSLP |
|---|---|---|---|---|---|---|---|---|---|---|
| PCTPASS | — | .37 | .72 | .69 | .49 | .52 | .49 | .39 | .39 | .38 |
| COHDIFF | | — | .85 | .86 | .44 | .43 | .44 | .32 | .29 | .31 |
| USLR | | | — | .99 | .60 | .61 | .60 | .45 | .43 | .43 |
| ASLR | | | | — | .60 | .60 | .60 | .44 | .42 | .43 |
| DLINT | | | | | — | .97 | 1.0 | .80 | .75 | .78 |
| UNINT | | | | | | — | .97 | .78 | .77 | .76 |
| AINT | | | | | | | — | .80 | .75 | .78 |
| DLINTSLP | | | | | | | | — | .94 | .98 |
| UNINTSLP | | | | | | | | | — | .97 |
| AINTSLP | | | | | | | | | | — |

Table 21: Correlations of Ranking across Models, Grade 5 Reading

| | PCTPASS | COHDIFF | USLR | ASLR | DLINT | UNINT | AINT | DLINTSLP | UNINTSLP | AINTSLP |
|---|---|---|---|---|---|---|---|---|---|---|
| PCTPASS | — | .36 | .72 | .66 | .44 | .53 | .44 | .36 | .44 | .35 |
| COHDIFF | | — | .85 | .86 | .48 | .45 | .48 | .37 | .30 | .35 |
| USLR | | | — | .96 | .56 | .61 | .56 | .44 | .45 | .42 |
| ASLR | | | | — | .58 | .56 | .58 | .45 | .39 | .43 |
| DLINT | | | | | — | .93 | 1.0 | .81 | .70 | .78 |
| UNINT | | | | | | — | .93 | .76 | .77 | .74 |
| AINT | | | | | | | — | .81 | .70 | .78 |
| DLINTSLP | | | | | | | | — | .89 | .97 |
| UNINTSLP | | | | | | | | | — | .92 |
| AINTSLP | | | | | | | | | | — |

Table 22: Average Correlation of School Rankings across Grades and Subjects

| | PCTPASS | COHDIFF | USLR | ASLR | DLINT | UNINT | AINT | DLINTSLP | UNINTSLP | AINTSLP |
|---|---|---|---|---|---|---|---|---|---|---|
| PCTPASS | — | .31 | .74 | .68 | .48 | .56 | .48 | .40 | .46 | .39 |
| COHDIFF | | — | .79 | .80 | .47 | .44 | .46 | .38 | .33 | .36 |
| USLR | | | — | .96 | .60 | .66 | .60 | .49 | .51 | .47 |
| ASLR | | | | — | .62 | .61 | .62 | .50 | .46 | .49 |
| DLINT | | | | | — | .93 | 1.0 | .83 | .74 | .80 |
| UNINT | | | | | | — | .93 | .78 | .80 | .76 |
| AINT | | | | | | | — | .83 | .74 | .80 |
| DLINTSLP | | | | | | | | — | .91 | .97 |
| UNINTSLP | | | | | | | | | — | .93 |
| AINTSLP | | | | | | | | | | — |

In order to further investigate the comparability of rankings calculated using each of the models, an Average Absolute Value (AAV) statistic was calculated to summarize the difference between rankings calculated using each of the models compared to the percent passing model. The absolute value of the difference between rankings between the percent passing model and each other model was calculated for each school and averaged for the entire data set.

As shown in Table 23, the total number of schools ranked using all models was 3,357 for grade 4 and 3,060 for grade 5. The average absolute value of the difference for a school ranked using the school status model and then the cohort difference model ranged from a low of 780 for grade 5 reading to a high of 989 for grade 4 mathematics. The minimum change between rankings was 0 for all grade/subject combinations, and the maximum difference between rankings under the school status model and the cohort

difference model was as high as 3,326 for grade 4 mathematics. Table 21 presents the AAV for the comparison of the results of rankings based on these two models.

Table 23: Average Absolute Value (AAV) of Difference Between Percent Passing (PCTPASS) Model and Cohort Difference (COHDIFF) Model

| Grade | Subject | Schools | AAV | Min | Max |
|---|---|---|---|---|---|
| 4 | Math | 3357 | 988.7 | 0 | 3326 |
| 4 | Reading | 3357 | 906.2 | 0 | 3316 |
| 5 | Math | 3060 | 783.3 | 0 | 3022 |
| 5 | Reading | 3060 | 780.2 | 0 | 3022 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

The difference between rankings using the percent passing model compared to the single-level regression models was less extreme than the difference between the percent passing and cohort difference models. These differences are presented in Table 24. The AAV between the percent passing and the unadjusted single-level regression model ranged from a low of 444 for grade 4 reading to a high of 550 for grade 4 mathematics. The AAV between the percent passing and the adjusted single-level regression model ranged from a low of 519 for grade 5 mathematics to a high of 600 for grade 4 mathematics. There was more variability in the rankings between the percent passing and adjusted regression model than there was between the percent passing and the unadjusted regression model. Table 22 presents the AAV for the comparison of the percent passing model results with the results calculated using the single-level regression models. The minimum change between rankings for the models was 0, with a maximum of 3,309 for the difference between rankings using the percent passing and unadjusted models for grade 4 mathematics.

Table 24: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Single-Level Regression Models (UNSLR & ASLR)

| Grade | Subject | Schools | PCTPASS - UNSLR | | | PCTPASS – ASLR | | |
|---|---|---|---|---|---|---|---|---|
| | | | AAV | Min | Max | AAV | Min | Max |
| 4 | Math | 3357 | 550.6 | 0 | 3309 | 600.6 | 0 | 3306 |
| 4 | Reading | 3357 | 444.5 | 0 | 3304 | 558.4 | 0 | 3302 |
| 5 | Math | 3060 | 493.2 | 0 | 3021 | 519.9 | 0 | 3021 |
| 5 | Reading | 3060 | 479.7 | 0 | 3020 | 551.5 | 0 | 3018 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

The AAV between rankings using the percent passing model and the random intercept multilevel models ranges from 650 to 793. Both extremes fall within the grade 4 reading ratings. The maximum number of rankings difference between the models was 3,140 between the percent passing model and the Dallas model for 4$^{th}$ grade reading. Table 25 presents the AAV for the comparison of results calculated using the percent passing model with the results calculated using the random intercept multilevel models.

Table 25: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept Multilevel Models (DLINT, UNINT & AINT)

| Grade/ Subject | Schools | PCTPASS-DLINT | | | PCTPASS-UNINT | | | PCTPASS-AINT | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AAV | Min | Max | AAV | Min | Max | AAV | Min | Max |
| 4 Math | 3357 | 776.3 | 0 | 3078 | 708.1 | 0 | 2950 | 775.9 | 1.0 | 3078 |
| 4 Reading | 3357 | 793.9 | 0 | 3140 | 650.5 | 0 | 3297 | 792.4 | 0 | 3139 |
| 5 Math | 3060 | 715.5 | 1.0 | 2972 | 687.3 | 1.0 | 2967 | 715.6 | 0 | 2973 |
| 5 Reading | 3060 | 752.7 | 0 | 2960 | 669.2 | 0 | 2945 | 752.3 | 0 | 2960 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Table 26 shows that the AAV between rankings using the percent passing model and the random intercept/random slope models are higher than those for the percent passing model and the random intercept models. The AAV for these models range from 742 to 856. Just as with the random intercept models, both extremes fall within the grade 4 reading ratings. The maximum number of rankings difference between the models was 3,147 between the percent passing model and the unadjusted random intercept/random slope model for 4$^{th}$ grade mathematics. Table 24 presents the AAV for the comparison of percent passing model results with the results calculated using the random intercept/random slope multilevel models.

Table 26: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept/Random Slope Multilevel Models (DLINTSLP, UNINTSLP & AINTSLP)

| Grade/ Subject | Schools | PCTPASS-DLINTSLP | | | PCTPASS-UNINTSLP | | | PCTPASS-AINTSLP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AAV | Min | Max | AAV | Min | Max | AAV | Min | Max |
| 4 Math | 3357 | 844.1 | 0 | 3169 | 804.4 | 0 | 3147 | 848.8 | 0 | 3147 |
| 4 Reading | 3357 | 844.6 | 0 | 3118 | 742.8 | 0 | 3046 | 856.9 | 1.0 | 3129 |
| 5 Math | 3060 | 804.1 | 0 | 2903 | 800.3 | 0 | 2886 | 813.9 | 0 | 2899 |
| 5 Reading | 3060 | 806.3 | 0 | 2853 | 751.7 | 0 | 2808 | 814.0 | 0 | 2909 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Extreme AAV values were then examined for each grade and subject combination. Because of concerns about very small schools influencing the results of this analysis, only schools with 30 or more students taking the examination for the grade/subject combination were studied. If the difference in rankings between any of the models was equal to 0, the school was examined for information that it might have in common with other schools with extreme AAV values. Very few campuses maintained their percent passing model ranking when ranked using another model. Most of the rankings of campuses retaining a percent passing model rank were quite low rankings. This finding indicates that model rankings are most consistent when performance is lowest. Schools which do well when ranked under the percent passing model are less likely to retain their ranking across models than schools which do not do well under the percent passing model. It is also apparent that the percent passing model rankings do not coincide with the random intercepts and random slopes models. There were no schools that had the same rank using the results of the school status model and the results of these models.

The change in ranking of a school based on which statistical model was used to calculate the rankings was unexpectedly large. Potentially the most severe impact of these changes would take place in schools eligible for rewards or subject to sanctions based on school ranking. The next analysis addressed schools which were ranked in the top 20 percent of schools using the percent passing model, and the change in rankings when they were ranked using other models. In some cases the different rankings not only moved the school out of the top 20 percent, but to the lowest rankings.

**Top 20 percent**

The 590 grade 4 and 534 grade 5 campuses with 30 or more students which were in the top 20 percent of campuses across the state as ranked by the percent passing model were analyzed separately. The absolute average change statistics for schools with 30 or more students which fall in the top 20 percent at each grade level and subject area are presented in tables 27-30. These tables show the potential impact on schools which not only no longer fall in the top 20 percent but which may fall to the bottom of the rankings, with average absolute values as high as 2,812.

Table 27: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Cohort Difference Model (COHDIFF) for Schools in the Top 20% with 30 or More Students

| Grade | Subject | Schools | PCTPASS – COHDIFF | | |
|---|---|---|---|---|---|
| | | | AAV | Min | Max |
| 4 | Math | 590 | 1133 | 6 | 2713 |
| 4 | Reading | 590 | 942 | 0 | 2812 |
| 5 | Math | 534 | 878 | 3 | 2560 |
| 5 | Reading | 534 | 803 | 1 | 2146 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Table 28: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Single-Level Regression Models (UNSLR & ASLR) for Schools in the Top 20% with 30 or More Students

| Grade | Subject | Schools | PCTPASS – UNSLR | | | PCTPASS – ASLR | | |
|---|---|---|---|---|---|---|---|---|
| | | | AAV | Min | Max | AAV | Min | Max |
| 4 | Math | 590 | 469 | 1 | 2165 | 544 | 2 | 2234 |
| 4 | Reading | 590 | 346 | 0 | 2068 | 486 | 0 | 1871 |
| 5 | Math | 534 | 488 | 2 | 2029 | 512 | 2 | 1952 |
| 5 | Reading | 534 | 382 | 2 | 1717 | 480 | 1 | 1791 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Table 29: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept Multilevel Models (DLINT, UNINT & AINT) for Schools in the Top 20% with 30 or More Students

| Grade/ Subject | Schools | PCTPASS-DLINT | | | PCTPASS-UNINT | | | PCTPASS-AINT | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AAV | Min | Max | AAV | Min | Max | AAV | Min | Max |
| 4 Math | 590 | 639 | 1 | 2478 | 555 | 0 | 2307 | 639 | 1 | 2481 |
| 4 Reading | 590 | 632 | 1 | 2357 | 467 | 0 | 2328 | 633 | 2 | 2362 |
| 5 Math | 534 | 601 | 5 | 2055 | 583 | 1 | 2094 | 602 | 5 | 2056 |
| 5 Reading | 534 | 598 | 1 | 2337 | 502 | 1 | 2333 | 598 | 1 | 2335 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Table 30: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept/Random Slope Models (DLINTSLP, UNINTSLP & AINTSLP) for Schools in the Top 20% with 30 or More Students

| Grade/Subject | Schools | PCTPASS-DLINTSLP | | | PCTPASS-UNINTSLP | | | PCTPASS-AINTSLP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AAV | Min | Max | AAV | Min | Max | AAV | Min | Max |
| 4 Math | 590 | 685 | 1 | 2575 | 627 | 1 | 2596 | 687 | 1 | 2658 |
| 4 Reading | 590 | 686 | 0 | 2725 | 539 | 4 | 2511 | 668 | 2 | 2645 |
| 5 Math | 534 | 676 | 7 | 2158 | 659 | 0 | 2252 | 675 | 1 | 2187 |
| 5 Reading | 534 | 647 | 1 | 2403 | 551 | 2 | 2347 | 630 | 0 | 2436 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Next, a comparison was conducted of which schools stay in the top 20% when ranked by models other than the percent passing model.

***Grade 4 Mathematics Analysis***

The number of schools remaining in the top 20 percent for the percent passing model and each other model varies from 37 percent to 46 percent, except for the cohort difference model. Only 14 percent of campuses ranked in the top 20 percent using the percent passing model are also in the top 20 percent using cohort difference model. There are 32 campuses which are ranked in the top 20 percent using every model. These 32 schools have an average of 72 (range=31 to 183) students with an average of 63 percent (range=4 percent to 97 percent) of students eligible for free and reduced price lunch.

***Grade 4 Reading Analysis***

The number of schools remaining in the top 20 percent for the percent passing model and each other model varies from 20 percent to 52 percent. There are 41

campuses which are ranked in the top 20 percent using every model. These 41 schools have an average of 72 (range-39-146) students with an average of 46 percent (range=0 to 95 percent) students eligible for free and reduced price lunch.

***Grade 5 Mathematics Analysis***

The number of schools remaining in the top 20 percent for the percent passing model and each other model varies from 24 percent to 45 percent. There are 42 campuses which are ranked in the top 20 percent using every model. These 42 campuses have an average of 74 (range=36-158) students with an average of 65 percent (range=0 to 100 percent) students eligible for free and reduced price lunch.

***Grade 5 Reading Analysis***

The number of schools remaining in the top 20 percent for the percent passing model and each other model varies from 21 percent to 47 percent. There are 31 campuses which are ranked in the top 20 percent using every model. These 31 campuses have an average of 69 (range=36-126) students with an average of 58 percent (range=0 percent - 100 percent) students eligible for free and reduced price lunch.

Differences in which schools are ranked in the top 20 percent could make a difference in state accountability systems if rewards are given to personnel in top-performing schools. A potentially even more serious situation could arise from differences in which schools are ranked in the bottom 20 percent, as these schools could face serious sanctions, including school closure. The following analysis explores whether the differences in ranking for schools falling in the bottom 20 percent are as drastic as the differences for schools in the top 20 percent.

**Bottom 20 percent**

The 590 grade 4 and 534 grade 5 campuses with 30 or more students which were in the bottom 20 percent of campuses across the state as ranked by the percent passing model were analyzed separately. The absolute average change statistics for schools with 30 or more students which fall in the bottom 20 percent at each grade level and subject area are presented in tables 31-34.

Table 31: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Cohort Difference Model (COHDIFF) for Schools in the Bottom 20% with 30 or More Students

| Grade | Subject | Schools | PCTPASS – COHDIFF | | |
|---|---|---|---|---|---|
| | | | AAV | Min | Max |
| 4 | Math | 590 | 1159 | 0 | 3258 |
| 4 | Reading | 590 | 997 | 1 | 3294 |
| 5 | Math | 534 | 791 | 1 | 2996 |
| 5 | Reading | 534 | 830 | 1 | 2893 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Table 32: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Single-Level Regression Models (UNSLR & ASLR) for Schools in the Bottom 20% with 30 or More Students

| Grade/Subject | Schools | PCTPASS - UNSLR | | | PCTPASS – ASLR | | |
|---|---|---|---|---|---|---|---|
| | | AAV | Min | Max | AAV | Min | Max |
| 4 Math | 590 | 443 | 0 | 3188 | 492 | 0 | 3191 |
| 4 Reading | 590 | 333 | 0 | 3246 | 438 | 0 | 2874 |
| 5 Math | 534 | 358 | 0 | 2619 | 389 | 0 | 2664 |
| 5 Reading | 534 | 400 | 0 | 2855 | 465 | 0 | 2831 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Table 33: Average Absolute Value (AAV) of Difference Between Models Percent Passing Model (PCTPASS) and Random Intercept Multilevel Models (DLINT, UNINT & AINT) for Schools in the Bottom 20% with 30 or More Students

| Grade/Subject | Schools | PCTPASS – DLINT | | | PCTPASS – UNINT | | | PCTPASS – AINT | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AAV | Min | Max | AAV | Min | Max | AAV | Min | Max |
| 4 Math | 590 | 684 | 2 | 3078 | 571 | 0 | 2950 | 683 | 4 | 3078 |
| 4 Reading | 590 | 759 | 0 | 3140 | 541 | 1 | 3297 | 756 | 1 | 3139 |
| 5 Math | 534 | 661 | 1 | 2972 | 613 | 1 | 2967 | 661 | 0 | 2973 |
| 5 Reading | 534 | 773 | 1 | 2960 | 657 | 0 | 2945 | 772 | 0 | 2960 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

Table 34: Average Absolute Value (AAV) of Difference Between Percent Passing Model (PCTPASS) and Random Intercept/Random Slope Models (DLINTSLP, UNINTSLP & AINTSLP) for Schools in the Bottom 20% with 30 or More Students

| Grade/ Subject | Schools | PCTPASS – DLINTSLP | | | PCTPASS – UNINTSLP | | | PCTPASS – AINTSLP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AAV | Min | Max | AAV | Min | Max | AAV | Min | Max |
| 4 Math | 590 | 849 | 1 | 3169 | 809 | 3 | 3147 | 881 | 0 | 3147 |
| 4 Reading | 590 | 862 | 2 | 3118 | 716 | 0 | 3046 | 898 | 4 | 3129 |
| 5 Math | 534 | 879 | 0 | 2903 | 891 | 1 | 2886 | 916 | 0 | 2899 |
| 5 Reading | 534 | 943 | 1 | 2853 | 878 | 2 | 2808 | 979 | 2 | 2909 |

Schools = Number of schools included in rankings
AAV = Average Absolute Value of difference in ranking
Min = Minimum difference in ranking
Max = Maximum difference in ranking

***Grade 4 Mathematics***

The number of schools remaining in the bottom 20 percent for the percent passing model and each other model varies from 41 percent to 77 percent. There are 151 campuses which are ranked in the bottom 20 percent using every model. These 151 schools have an average of 76 (range=30 to 185) students with an average of 65 percent (range=3 percent to 100 percent) students eligible for free and reduced price lunch.

***Grade 4 Reading***

The number of schools remaining in the bottom 20 percent for the percent passing model and each other model varies from 48 percent to 82 percent. There are 156 campuses which are ranked in the bottom 20 percent using every model. These 156 schools have an average of 78 (range=30-210) students with an average of 71 percent (range=4 percent-98 percent) students eligible for free and reduced price lunch.

***Grade 5 Mathematics***

The number of schools remaining in the bottom 20 percent for the percent passing model and each other model varies from 48 percent to 80 percent. There are 136 campuses which are ranked in the bottom 20 percent using every model. These 136 schools have an average of 83 (range=30 – 200) students with an average of 60 percent (range=0 percent – 100 percent) students eligible for free and reduced price lunch.

***Grade 5 Reading***

The number of schools remaining in the bottom 20 percent for the percent passing model and each other model varies from 42 percent to 79 percent. There are 118 campuses which are ranked in the bottom 20 percent using every model. These 118 schools have an average of 76 (range=30-155) students with an average of 72 percent (range=22 percent-100 percent students eligible for free and reduced price lunch.

When rankings based on each of the different models were compared to each other, it was found that rankings based on the percent passing model were most different from rankings based on the cohort difference model. Rankings were most consistent for the worst-performing schools. When schools were consistently ranked using different models, they tended to be schools that were in the bottom half of the rankings.

The investigation of research questions one through three showed that choice of statistical models made a difference in the ranking of schools. Results calculated using some models were vastly different from results using other models, and there were differences in the extent to which ranking showed a relation with percent of low SES students in a school. Further investigation of the extent to which ranking and percent of low SES students were related was investigated through a simulation study in which the relation between SES and student test score was varied along with the clustering of

student scores within schools, in order to determine the impact of these factors on school ranking.

## Simulated Data Sets

Next, the results using simulated data sets will be discussed. As displayed in Table 35, there were 24 conditions simulated varying the relation between student SES and student test score in 250 simulated schools. Ten datasets were simulated for each of the 24 conditions. Rankings of the 250 simulated schools were then compared across the ten iterations of each condition, as well as across all of the 24 conditions.

Table 35: Simulation Conditions

| *pct* a | *pre*b | *post*c | ICC d |
|---|---|---|---|
| 0.00 | 0.00 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| | -0.10 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| 0.01 | 0.00 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| | -0.10 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| -0.10 | 0.00 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| | -0.10 | 0.00 | 0.05 |
| | | -0.10 | 0.05 |
| 0.00 | 0.00 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| | -0.10 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| 0.01 | 0.00 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| | -0.10 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| -0.10 | 0.00 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |
| | -0.10 | 0.00 | 0.15 |
| | | -0.10 | 0.15 |

a *pct* =coefficient for relation between percent low SES students and $Y_{ij}$
b *pre* =coefficient for relation between student first year test score and *SES*
c *post* =coefficient for relation between student second year test score and SES
d ICC = intraclass correlation

Data were simulated ten times for each condition, resulting in 240 data sets. School rankings were computed for each of the 240 data sets using each of the seven statistical models. Unlike the real datasets, in the simulated data every school had results

for 30 students, therefore every school was ranked using the results of each of the models. The average proportion of students receiving free and reduced price lunch services ranged from 0.46 to 0.54 in the 240 datasets, with averages across the ten iterations for the 24 conditions ranging from 0.49 to 0.51. Across all conditions the average proportion of students receiving free and reduced price lunch services was 0.50.

The ten iterations for each condition were then combined and averaged. Average percent of students receiving free and reduced price lunch was computed based on the iterations for each condition. Average rankings were computed for each condition, and schools were ranked for each condition based on the average ranking across iterations. Average rankings for schools under each of the models were correlated using Spearman rank-order correlations.

After all datasets were created, rankings calculated using each of the statistical models were compared across simulation conditions. First to be addressed was Research Question Four, in which rankings were compared for their differences when the simulated relation between student SES and student achievement changed. Differences were then examined according to Research Question Five when the simulated relation between percent low SES students in a school and student achievement changed. Finally, to address Research Question Six, differences between the multilevel models and the other models were examined when the proportion of variation attributed to schools varied.

First the relation between ranking and percent of low SES students in a school was examined. Interestingly, for the two conditions in which there was no simulated relation between SES and student test score, results were different. As shown in Table 36, when a smaller proportion of variance in student scores was between schools, (ICC=0.05), there was virtually no relation between SES and school ranking when

calculated using the percent passing model, while there was a slight relation between SES and school ranking in all other models. When a larger proportion of variance in student scores was simulated between schools (ICC=0.15), just the opposite occurred. There was no relation between SES and school ranking in any of the models except the percent passing model. When no relation between student test score and SES was simulated, rankings calculated using the NCLB-type percent passing model were more highly correlated with percent of low SES students when a larger proportion of the variance was explained by schools. When a smaller proportion of the variance was explained by schools, the percent passing model was less highly correlated with percent of low SES students than the other models.

Table 36. Correlation Between Ranking and Percent Low SES Students When No Relation Between SES and Test Score is Simulated

| Condition | Model | | | | | | |
|---|---|---|---|---|---|---|---|
| | PCTPASS | COHDIFF | UNSLR | ASLR | DLINT | UNINT | AINT |
| ICC=.05 | 0.06 | 0.14 | 0.14 | 0.14 | 0.15 | 0.14 | 0.15 |
| ICC=.15 | 0.11 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |

When the proportion of student score variance explained by schools increased, the percent passing model showed results more highly correlated with percent of low SES students, while results for the other models showed results more highly correlated when

the smaller portion of the variance was explained by schools. However, as expected, all correlations between ranking and percent low SES students were relatively low.

Results were then compared when there was a simulated relation between SES and student test score. It was expected that these conditions would show a stronger relation between school ranking and school percent low socioeconomic students than the no-relation conditions discussed above. Two conditions had the highest correlation between ranking and percent low SES students. One condition, in which there was no relation between school percent low SES students and student test score but there was the maximum relation between student SES and both first and second year test scores, showed correlations ranging from 0.13 (adjusted single-level and multilevel regression) to 0.22 (unadjusted single-level and multilevel regression). The percent passing model produced results more like the adjusted regressions, at 0.14, while the cohort difference model performed more like the unadjusted regressions, at 0.21. When this condition was simulated with a greater proportion of variance between schools, results were dramatically different. The strongest correlations between school ranking and school percent low SES students were found using the adjusted school-level regression and the unadjusted multilevel models. Results from these models were negatively correlated with percent of free and reduced price lunch students. In other words, they were biased against schools with low percents of low SES students.

In the second condition, which had the strongest relation between school percent low SES and test score along with the strongest correlation between student SES and first year test score, the correlations between ranking and school percent low SES students was 0.23 for the unadjusted regression models, and 0.16 for the adjusted regression models. The percent passing model was most highly correlated with percent low SES students, at 0.26, and the cohort difference model was similar to the unadjusted

regression models at 0.23. When this condition was simulated with a greater proportion of variance between schools, results were very similar. Correlations between school percent low SES students and school ranking were among the highest of the models with 0.15 ICCs. These results are reported in Table 37.

Table 37. Simulated Conditions Under Which Ranking was Most Highly Correlated with Percent of Low SES Students

| Condition | | | | Model | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma_{01}$ a | $\gamma_{10}$ b | $\gamma_{20}$ c | ICC d | PCTPASS | COHDIFF | UNSLR | ASLR | DLINT | UNINT | AINT |
| 0 | -0.10 | -0.10 | 0.05 | 0.14 | 0.21 | 0.22 | 0.13 | 0.13 | 0.22 | 0.13 |
| -0.10 | -0.10 | 0.00 | 0.05 | 0.26 | 0.23 | 0.23 | 0.16 | 0.16 | 0.23 | 0.16 |

a $\gamma_{01}$ = relation between percent low SES students and $Y_{ij}$
b $\gamma_{10}$ = relation between student SES and student first year test score
c $\gamma_{20}$ = relation between student SES and student second year test score$_j$
d ICC = intraclass correlation

The two cases involving the highest correlations between ranking and percent low SES students were conditions in which the relation between student SES and first year test score was the strongest, and either the relation between school percent low SES students or the relation between student SES and second year test score was the strongest. Also, in both conditions the proportion of variance attributable to schools was the lower ICC=.05. When the proportion of variance between schools was higher, the case in which SES was related to first and second year test score showed dramatically different results form the first condition. The case in which test score was related to percent low SES

students and first year test score showed consisted results between the two variance conditions.

While the correlations between ranking and percent low SES students were in the direction expected, the correlations were not as strong as expected. The difference between models in correlation with percent low SES students was also not as great as expected. It may be that the numbers generated, which were intended to look at differences between models when the parameters varied, did not create a situation in which there was enough power to examine differences. Correlations of all model results under all conditions are displayed in Table 38.

Table 38: Correlation Between School Rank and Percent Low SES Students using Simulated Data

| Condition | | | | Model | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *pct* a | *pre* b | *post*c | ICC d | PCTPASS | COHDIFF | UNSLR | ASLR | DLINT | UNINT | AINT |
| 0.00 | 0.00 | 0.00 | 0.05 | 0.06 | 0.14 | 0.14 | 0.14 | 0.15 | 0.14 | 0.15 |
| | | -0.10 | 0.05 | 0.06 | 0.09 | 0.08 | -0.00 | 0.00 | 0.09 | 0.00 |
| | -0.10 | 0.00 | 0.05 | 0.09 | 0.05 | 0.05 | 0.06 | 0.07 | 0.05 | 0.07 |
| | | -0.10 | 0.05 | 0.14 | 0.21 | 0.22 | 0.13 | 0.13 | 0.22 | 0.13 |
| 0.01 | 0.00 | 0.00 | 0.05 | 0.14 | 0.06 | 0.06 | 0.04 | 0.04 | 0.06 | 0.04 |
| | | -0.10 | 0.05 | 0.04 | -0.01 | -0.00 | -0.06 | -0.06 | -0.02 | -0.06 |
| | -0.10 | 0.00 | 0.05 | 0.01 | 0.02 | 0.03 | 0.02 | 0.01 | 0.02 | 0.02 |
| | | -0.10 | 0.05 | -0.05 | -0.02 | -0.02 | -0.08 | -0.06 | -0.02 | -0.06 |
| -0.10 | 0.00 | 0.00 | 0.05 | 0.03 | -0.07 | -0.06 | -0.12 | -0.13 | -0.07 | -0.13 |
| | | -0.10 | 0.05 | 0.07 | 0.03 | 0.02 | -0.15 | -0.15 | 0.02 | -0.15 |
| | -0.10 | 0.00 | 0.05 | 0.26 | 0.23 | 0.23 | 0.16 | 0.16 | 0.23 | 0.16 |
| | | -0.10 | 0.05 | 0.24 | 0.09 | 0.09 | -0.05 | -0.06 | 0.09 | -0.06 |
| 0.00 | 0.00 | 0.00 | 0.15 | 0.11 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| | | -0.10 | 0.15 | 0.04 | 0.09 | 0.09 | 0.03 | 0.03 | 0.09 | 0.03 |
| | -0.10 | 0.00 | 0.15 | 0.04 | 0.05 | 0.04 | 0.04 | 0.05 | 0.05 | 0.05 |
| | | -0.10 | 0.15 | 0.02 | -0.02 | -0.02 | -0.08 | -0.08 | -0.02 | -0.08 |
| 0.01 | 0.00 | 0.00 | 0.15 | -0.06 | -0.02 | -0.01 | -0.01 | -0.02 | -0.02 | -0.02 |
| | | -0.10 | 0.15 | 0.09 | 0.03 | 0.04 | 0.02 | 0.02 | 0.03 | 0.02 |
| | -0.10 | 0.00 | 0.15 | -0.04 | -0.11 | -0.12 | -0.11 | -0.10 | -0.11 | -0.10 |
| | | -0.10 | 0.15 | 0.00 | 0.05 | 0.04 | -0.03 | -0.03 | 0.05 | -0.03 |
| -0.10 | 0.00 | 0.00 | 0.15 | 0.08 | 0.03 | 0.03 | -0.03 | -0.03 | 0.03 | -0.03 |
| | | -0.10 | 0.15 | 0.19 | 0.22 | 0.22 | 0.11 | 0.12 | 0.22 | 0.12 |
| | -0.10 | 0.00 | 0.15 | 0.22 | 0.21 | 0.21 | 0.14 | 0.14 | 0.21 | 0.14 |
| | | -0.10 | 0.15 | 0.09 | 0.09 | 0.09 | 0.00 | 0.01 | 0.09 | 0.01 |

a *pct* =coefficient for relation between percent low SES students and $Y_{ij}$
b *pre* =coefficient for relation between student first year test score and *SES*
c *post* =coefficient for relation between student second year test score and SES
d ICC = intraclass correlation

There were conditions in which results of the adjusted regression models were clearly less highly correlated with percent of low SES students in the school than results of the unadjusted regression models. These are the ICC=0.05 conditions with no relation between school percent low SES and test score or student SES with first year test score, but a strong relation between student SES and second year test score, and the condition in which the relation between percent low SES, student SES and first year score as well as second year score were the strongest, as well as the ICC=0.15 conditions in which there was no relation between school percent or student SES with first year test score, but there was between student SES and second year test score, and the condition in which the relation was strongest between school percent low SES, and either first or second year test score, or both. When a regression model is used to determine school rankings, including SES as a predictor in the regression model is most important when there is a strong relation between student SES and second year test score or a strong relation between school percent low SES students and test score along with student SES and first year test score. Results are displayed in Table 39.

Table 39. Simulated Conditions Under Which Unadjusted and Adjusted Regression Models Differ Most

| Condition | | | | Model | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *pct* a | *pre* b | *post* c | ICC d | PCTPASS | COHDIFF | UNSLR | ASLR | DLINT | UNINT | AINT |
| 0.00 | 0.00 | -0.10 | 0.05 | 0.06 | 0.09 | 0.08 | -0.00 | 0.00 | 0.09 | 0.00 |
| -0.10 | -0.10 | -0.10 | 0.05 | 0.24 | 0.09 | 0.09 | -0.05 | -0.06 | 0.09 | -0.06 |
| 0.00 | 0.00 | -0.10 | 0.15 | 0.04 | 0.09 | 0.09 | 0.03 | 0.03 | 0.09 | 0.03 |
| -0.10 | 0.00 | -0.10 | 0.15 | 0.09 | 0.22 | 0.22 | 0.11 | 0.12 | 0.22 | 0.12 |
| -0.10 | -0.10 | 0.00 | 0.15 | 0.22 | 0.21 | 0.21 | 0.14 | 0.14 | 0.21 | 0.14 |
| -0.10 | -0.10 | -0.10 | 0.15 | 0.09 | 0.09 | 0.09 | 0.00 | 0.01 | 0.09 | 0.01 |

a *pct* =coefficient for relation between percent low SES students and $Y_{ij}$
b *pre* =coefficient for relation between student first year test score and *SES*
c *post* =coefficient for relation between student second year test score and SES
d ICC = intraclass correlation

Rankings across the different conditions were next compared to each other. As the relation between free and reduced price lunch and student achievement changed, the difference in rankings for the different models was examined. For each condition, an AAV was calculated to measure the difference in the ranking of each school by two different models. An AAV was therefore calculated for each school for the difference in rankings under the percent passing model and the cohort difference model, for instance. Rankings calculated using each model were first compared to results from the percent passing model.

Results of the percent passing model were most similar to the cohort difference model in situations in which the variability between schools was the greatest (ICC=0.15) with either a strong negative relation between SES and first and second year test score or

a strong negative relation between percent low SES students and achievement, with no relation between student SES and test score. Results of these two models were the most different when the variability between schools was smaller (ICC=0.05) with a negative relation between SES and first year test score.

Results of the percent passing model were most similar to the unadjusted single level and multilevel regression in the condition with ICC=0.15 with the strongest negative relation between SES and first year test score. Results of the adjusted regression models were more similar to the percent passing model when there was a strong negative correlation between percent low SES and achievement, with no relation between student SES and achievement. The percent passing and adjusted or unadjusted single level regression models included were most different in the ICC=0.05 condition when there was a relation only between student SES and first year test score.

The results of all models were most different from the percent passing model when the variability between schools was at the lower condition (ICC=0.05) and the results of all models were most similar to the percent passing model when the variability between schools was at the higher condition (ICC=0.15) and there was a strong negative relation between student SES and student first year test score or a strong negative relation between percent of low SES students and no relation between student SES and test score.

Absolute value differences were then calculated between the cohort difference model and the regression models. The cohort difference model was most different from the unadjusted single level regression model when the ICC was 0.05 and there was a small positive correlation between percent low SES and test score, with a strong negative correlation between student SES and both first and second year test score. Results were the most similar in the very same condition with the larger, 0.15 ICC. The largest median difference was 5 and the smallest was 2. The results of these models were very similar.

When rankings determined using the cohort difference model were compared to the adjusted single level regression model results were most different in three conditions, strong relation between percent SES and student second year test score (ICC=0.05), strongest relation between all three SES variables and achievement (ICC=0.15), and strongest relation between percent SES and student second year test score (ICC=0.15). Again, the results are not too different, the smallest is 3 and the largest is 7.

When the cohort difference model was compared to the multilevel regression models, in all cases the biggest difference in rankings occurred in the condition with a slightly positive relation between percent SES and student achievement, and strongest negative relation between student SES and second year test score. These differences were much higher than above, at 16 in all three comparisons. The cohort difference model was the most similar to the adjusted multilevel models when there was a slightly positive relation between percent SES and student achievement, and no relation between SES and first year test score with either no relation or a negative relation between student SES and second year test score. The difference between the cohort difference model and the unadjusted multilevel model was negligible (1) in all cases except as mentioned above.

Rankings determined using each of the statistical models were compared for consistency across models. These results varied greatly depending on the models being compared. The NCLB-type percent passing model produced results least like the other models. In addition to most of the other models being regression models, another difference is that this is the only model using percent of students passing the examination. The cohort difference model and the regression models all use some form of average score in determining rankings. This appears to make a bigger difference between model results than the regression/non-regression difference, in that the cohort difference model

results were more like the regression model results than the percent passing model results. The cohort difference model was recommended by Kingston and Reidy (1997) as a good compromise between the sophistication of the regression models and the simplicity of the percent passing model. From the results of the real data in this study, results of this model are more highly correlated with the simple regression models than the multilevel or percent passing models.

Comparability was also shown to be lacking when average change in rankings was calculated. Of the over 3,000 schools ranked in this study, the average change in rankings between models was as much as 988 positions. This finding could have an important impact on the results of state rankings of schools for the purposes of rewards or sanctions. Schools ranked in the top 20 percent using the percent passing model changed rankings as many as 1,133 places when ranked using the cohort difference model. Indeed, only an average of 35 campuses retained their place in the top 20 percent across all models. On the other hand, on average 140 schools maintained their status in the lowest 20 percent using all models. While this number is higher than the number of schools maintaining their status in the top 20 percent, it is alarmingly low as a percent of an average of 560 schools which might be subject to sanctions based on student performance.

# CHAPTER 5: DISCUSSION

The intent of this dissertation was to examine the impact of ranking schools when rankings were determined using different statistical models. Rankings determined using the various models were compared both in relation to their similarity with each other and also the correlation between the ranking of a school and the percent of students in the school receiving free and reduced price lunch services. Analyses were first conducted using real data and then using simulated data.

There were four model types used to rank schools using both the real data and the simulated data. First, a percent passing model ranked each school based on percent of students passing an examination. A cohort difference model was then used to rank schools based on school average test score minus school average test score the previous year. The third group of models comprised a school-level regression model, one of which was an unadjusted model which predicted school average examination score using previous year school average examination score, and the other an adjusted model using previous year school average examination score and school percent low SES students. The fourth group of models consisted of two groups of multilevel models. One group used random intercepts and fixed slopes, while the other group, which was used in the real data study only, used random intercept and slopes. Within each group of multilevel models there were three individual models, including one two-stage Dallas-type model, one unadjusted model and one model adjusted for student SES.

The first research question addressed differences between rankings produced by adjusted regression models, which included student SES as a predictor, and rankings produced by the same regression models but which did not include SES. Tekwe, et al. (2004) found agreement between an adjusted and an unadjusted multilevel model applied

to real data to be between 0.61 and 0.95, with four of six correlations below 0.80. In the current study, correlations of real data rankings using unadjusted and adjusted models were stronger than expected. Correlation of rankings calculated using unadjusted and adjusted single-level regression models were between 0.92 and 0.99, and results between the adjusted and unadjusted multilevel models ranged from 0.88 to 0.97. These are very high correlations; the differences in rankings based on adjusted or unadjusted models may or may not be considered important, depending on the use of the rankings and the extent to which rankings are correlated with percent low SES students in a school.

Rankings using adjusted and unadjusted models to rank schools in the real dataset were less highly correlated for one grade/subject combination, fourth grade reading, than for the other grade/subject combinations. Inclusion of SES in a model made more of a difference for this analysis than for any of the other grade/subject combinations. Further investigation revealed a stronger relation between achievement (as measured by test score) and SES for reading in fourth grade, at 0.10, than for the other grade/subject combinations. The correlation for fifth grade reading was 0.09, whereas for mathematics in both grades the correlation was around 0.04. The reason for a stronger relation between achievement and SES in reading is unknown, but probably is connected to differing exposure to language for low and high SES students. That would explain a difference in reading, but not mathematics, which is less language dependent.

The second research question addressed the extent to which model rankings for the real dataset were correlated with school percent of students receiving free and reduced price lunch. A strong positive correlation indicates that results of a model are biased against schools with high numbers of low socioeconomic students. As expected, models which did not adjust for SES produced school rankings which were more highly correlated with the percent of free and reduced price lunch students in the school. Those

models which did include SES as a predictor showed no correlation between percent of free and reduced price lunch students and the school rankings. Rankings based on the percent passing model were correlated as strongly as 0.41 with percent of low SES students at all grades and subjects.

The exception to this finding using real data was the cohort difference model. This model, although not including SES as a predictor, showed either no relation between rankings and school percent free and reduced price lunch students, or a slightly negative relation. The negative correlation indicates a potential bias against schools with low percentages of free and reduced price lunch students. This model uses the difference in average score from one year to the next of a pseudo-cohort of students, not students matched across the years. It is most likely that these negative correlations result from intense efforts, especially in certain grades, to raise the examination scores of low-performing students. If low performing students made greater gains in average scale score than other students because of these efforts, and if they represent a high proportion of students receiving free and reduced price lunch services, then the year-to-year change in average score would be higher in schools with high percentages of such students. This finding is especially likely when student achievement is measured through scores on a criterion-referenced examination which does not measure achievement at a level higher than the basic expectations for each grade level. The adjusted school level regression model also showed ranking and percent low SES students to be negatively correlated, although less strongly than the cohort difference model.

Comparison of rankings calculated using each of the different types of models was the focus of the third research question. It was expected, based on the findings of Clotfelter and Ladd (1996) that correlation of rankings calculated using the different models would be low, and the present study found correlations as low as 0.19 between

results using the percent passing and the cohort difference model.. Percent passing model results were not correlated strongly with results of any of the other models, because it is the only model which uses percent passing as a criterion rather than test score. Cohort difference model results were more like the school-level regression model than the other models, which was expected because both the cohort difference and school-level regression models use school average examination score for two years to calculate school rankings. The high correlation between results of the multilevel models was expected, based on the research of others (Tekwe, et al., 2004; Webster, 1998). Results from the real data analysis supported the recommendations of Thum and Bryk (1997) in favor of a one-stage multilevel model over a two-stage model. Their argument was that one-stage multilevel models produce results which are more reliable than the two-stage Dallas multilevel model, in part because of the use of residuals in the Dallas model. The current study did not address model precision or reliability, but the extremely high correlation of results using the Dallas model with results using adjusted multilevel models (0.97 to 1.0) supports the use of one-stage models.

Results of the adjusted school-level regression applied to real data were more similar to the adjusted multilevel models with random intercepts than the adjusted multilevel models with random intercepts and slopes. There were larger than expected differences between results of the random intercept and the random intercept/slope models. The multilevel model with random intercepts uses the equations

$$\beta_{0j}=\gamma_{00}+u_{0j} \qquad (14)$$

$$\beta_{1j}=\gamma_{10} \qquad (15)$$

which are combined to yield the following equation for predicting student achievement:

$$Y_{ij}=\gamma_{00}+\gamma_{10}(X_{ij}-\overline{X}\ldots)+u_{0j}+r_{ij} \qquad (16)$$

where $Y_{ij}$ is student examination score, $\gamma_{00}$ is the average achievement level across schools, given $X$ (SES) is zero, and $u_{0j}$ is the residual unique to the intercept, or average achievement, for school $j$. Similarly, $\gamma_{10}$ is the slope, or average change in achievement across schools. Note that in the above case $\beta_{1j}$ is modeled as a fixed coefficient and does not vary across schools.

A random intercept/random slopes model allows for variation in slope including an error term, as follows:

$$\beta_{0j}=\gamma_{00}+u_{0j} \qquad (17)$$

$$\beta_{1j}=\gamma_{10}+u_{1j} \qquad (18)$$

and yields the following:

$$Y_{ij}=\gamma_{00}+\gamma_{10}(X_{ij}-\overline{X}\ldots)+u_{0j}+u_{1j}(X_{ij}-\overline{X}\ldots)+r_{ij} \qquad (19)$$

where $\gamma_{00}$ is the average achievement level across schools, given $X$ (SES) is zero, and $u_{0j}$ is the residual unique to the intercept, or average achievement, for school $j$. Similarly, $\gamma_{10}$ is the slope, or average change in achievement across schools, and $u_{1j}$ is the residual unique to the slope, or change in achievement due to $X$ (SES), for school $j$. In this case, both intercept and slope can vary across schools.

The difference in rankings calculated using the model with random intercepts and slopes indicates that the slope, or effect of SES on achievement, varies between schools. In the model with random intercepts only, the slopes are not specified correctly, consequently the adjusted mean is not estimated well, and the residuals used to rank schools are different from the residuals used from the random intercept/random slope model. Further research is warranted regarding the differences between random intercepts and random intercepts/random slopes models results using the real data in this study.

Average Absolute Value statistics created to measure the difference between rankings of each school based on different models show that a school can change from a

top ranking to a bottom ranking or from a bottom ranking to a top ranking depending upon the statistical model used to determine rankings. Even if there is a set standard and schools are rated passing or not passing, rather than ranked, such large differences could have a serious impact on students as well as teachers and administrators.

If there are consequences for performing at the top or bottom of the rankings, the volatility between model results could have the most serious impact at the top end. Over 3,000 schools were rated by the different models for purposes of this dissertation, and 14 to 46 percent stayed in the top 20 percent when ranked by different models. An average of 37 schools, just over 1 in 100, stayed in the top 20 percent when ranked by all models. At the bottom end, 41 to 82 percent of schools stayed in the bottom 20 percent when ranked by different models, and an average of 140 schools stayed in the bottom 20 percent when ranked by all models.

It is disconcerting to think that so many different schools would be the recipient of rewards or the target of sanctions if a different statistical model was used to calculate school rankings. The impact of sanctions for schools at the bottom end could be as serious as school closure, yet only 41 to 82 percent of schools stayed in the bottom 20 percent of schools when ranked using different models. This means that large numbers of schools might be sanctioned by a state or federal education agency using one statistical model, while the same school would escape sanctions or even be rewarded if the agency used a different statistical model to rank schools.

In 2002, the second year for which data were used in this study, the state system of accountability in Texas required 90 percent of students to pass all subjects tested in order to be rated *exemplary*. Schools were required to have 80 percent of student pass all examinations in order to be rated *recognized*, and 50 to 55 percent passing in order to be rated *acceptable* (Texas Education Agency, 2002). The present study used schools in the

top or bottom 20 percent of the rankings to examine potential effects of differences in rankings on highest and lowest performing schools. The top 20 percent of schools included schools with 95 to 100 percent of students passing the mathematics examination, and the bottom 20 percent included schools with 18 to 83 percent of students passing the reading examination. It is difficult to imagine sanctions being fairly assessed against schools with 83 percent of students passing a statewide examination. The state has not proposed a relative method, such as top 20 percent or bottom 20 percent, as a way of rewarding or sanctioning schools. It is clear, however, that the methods used must be thoroughly thought out, to avoid unintended consequences.

Further investigation of the relation between achievement, free and reduced price lunch and model type was conducted using a simulation study to investigate research questions four through six. There were three simulated relations between school percent low SES students and individual student achievement, and two simulated relations between student SES and achievement. In addition, there were two simulated proportions of variance between schools. It was expected that as the relation between SES and achievement was simulated to be stronger, rankings would be more different between models, and unadjusted models would produce rankings more highly correlated with percent low SES students in a school.

Research question four addressed the differences in rankings when the relation between student SES and student achievement varied. As expected when the relation between student SES and student achievement was simulated only in first or second year test score, differences were greater between the percent passing and the other models, because the percent passing model uses only the second year test score, while the other models use both first and second year test score.

When there was a relation between student SES and student score on both first and second year test, the cohort difference and unadjusted regression models produced results most highly correlated with percent of low socioeconomic students, 0.22. However, in this case even the adjusted regression models produced results correlated at 0.13. The percent passing model rankings were most like the adjusted regression model results in this case, because the percent passing model uses only student test score from the second year.

Research question five addressed the impact on rankings of changes in the relation between percent low SES students and an individual student's test score. When this relation was most strong and student SES and student first year test score were also most highly related, rankings were most highly correlated with percent low SES students. Results of the percent passing model were the most highly correlated with percent of low SES students in this condition, despite the fact that this model does not use student first year test score.

Overall, results addressing research questions four and five were not quite as expected. The relation between SES and student test score was expected to have more impact on school rankings in the form of differences between model results. Additional conditions were simulated which introduced a stronger relation, -0.5, between student SES (high SES = 0 and low SES = 1) and student achievement as well as between school percent low SES students and individual student achievement. In this case, correlations between school ranking and percent low SES students were as strong as 0.51 for the percent passing model, and over 0.3 for the cohort difference and unadjusted models, while remaining at 0.0 for the adjusted regression models. While the relations used to generate the original simulated data were similar to those occurring in the real data sets used in the first part of this study, more information on the impact of SES-achievement

relations on school rankings would have been generated if more extreme relations had been modeled.

The sixth research question addressed differences in rankings using the different models when the proportion of variance between schools varied through manipulation of the intraclass correlation. The intraclass correlation measures the portion of total variation in the dependent variable (student achievement score) that is explained by the second level groups (schools). Results of all models were less strongly correlated with percent low SES students when the greater proportion of variance was between schools. The only exception to this was when there was no simulated relation between SES and student test score, and there was a smaller simulated proportion of variance between schools. As cited earlier, research has generally found the proportion of variance in student test scores between schools to be between 0.10 and 0.20 (Kreft & Yoon, 1994). The current study used intraclass correlations of 0.05 and 0.15 based on the real data used in this study, and more information on the value of using multilevel models, rather than single-level models which assume the statistical independence of scores of students within a school, would have been obtained if a higher intraclass correlation had been simulated.

The differences in rankings depending on statistical model used point out the importance of carefully selecting a model for calculation of accountability rankings. The suitability of a particular model depends on the purpose for which it is employed. When the purpose is to isolate the influence of *school* effects, which are referred to by Raudenbush and Willms (1995) as Type B effects, the adjusted regression models are the most appropriate. Type B effects isolate the influence of school practice, including curricular content, classroom instruction, administrative leadership and utilization of resources from other influences on student achievement, such as student background and school context (Raudenbush and Willms, 1995). Alternatively, if the purpose is to assist

parents in selecting the best of all possible schools which their children might attend, then both *school* and *other* effects can be important. These are called Type A effects by Raudenbush and Willms (1995) and are measured by the percent passing, cohort difference and unadjusted regression models. Such effects do not separate different influences on student achievement. Type A effects include student background and school context (such as average socioeconomic background) as well as school practice.

Type A effects were not consistently measured by all models, as seen in the differences between ratings calculated using the different models. Type A effects indicate overall quality of schools, in that they do not separate effects on achievement that are not under the control of the school. Yet the different models employed in this study were not consistent in measuring overall quality of schools. The percent passing and cohort difference models, the simplest of the statistical models employed, showed the least consistency between results of pairs of models, in part because of the use of different criteria (percent passing vs. change in average test score) between the two models. The correlation of rankings ranged from 0.19 to 0.37 using real data. The cohort difference model performed most similarly to the single level regression models. The correlation of results between the cohort difference and the unadjusted school-level regression model averaged 0.79 and 0.80 between the cohort difference model and the adjusted school-level regression model.

Type B effects isolate the effect of *school* influences on student achievement from *other* influences such as SES, which are not controlled by the school. These effects were measured by the adjusted school-level and multilevel regression models. When using real data, rankings from the adjusted school-level regression model were correlated between 0.49 and 0.62 with the adjusted multilevel models. Results from the adjusted multilevel models correlated very highly with each other, between 0.83 and 1.0.

Simulated results showed all school-level and multilevel regression model results to be very highly correlated. All conditions showed results correlated between 0.95 and 0.99.

Findings from the two studies discussed above, one using real data and the second using simulated data, show that choice of statistical model makes a large difference in what ranking is assigned to a school. Decisions about the appropriateness of any model depend on the purpose of the ranking. This dissertation has discussed two ways of looking at the effectiveness of accountability models, determining the best schools overall (Type A effects) or isolating the influence of schools (Type B effects). It is impractical to think that these are the only criteria that are important to the stakeholders with an interest in public schools, however. School teachers and administrators will be interested in results that are fair to them, and easily understandable. A school principal would most likely prefer a method that allows him or her to calculate their accountability rating upon receipt of student test scores, rather than waiting for the scores to be input in a "giant black box" somewhere in the state capital with school ratings issued months later, despite the statistical precision of the ensuing ratings. Likewise, parents of students in struggling schools should be able to understand the criteria used for labeling the school unacceptable. If the application of a statistical model result in sanctions against schools with high numbers of low socioeconomic status students, then the parents of those same students are the stakeholders in need of understanding the problems at the school. In order for all stakeholders to understand judgments about the performance of their schools, a model must be easily understood by the variety of interested parties in addition to being statistical efficient and accurate.

The percent passing model, which is the model currently used for federal accountability, is the most easily understandable model, but produces results least like the results of other, more complex models. In part this is due to the use of percent of students

passing the examination, rather than the actual test scores used in all of the other models. More importantly, rankings calculated using the percent passing model are most strongly correlated with percent of low SES students in a school, suggesting that sanctions against low-performing schools are destined to affect the poorest schools. A "hybrid model" which allows schools to meet accountability standards in two ways may be preferable to a single path to acceptability. This hybrid could consist of the current percent passing model combined with a cohort different model that allowed schools not meeting the percent passing standard to calculate difference in average score over two years and use that to demonstrate progress for accountability purposes. While it is admirable to set high standards for all students, regardless of socioeconomic status, there is plenty of evidence that students who differ on socioeconomic status also differ on achievement test scores. Until this difference between groups of students in eliminated, either differences in achievement associated with socioeconomic status should be accounted for in statistical models used to rank schools, or a method such as the cohort difference model with results more highly correlated with the student change models should be employed.

While 0.80 correlation between results of the cohort difference model and the school level regression models are quite high, there are still large numbers of schools that will be ranked differently when a cohort difference model is used. One avenue of exploration should be a cohort difference model which uses real cohorts, based on students matched across school years, to determine rankings. These results may or may not be closer to the actual adjusted school level regression models.

## Conclusions

In summary, this dissertation showed that:

- As expected, models which did not include socioeconomic status as a predictor produced school rankings which were more highly correlated with the percent of free and reduced price lunch students in the school. Those models which did include socioeconomic status as a predictor showed no correlation between percent of free and reduced price lunch students and the school rankings.
- The exception to this finding was the cohort difference model. This model, although not including socioeconomic status as a predictor, showed either no relation between rankings and school percent free and reduced price lunch students, or a slightly negative relation.
- Average Absolute Value statistics created to measure the difference between rankings of each school based on different models show that a school can change from a top ranking to a bottom ranking or from a bottom ranking to a top ranking depending upon the statistical model used to determine rankings.
- Type A effects were not consistently measured by all models. The percent passing and cohort difference models, the simplest of the statistical models employed, showed the least consistency between results of pairs of models. The correlation of rankings ranged from 0.19 to 0.37 using real data.
- The cohort difference model performed most similarly to the single level regression models. The correlation of results between the cohort difference and the unadjusted school-level regression model was 0.79.
- When the relation between percent low SES students and an individual student's score was most strong and student SES and student first year test score were also

most highly related, rankings were most highly correlated with percent low SES students.

- When there was a relation between student SES and student score on both first and second year test, the cohort difference and unadjusted regression models produced results most highly correlated with percent of low socioeconomic students, 0.22. However, in this case even the adjusted regression models produced results correlated at 0.13.
- Results of all models were less strongly correlated with percent low SES students when the greater proportion of variance was between schools.
- All of the models measuring Type A results produced rankings that were more similar when there was a larger proportion of variance between schools. As student scores were less independent from each other, models measuring Type A results were more consistent in their performance.
- Type B effects, isolating the influence of *school* from *other*, were measured by the adjusted school-level and multilevel regression models. When using real data, rankings from the adjusted school-level regression model were correlated between 0.49 and 0.62 with the adjusted multilevel models.
- Results from the adjusted multilevel models correlated very highly with each other, between 0.83 and 1.0. Simulated results showed all school-level and multilevel regression model results to be very highly correlated. All conditions showed results correlated between 0.95 and 0.99.
- The cohort difference model may be a worthwhile compromise between the simplicity of the percent passing model and the precision of regression models

Findings from this dissertation show that, in fact, the cohort difference model may be a worthwhile compromise between the simplicity of the percent passing model and the

precision of regression models. While the cohort difference model does not attempt to isolate Type B effects, rankings calculated using this model are similar to those calculated using models which do attempt to isolate Type B effects. A caution regarding use of the cohort different model results from the bias against schools with small percentages of low SES students resulting from use of the cohort difference model. States using a criterion-referenced examination to measure student achievement might be better off using the adjusted school-level or a multilevel regression model. Comparison of the cohort difference model and the other types of models using results from a norm-referenced examination could shed further light on this problem.

# GLOSSARY

| | |
|---|---|
| PCTPASS | Percent Passing Model |
| COHDIFF | Cohort Difference Model |
| UNSLR | Unadjusted Single Level Regression Model |
| ASLR | Adjusted Single Level Regression Model |
| DLINT | Dallas-Type Two-Stage Multilevel Model with Random Intercepts |
| UNINT | Unadjusted One-Stage Multilevel Model with Random Intercepts |
| AINT | Adjusted One-Stage Multilevel Model with Random Intercepts |
| DLINTSLP | Dallas-Type Two-Stage Multilevel Model with Random Intercepts and Random Slopes |
| UNINTSLP | Unadjusted One-Stage Multilevel Model with Random Intercepts and Random Slopes |
| AINTSLP | Adjusted One-Stage Multilevel Model with Random Intercepts and Random Slopes |

# REFERENCES


Alban, T. R. (2002). Evaluating School and Teacher Effectiveness: A Comparison of Analytic Models. University of Maryland, College Park.

Baker, A. P., & Xu, D. (1995). The Measure of Education: A Review of the Tennessee Value Added Assessment System.

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. Journal of Educational and Behavioral Statistics, 29(1).

Betts, J. R., & Danenberg, A. (2002). School Accountability in California: An Early Evaluation. Brookings Papers on Education Policy, 2002, 123-197.

Buckendahl, C. W., Impara, J. C., & Plake, B. S. (2000). Computing composite scale scores for accountability: A validation study of Nebraska's district evaluation model. Paper presented at the Mid-Western Educational Research Association, Chicago, IL.

California Department of Education. (2002). Explanatory Notes for the 2002 Academic Performance Index Base Report. Policy and Evaluation Division. Retrieved December 5, 2003, from the World Wide Web:

California Department of Education. (2002-2003). Explanatory Notes For the 2002-03 Academic Performance Index Growth Report. Policy and Evaluation Division. Retrieved December 5, 2003, from the World Wide Web:

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and Rewarding Success in Public Schools. In H. F. Ladd (Ed.), Holding Schools Accountable. Washington, D.C.: The Brookings Institution.

Coleman, J. S., Campbell, E. Q., Hobson, C. F., McPartland, J., Mood, A. M., Weifeld, F. D., & York, R. L. (1966). Equality of Educational Opportunity: U. S. Department of Health, Education and Welfare, Office of Education.

Consortium for Policy Research in Education. (2000). Assessment and Accountability in the Fifty States: 1999-2000. Consortium for Policy Research in Education (CPRE). Retrieved March 15, 2004, 2004, from the World Wide Web: http://www.cpre.org/Publications/wv.pdf

Darlington, R. B. (2004). The Tennessee Value-Added Assessment System. In J. Millman (Ed.), Grading Teachers, Grading Schools. Thousand Oaks, CA: Corwin Press, Inc.

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

Fletcher, S. H., & Raymond, M. E. (2002). The future of California's academic performance index: CREDO, Hoover Institution, Stanford University.

Graybill, F. A., & Iyer, H. K. (1994). Regression Analysis: Concepts and Applications. Belmont, alifornia: Duxbury Press.

Hanushek, E. A., & Raymond, M. E. (2002). Lessons abou the Design of State Accountability Systems. Paper presented at the Taking Account of Accountability: Assessing Policy and Politics, Harvard University.

Hess, F. M. (2002). Reform, Resistance,...Retreat? The Predicatobe Politics of Accountability in Virginia. Brookings Papers on Education Policy, 2002, 69-122.

Hill, P. T., & Lake, R. J. (2002). Standards and Accountability in Washington State. Brookings Papers on Education Policy, 2002, 199-234.

Hox, J. (2002). Multilevel Analysis Techniques and Applications. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kane, T. J., & Staiger, D. O. (2002). Volatility in School Test Scores: Implications for Test-Based Accountability Systems. Brookings Papers on Education Policy.

Kingston, N., & Reidy, E. (1997). Kentucky's Accountability and Assessment Systems. In J. Millman (Ed.), Grading Teachers, Grading Schools. Thousand Oaks: Corwin Press, Inc.

Kreft, I., & De Leeuw, J. (1998). Introducing Multilevel Modeling. Thousand Oaks, California: Sage Publications, Inc.

Kreft, I., & Yoon, B. (1994). Are Multilevel Techniques Necessary? An Attempt at Demystification. New Orleans, LA: American Educational Research Association.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel Modeling of Individual and Group Level Mediated Effects. Multivariate Behavioral Research, 36(2), 249-277.

Kupermintz, H. (2002). Value-Added Assessment of Teachers. In A. Molnar (Ed.), School Reform Proposals: The Research Evidence. Greenwich, CT: Information Age Publishing.

Lee, K., & Weimer, D. (2002). Building Value-Added Assessment into Michican's Accountability System: Lessons from Other States. East Lansing, Michigan: Education Policy Center at Michigan State University.

Linn, R. L. (1998). Assessments and Accountability. Los Angeles, California: Center for the Study of Evaluation, University of California, Los Angeles.

Linn, R. L. (2001a). The Design and Evaluation of Educational Assessment and Accountability Systems (CSE Technical Report 539). Los Angeles: Center for the Student of Evaluation, National Center for Research on Evaluation, Standards, and Student Teaching, University of Californa, Los Angeles.

Linn, R. L. (2001b). Reporting School Quality in Standards-Based Accountability Systems. CRESST Policy Brief 3, Spring, 2001.

Linn, R. L., & Haug, C. (2002). Stability of School Building Accountability Scores and Gains (CSE Technical Report 561). Los Angeles: University of California.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating Value-Added Models for Teacher Accountability. Santa Monica, California: RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A., & Hamilton, L. S. (2004). Models for Value-Added Modeling of Teacher Effects. Journal of Educational and Behavioral Statistics, 29(1).

Mendro, R. L. (1998). Student Achievement and School and Teacher Accountability. Journal of Personnel Evaluation in Education, 12(3), 257-267.

Meyer, R. H. (2000). Value-Added Indicators: A Powerful Tool for Evaluating Science and Mathematics Programs and Policies. NISE Brief, 3(3).

Meyer, R. H. (2002). Value-Added Indicators: Do They Make an Important Difference? Evidence From the Milwaukee Public Schools. Paper presented at the Annaul Meeting of the American Educational Research Association, New Orleans.

North Carolina (Public Schools of) Division of Accountability Services Reporting Section. (2003a). Setting Annual Growth Standards: 'The Formula'. Retrieved December 29, 2003, 2003, from the World Wide Web: http://www.ncpublicschools.org/Accountability/reporting/2003memo/Standardsfeb2003.PDF

North Carolina (Public Schools of) Division of Accountability Services Reporting Section. (2003b). Determining Composite Scores in the ABCs Model. Retrieved December 29, 2003, from the World Wide Web:

http://www.ncpublicschools.org/Accountability/reporting/2003memo/Standardsfeb2003.PDF

Porter, A., & Chester, M. (2002). Building a High-Quality Assessment and Accountability Program: The Philadelphia Example. Brookings Papers on Education Policy(2002), 285-337.

Public Schools of North Carolina, D. o. A. S., Reporting Section. (2003). Determining Composite Scores in the ABCs Model. Retrieved December 29, 2003, 2003, from the World Wide Web: http://www.ncpublicschools.org/Accountability/reporting/2003memo/Composite.PDF

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models (Vol. 1). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. Journal of Educational and Behavioral Statistics, 20(4), 307-335.

Rouk, U. (2000). 'Tough Love': State Accountability Policies Push Student Achievement. Austin, Texas: Southwest Educational Development Laboratory.

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

Sanders, W. L., & Horn, S. P. (1995). Educational Assessment Reassessed: The Usefulness of Standardized and Alternative Measures of Student Achievement as Indicators for the Assessment of Educational Outcomes. Education Policy Analysis Archives, 3(6).

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System. In J. Millman (Ed.), Grading Teachers, Grading Schools. Thousand Oaks: Corwin Press, Inc.

Stevens, J., Estrada, S., & Parkes, J. (2000). Measurement Issues in the Design of State Accountability Systems. Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, LA.

Stufflebeam, D. L. (1997). Overview and Assessment of the Kentucky Instructional Results Information System. In J. Millman (Ed.), Grading Teachers, Grading Schools. Thousand Oaks, CA: Corwin Press, Inc.

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An Empirical Comparison of Statistical

Models for Value-Added Assessment of School Performance. Journal of Educational and Behavioral Statistics, 29(1), 11-36.

Texas Education Agency Department of Accountability Reporting and Research. (2002). 2002 Accountability Manual. Austin, Texas: Texas Education Agency.

Texas Education Agency Department of Accountability Reporting and Research. (2005). 2005 Accountability Manual. Austin, TX: Texas Education Agency.

Texas Education Agency Pearson Educational Measurement Harcourt Educational Measurement BETA Inc. (2003). Texas Student Assessment Program Technical Digest. Retrieved June 14, 2004, from the World Wide Web: http://www.tea.state.tx.us/student.assessment/resources/techdig/contents.pdf

Thum, Y. M. (2001). Measuring Progress Towards a Goal: Estimating Teacher Productivity using a Multivariate Multilevel Model for Value-Added Analysis. Los Angeles, CA: University of California, Los Angeles.

Thum, Y. M. (2003). No Child Left Behind: Methodological Challenges & Recommendations for Measuring Adequate Yearly Progress (CSE Tech Report 590). Los Angeles, CA: University of California, Los Angeles.

Thum, Y. M., & Bryk, A. S. (1997). Value-added productivity indicators. In J. Millman (Ed.), Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure? Thousand Oaks, CA: Corwin Press, Inc.

U.S. Department of Education. (2001). Elementary and Secondary Education Act as Reauthorized by the No Child Left Behind Act of 2001, 115 STAT. 1425 (107th Congress ed.).

Webster, W. J. (1998). A Comprehensive System for the Evaluation of Schools. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Webster, W. J., & Mendro, R. L. (1997). The Dallas Value-Added Accountability System. In J. Millman (Ed.), Grading Techers, Grading Schools: Is Student Achievement a Useful Evaluation Measure? Thousand Oaks, CA: Corwin Press, Inc.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1998). An Application of Hierarchical Linear Modeling to the Estimation of School and Teacher Effect. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1998). An Application of Hierarchical Linear Modeling to the Estimation of School and Teacher Effect.

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Webster, W. J., Mendro, R. L., Orsak, T. H., Weerasinghe, D., & Bembry, K. (1997). Little Practical Difference and Pie in the Sky. In J. Millman (Ed.), Grading Teachers, Grading Schools. Thousand Oaks, California: Corwin Press, Inc.

Weerasinghe, D., Orsak, T. H., & Mendro, R. L. (1997). Value Added Productivity Indicators: A Statistical Comparison of the Pre-Test/Post-Test Model and Gain Model. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, Texas.

West Virginia Education Information System.NCLB - WV Report Cards. West Virginia Department of Education. Retrieved January 23, 2004, from the World Wide Web:

Wisconsin Department of Public Instruction. (2003, July 3, 2003). Annual Review of School Performance

Year of Testing: 2002-03. Office of Educational Accountability,. Retrieved March 7, 2004, from the World Wide Web: http://www.dpi.state.wi.us/dpi/oea/annrvw03.html

Zvoch, K., & Stevens, J. (2003). A Multilevel, Longitudinal Analysis of Middle School Math and Language Achievement. Education Policy Analysis Archives, 11(20). [Campbell, W. G. 1990. Form and Style in Thesis Writing, a Manual of Style. Chicago: The University of Chicago Press.

Turabian, K. L. 1987. A Manual for Writers of Term Papers, Theses, and Dissertations. 5th ed. Chicago: The University of Chicago Press.

# VITA

Judith Ann Jennings was born in Madison, Wisconsin on February 13, 1954, the daughter of Jan Lee Sime and Germain Melvin Staebell. After completing her work at Monona Grove High School, Monona, Wisconsin, in 1972, she lived in Omaha, Nebraska; Boca Raton, Florida; Tokyo, Japan; and Hong Kong, pursuing her studies throughout. She received the degree of Bachelor of Arts from the University of Texas in December 1995. In September 1996, she entered the Graduate School of The University of Texas. During graduate school she was employed at ESP, Inc., and Texas Education Agency. Since March 2005 she has been employed at Resources for Learning in Austin, Texas.

Permanent address: 5327 Bull Run, Austin, Texas 78727

This dissertation was typed by the author