ESTABLISHING THE UTILITY OF A CLASSROOM EFFECTIVENESS INDEX AS A TEACHER ACCOUNTABILITY MEASURE

Karen L. Bembry

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2002

APPROVED:

Randall Schumacher, Major Professor
Michael Saylor, Committee Member
Robin Henson, Committee Member
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Bembry, Karen L., Establishing the utility of a classroom effectiveness index as a teacher accountability system. Doctor of Philosophy (Educational Research), May 2002, 102 pp., 30 tables, 6 figures, references, 37 titles.

How to identify effective teachers who improve student achievement despite diverse student populations and school contexts is an ongoing discussion in public education. The need to show communities and parents how well teachers and schools improve student learning has led districts and states to seek a fair, equitable and valid measure of student growth using student achievement. This study investigated a two stage hierarchical model for estimating teacher effect on student achievement. This measure was entitled a Classroom Effectiveness Index (CEI). Consistency of this model over time, outlier influences in individual CEIs, variance among CEIs across four years, and correlations of second stage student residuals with first stage student residuals were analyzed. The statistical analysis used four years of student residual data from a state-mandated mathematics assessment (n=7086) and a state-mandated reading assessment (n=7572) aggregated by teacher. The study identified the following results.

Four years of district grand slopes and grand intercepts were analyzed to show consistent results over time. Repeated measures analyses of grand slopes and intercepts in mathematics were statistically significant at the .01 level. Repeated measures analyses of grand slopes and intercepts in reading were not statistically significant. The analyses indicated consistent results over time for reading but not for mathematics.

Data were analyzed to assess outlier effects. Nineteen statistically significant outliers in 15,378 student residuals were identified. However, the impact on individual teachers was extreme in eight of the 19 cases. Further study is indicated.

Subsets of teachers in the same assignment at the same school for four consecutive years and for three consecutive years indicated CEIs were stable over time. There were no statistically significant differences in either mathematics or reading.

Correlations between Level One student residuals and HLM residuals were statistically significant in reading and in mathematics. This implied that the second stage of the model was consistent for all students.

Much is still unknown concerning teacher effect on student achievement, especially when confined to teacher activity within one school year. However, results indicate the utility of using statistical modeling of student achievement within the context of teacher accountability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Jason Millman (1997) in the book, *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* summarized a current disagreement among public educators and constituents:

> The disagreement is not whether student learning is an important goal of teaching. Student learning is. Rather, the split is over how best, in high stakes contexts, to evaluate how well teachers and schools accomplish this task. Usually parents and legislators support the use of gains in student achievement as the criterion of student learning. The vast majority of educational professionals favor measures of teacher knowledge and skills as preferred criteria of the likelihood that student learning is taking place. (p. 3)

How to identify effective teachers who improve student achievement despite extremely diverse student populations and differing school contexts has been a topic of investigation in public education at least since the 1980's (Raudenbush, 1988). The increasing need for public education to indicate to school communities and parents how well teachers and schools improve student learning has led school districts and states to seek a fair, equitable and valid way of measuring student growth using student achievement data.

Raw test scores of individual students are inadequate for assessing teacher effectiveness (Bingham, Heywood, & White, 1991). Raw scores do not factor out

influences on student achievement outside the influence of the school. Some correlates that affect student achievement are beyond the control of the school, e.g., the student's proficiency in English or socio-economic status, and have been well documented (Bingham, et al., 1991).

Nevertheless, the use of some data on student achievement as a measure of teacher effectiveness is being utilized in several public school districts (Millman & Shalock, 1997). Tennessee initiated a statewide accountability system, the Tennessee Value-Added Accountability System (TVAAS), in 1993 that includes an assessment of teacher effectiveness in the content areas of mathematics, science, social studies, language arts and reading. A growing number of school districts, such as the Seattle School District in Washington and the Prince George County Schools in Maryland also employ some form of a value-added, multilevel regression accountability model (Webster, Mendro, Bembry, & Orsak, 1995).

Since 1996, the Dallas Independent School District in Dallas, Texas, has been producing locally-defined Classroom Effectiveness Indices (CEIs) for teachers using individual student test scores after controlling for both student and school characteristics with a multi-level regression model. The CEIs were produced primarily to assist teachers and administrators in planning instruction for students. The CEIs were to initiate discussions of successful and unsuccessful instructional practices with the intent that instruction and therefore student achievement would continue to improve throughout the district. However, repeated requests from the community for a process with which teachers can be held accountable for student

achievement may change the intent of the CEIs. A Dallas ISD Teacher Evaluation Task Force developed a plan that asked for the dismissal of a teacher after three consecutive years of low CEIs. The plan was voted down and not implemented. However, in this increasingly high-stakes context of student achievement, there is a need for a teacher accountability system. Currently, the Dallas School Board has initiated the process of developing a teacher incentive pay plan that must include a student achievement component. The CEIs were developed and may become one measure in that incentive pay plan.

This research study investigated the utility of using a CEI as a measure of teacher effectiveness. The study investigated the consistency of the regression model over time, the influence of outliers to individual teachers' CEIs, the variance among teachers' CEIs over time, and the correlation of the final student residual to the initial raw score using multiple years of data. By investigating the consistency of the regression model over time, it was established whether or not teachers and schools were being measured by similar standards from year to year. Investigating the influence of outliers on individual CEIs established the degree of stability necessary to include the CEI as an indicator of adjusted student achievement. Investigating the change in the Classroom Effective Index of a teacher over time further established the degree of stability of a CEI as an estimator of teacher effectiveness. Finally, a correlation between the initial student residual and the adjusted Classroom Effectiveness Index score indicated consistency between the initial and final stage of the process and impartiality of the measure.

Rationale for Study

An effort to connect student achievement data to teacher effectiveness in the United States has spanned the last 50 years (Millman & Shalock, 1997). The use of sophisticated statistical techniques to identify school and teacher effects, however, is a relatively recent phenomenon. A common methodological criticism of educational research had been the failure to account for the nested design inherent in public school data, i.e., students within classes within schools (Raudenbush, 1988). Influences on student achievement outside the control of the school must also be considered in any analysis used for identifying teacher effectiveness. It has only been in the last twenty years that greater computer capabilities, new statistical analyses and newly-developed software capable of modeling nested design data that statisticians have been able to model student achievement as a measure of teacher effectiveness (Bryk & Raudenbush, 1992). The components of a multi-level regression model as well as the outcomes of the model must be studied over time, because of new analytical procedures and the consequences of relating teacher effectiveness to student achievement.

The importance of student achievement in the Dallas Independent School District was of such concern to its constituents that in 1991 a community task force (Commission for Educational Excellence) directed school personnel to identify a way to measure school and teacher effectiveness fairly across all schools (Commission Report, 1991). The Commission's final report asked the District to develop a

comprehensive, results-oriented accountability system that would measure several school outcomes fairly. A major component of the accountability system was the need to develop a fair measure of school and teacher effectiveness. Large differences in student populations and in the school environment across schools required a measure that adjusted for these differences if schools and teachers were to be fairly compared. These differences are indicated in the variables included in the model.

After investigating numerous statistical models, a two-level hierarchical regression model (HLM) for measuring school and teacher effectiveness was identified (Webster et. al., 1995). In keeping with the Commission's directive, the Dallas HLM model generated both school and teacher level measures, the School Effectiveness Indices and the Classroom Effectiveness Indices, that were adjusted for student and school characteristics.

The teacher-level accountability measure, the Classroom Effectiveness Index, has been investigated for bias due to the ethnicity or gender of the teacher, and number of years in teaching with no statistically significant bias identified for any of these teacher characteristics. Also, the students' previous levels of achievement, teaching low achieving students rather than high achieving students has been investigated with no statistically significant bias identified (Bembry, Weerasinghe, & Mendro, 1997). However, the stability of the Classroom Effectiveness Index had not been investigated over time.

The Classroom Effectiveness Index has been used as a planning tool for teachers and administrators. The CEIs were additional measures of prior student

achievement that teachers and school administrators utilized to assess previous instruction and to plan instructional adjustments for the upcoming school year. This was considered a low-stakes application of the Dallas HLM model and its outcomes, and consequences to individual teachers were kept to a minimum. However, in the current climate of increasing requests for school and teacher accountability, the possibility of including the Classroom Effectiveness Index as a measure within an accountability system with consequences for teachers is feasible. Including the Classroom Effectiveness Index in such an accountability system would create a high-stakes situation in which the nature of the CEI would need to be scrutinized.

There are other implications to establishing the Classroom Effectiveness Index as a measure of teacher accountability that extend beyond the Dallas Independent School District. If the measure is established as a consistent, stable and reliable measure that is transferable to other districts and educational settings, it may contribute to the understanding that all students need equal access to a quality education, may influence student remediation policies as a school's responsibility to modify the impact of an ineffective teacher, and may contribute additional information to the discussion of whether or not an ineffective teacher can improve. Policies governing teacher recruitment, teacher evaluation, and teacher retention may also be affected.

Understanding the statistical properties of the Classroom Effectiveness Index is important in establishing the utility of the CEI as a viable measure of a teacher's influence on student achievement. It is important to investigate the consistency of the

Dallas HLM model over time, the influence of outliers on individual teachers' CEIs, the variance of the model's components over time, and the correlation of the final CEI and the initial student residual.

Problem Statement

The most important classroom influence on student achievement is the teacher (Sanders & Rivers, 1996). It has also been established that the detrimental effect of a poor teacher on student achievement lasts beyond the initial school year (Sanders & Horn, 1995). This effect of poor teaching on student achievement creates a responsibility for school districts to identify and promote teaching that improves student learning. The diversity of the student population and the differences in the quality of instructional settings add to the complexity of the issue. Statistical models must be developed to isolate teacher effectiveness from the confounding influences of other measurable factors; therefore Dallas Independent School District created the Dallas HLM model.

The Dallas Independent School District has developed a two-level HLM model that controls for both student influences (gender, ethnicity, language proficiency, and socio-economic status) and school influences (mobility, overcrowdedness, average school-level socio-economic status, percent minority students, and percent limited English students). However, the consistency of the Dallas HLM model over time has not been established. The utility of the Classroom Effectiveness Index also needs to be confirmed if it is to be used as a measure of teacher accountability.

Purpose of Study

The purpose of this study was to examine the statistical properties of the Dallas HLM model Classroom Effectiveness Index over time to determine its utility. In addition, the correlation of the predicted value and the final statistic were investigated. This study therefore investigated the Dallas HLM model data across teachers within schools and across schools over time.

Research Questions

This study evaluated several aspects of the Dallas HLM model over a period of four years. The study analyzed student residuals from a Level One multiple regression equation, where student level characteristics were removed from both previous and current test scores. This analysis included an examination of the influence of student residual outliers on individual teacher CEIs. This study also examined four years of the Dallas HLM model for district-wide slopes and intercepts after individual school characteristics were removed. Finally, the study investigated the strength of the correlation between the individual adjusted student residual used in the predicted CEI and the Level One residual for each of the four years. The study therefore investigated the following research questions:

1. Do CEIs produce consistent results over time?
2. How do outliers affect Classroom Effectiveness Indices?
3. Do teachers differ in average CEIs over time?
4. What is the correlation between the predicted CEI and the Level One residual?

Definition of Terms

Classroom Effectiveness Index (CEI): the mean scaled score of a teacher summed from the individual student residuals after student characteristics (Level One) and school characteristics (Level Two) are removed. Student residuals for the analysis are assigned to teachers by the final grade report of the year.

Continuously enrolled: students who are registered on a campus by the first day of the second six weeks and remains on that campus through the day of testing.

Effective teacher: a teacher "who raises the achievement level of his or her students significantly above the predicted" achievement after accounting for the influence of school and student level characteristics (Bingham, Heywood & White, 1991, p. 192).

Ethnicity: student ethnicity is defined as African American, Hispanic, and other, reflecting the major ethnic groups of students in the district.

Hierarchical Model: a statistical model for nested research designs which reflects the influence variables at one level have on variables at another level.

Language proficiency: students are grouped as English proficient or non-English proficient using information from the campus-level Language Proficiency Assessment Committee.

Level One residuals: the remaining error terms from a multiple regression procedure controlling student outcome and predictor variables for gender, ethnicity, language proficiency, and socio-economic status.

Level Two residuals: the residualized student predicted score after adjusting for school characteristics of mobility, overcrowdedness, average school-level socio-economic status, percent of minority students, and percent of limited English students.

Mobility: a school percent computed by the average number of students entering or leaving a campus throughout the school year (average yearly transactions) divided by the average number of students enrolled for that school year (average daily membership).

Overcrowdedness: a school percent computed by dividing the average daily student membership by the optimum number of students established for the campus by the district (capacity of the building).

Socio-Economic Status: a set of student indicators that include whether or not a student is on free or reduced lunch, the block-level average family income for that student, the block-level average family education, and the block level family poverty index. The block level variables are established by census information.

School average socio-economic status: a set of indicators for a school including the percent of students on free or reduced lunch, the school average family income, the school average family education, and the school average family poverty index.

TAAS: the Texas Assessment of Academic Skills (TAAS) - state criterion-referenced tests in mathematics and reading administered in grades three through eight

and in grade ten. The TAAS was first administered at grade eight and used for school accreditation in 1994 in Texas.

Utility: the level of statistical consistency and integrity sufficient for using an HLM model to identify effective and ineffective teachers.

Limitations of the Study

Although the Classroom Effectiveness Indices included in the study were computed for two content areas, reading and mathematics, as assessed by the Texas Assessment of Academic Skills (TAAS), and over four years, 1997 - 2000, only one grade level, Grade 8, was included in the study. This study therefore did not encompass possible factors existing at other grade levels. Similarly, the study was confined to a Texas state student achievement test (TAAS). This study therefore did not assess possible differences in the regression solutions using any other measure of student achievement, such as the nationally normed Stanford 9 or the Iowa Test of Basic Skills (ITBS). Also, students with excessive absences, defined by the Dallas ISD Accountability Task Force as 20 absences or more per year, were removed from the sample. Any achievement data from students with more than 20 absences were not included in the database, and any potential information from this student population was not included in the study.

Delimitations of the Study

Changes to the Texas Assessment of Academic Skills (TAAS) occur each school year. The number of test items, item difficulty and the objectives tested change from school year to school year such that the ability to develop definite information

concerning student achievement using the TAAS over time is compromised . In addition, only four years of student and school level data for replicating the Dallas HLM model have been retained and used in this study.

CHAPTER TWO

REVIEW OF THE LITERATURE

This study investigated the statistical properties of a two stage hierarchical model estimating teacher effectiveness using student achievement data. Both the statistical validity of hierarchical models and the use of the models in educational research have been researched for the past two decades (Raudenbush, 1988). This chapter consists of three components. First, research concerning the use of student achievement in teacher assessment and accountability are detailed. Second, research concerning hierarchical modeling is summarized. Finally, research concerning the Dallas HLM model is reviewed.

Student Achievement and Teacher Accountability

The use of student achievement data to assess teacher accountability in public schools is a recent phenomenon (Kingston & Reidy, 1997; Mendro, 1998; Sanders & Horn, 1993; Schalock, Schalock, & Girod, 1997; Webster, Mendro, Orsak, & Weerasinghe, 1997). Among the first to attempt to identify teacher effectiveness that would lead to teacher accountability is a 1991 article by Bingham, Heywood, and White. The article reports on an "empirical investigation designed to determine if it is possible to hold teachers accountable for the academic performance of their students" (p. 192). In order to accomplish this, the authors identified a synthesis of 35 variables that may have an effect on student achievement, including individual

student characteristics, family characteristics, peer group characteristics, teacher characteristics, and school characteristics (Bridge, Judd, & Moock, 1979). In addition to the variables identified in the research, previous level of achievement was also included, since the authors wanted to look at measuring achievement within one school year. Identified variables were entered into a multiple regression equation, with ITBS reading scores as the dependent variable, to predict scores for students. The predicted score was subtracted from the actual score to create a residual. Residuals were aggregated by school and classroom to see whether or not differences existed at the school or classroom level.

Each of the variables was included in a series of regression equations in order to identify their predictive value. Three sets of characteristics were among those empirically identified as strong predictors influencing student achievement: student characteristics, including ethnicity, gender, and socioeconomic status; previous level of achievement; and school composition variables (Bingham, Heywood & White, 1991). Using these predictors as well as others available in the local database (number of years in 5^{th} grade, attendance, earlier test scores) in a two-stage multiple regression analysis generated sets of residuals for each school and classroom in the study. It was found that:

> the most important conclusion is that teachers can be evaluated using this method of predicting student performance and comparing it with actual outcomes. Our experiment showed that we can differentiate among teachers on the basis of how much their students have learned in comparison with what comparable students ordinarily learn. (p. 214)

In the attempt to utilize student achievement data for assessing teacher effectiveness, there has been considerable debate concerning the methodology and the appropriateness of using student data at all (Darling-Hammond, 1997; Glass, 1990; Millman, 1981; Raudenbush & Bryk, 1989, Thum & Bryk, 1997; Webster & Mendro, 1997). The influence of factors other than the teacher on student achievement is the basis for this debate, with one faction believing that the influences on student achievement cannot be measured (Darling-Hammond, 1997). It is within this context that the investigation of statistical models that measure student achievement continues.

However, the successful application of statistical models within schools and school districts such as the one posited by Bingham, Haywood and White (1991) for determining teacher effectiveness has been limited to a few locations. One of the most prominent statewide systems is the Tennessee Value Added Assessment System (TVAAS) for the state of Tennessee (Sanders & Horn, 1994). This model produces district, school, and teacher level information on student achievement gains using data from state standardized achievement tests in grades 2-8 (Baker, Xu, & Detch, 1995). Student and district data over three years are included in the computations.

The Dallas Independent School District in Texas and the Prince George County Schools in Maryland are operating districtwide, value-added student achievement models (Webster et. al., 1994; Phillips & Adcock, 1997). Both districts utilize a two-level HLM model with student and school characteristics being

accounted for. Student and school level data for two years are used in the analyses in the two systems, one year of outcome data and one year of predictor data..

Accountability systems specifically measuring the effect of the teacher are less prevalent. However, how a teacher affects student achievement has now been documented (Sanders & Rivers, 1996). Following cohorts of students over four years (1992 – 1995) in three urban school districts in Tennessee, Memphis, Knoxville and Nashville, Sanders and Rivers utilized a multilevel longitudinal analysis across teachers to compute "teacher effects" (p.2). Using the teacher effects, the distribution of teachers was divided into quintiles for each year, with the least effective teachers in the first quintile and the most effective teachers in the fifth quintile. Students were then tracked in a post hoc process through varying series of effective and less effective teachers. Results of the analysis indicated that regardless of a student's initial achievement level, the top quintile teachers indicated academic progress for all students, while students of the first quintile teachers, regardless of their initial level of academic achievement, made fewer gains. The study further concluded that "teacher effects are both additive and cumulative with little evidence of compensatory effects of more effective teachers in later grades" (p.6). Students who had less effective teachers did not catch up to their academic peers with an effective teacher at a later date.

A second study found that ineffective teachers had a long term effect on a student's achievement (Jordan, Mendro, & Weerasinghe, 1997). Student cohorts were identified using similar initial achievement levels across a district in grades 1, 2,

3, 4, and 5. Similar to the Tennessee study, the distribution of the Classroom Effectiveness Indices was used to assign teachers to quintiles with the least effective teachers assigned to the first quintile and the most effective teachers assigned to quintile five for each of the three years in the study. The students' achievement was then assessed three years later after the students were taught by a series of effective and ineffective teachers. The students who had three ineffective teachers had lower achievement gains by as much as forty percentile points. Overall, the range of achievement gains for the three years was - 42.29 to 24.41. Of the students who progressed through a series of first quintile teachers (111), or a series of first and second quintile teachers (112, 121, 122 or 211) the range of achievement gains for the three years was - 42.49 to - 13.46 percentile points. The average loss of achievement was - 23.00 percentile points for this group of students.

The two studies utilized student populations from different states and grade levels, and employed differing state tests, statistical methodology and analysis procedures. However, the similar results indicated that less effective teachers had a long-term effect on student achievement.

Hierarchical Linear Modeling

McLean, Sanders and Stroup (1991) identified the combined development of "existing computer hardware and theoretical knowledge about mixed linear models" (p. 62) thereby creating interest in multi-level modeling in the early 1990s. In addition, the development of software capable of incorporating HLM statistical models allowed greater access for educators. Raudenbush and Bryk (1992) called

this process Hierarchical Linear Modeling "because it conveys an important structural feature of data that is common in a wide variety of (educational) applications" (p. 3-4). According to Bryk and Raudenbush, "these submodels express relationships among variables within a given level, and specify how variables at one level influence relations occurring at another" (p. 4). Hierarchical linear models contain complex residual error structures due to the nested aspect of the model. In their 1992 book, Bryk and Raudenbush examined the assumptions of the multi-level models and recommended procedures for model building based on these assumptions.

Sample Size

Standard errors and adequate sample size for hierarchical linear models have also been investigated. Snijders and Bosker (1993) investigated a two-level model of students within schools to estimate optimal sample sizes at both levels to minimize standard errors. In their investigation of the standard errors of the regression coefficients, they concluded that $n>10$ was the minimum size acceptable for classroom level investigations. Prior to the early nineties, only large national data sets were used in educational research in hierarchical linear modeling. The estimate of smaller adequate sample sizes allowed researchers to extend the use of the models to include individual school districts and individual classrooms.

Centering

Within the nested design of HLM, the interpretation of each variable's value also needed to be clarified. The centering options available in HLM produced different interpretations of the Level One residuals. Level One predictors must be

centered in one of four ways: X metric (uncentered), grand mean, group mean, or a cut score (Bryk & Raudenbush 1992). Centering at Level One determines the meaning of the level-one intercept. Although the choice of which centering method should be used is driven by the research question, the stability of the centering methods needs to be explored in order to create a context for interpreting outcomes.

Three of the centering methods were investigated for stability (Schumacker & Bembry, 1995). In this study, the effects of the centering options on the level-one intercept was assessed using student level (level-one) variables of ITBS reading test scores for 1993 and 1994, and free or reduced lunch status. School level (level two) variables were graduation rate and the percent of students in advanced diploma plans for the 26 high schools included in the study. It was found that Level One variables centered on either the grand or group mean appeared more stable than when centered on the X metric or when uncentered. The group mean centering method for both level-one predictors identified the same intercept and reliability estimate ($\beta_0 = 16.85$; $r = .99$) as the initial null model ($\beta_0 = 16.85$; $r = .98$). The grand mean centering method was similar ($\beta_0 = 16.77$; $r = .89$), while the uncentered method yielded different results ($\beta_0 = 6.46$; $r = .47$). Therefore, "a researcher will typically center some or all Level 1 (student-level) predictors at either the grand mean or group mean to add stability to the estimation process and provide for intercepts that can be meaningfully interpreted" (p.7).

Centering in HLM was one of several issues addressed in a 1996 journal article, *Measuring School Effects with Hierarchical Linear Modeling: Data*

Handling and Modeling Issues (Adcock & Phillips, 1997). According to Adcock and Phillips, centering on the group mean is used in school effectiveness research when studying differences in school means. Grand mean centering should be used when investigating school differences against a district mean. The importance of establishing the unit of analysis within a nested design, equations for estimating school effects, and clarifying the difference between the Empirical Bayes estimates and OLS estimates were included in the article. Establishing a unit of analysis is important in HLM because it identifies not only the level of data to be used in the analysis but the relationship among the variables within the nested design (i.e. students within schools, schools within districts). Equations for estimating school effects need to determine the specifications necessary for the inclusion of variables within the design. Prior research as well as district level information may establish the viability of the variables included. The difference between the Empirical Bayes estimates and the OLS estimates is an important differentiation in HLM. The study recommended that the Empirical Bayes residual be used in school effectiveness research because it included information from other similar schools in the analysis, countering the small sample sizes often encountered at the school or classroom level.

Outliers

The detection and modeling of outliers in two-level models has only begun to be investigated (Sheehan & Han, 1996). In this study, a cross-level exploratory analysis investigated the type of influence discrepant schools had on the estimation of HLM. Data were generated for three sets of schools: those with no outliers, those

with 10% outliers and those with a single outlier. Analyses indicated that the intercept changed little from a single outlier in intercept, slope or in the combination of intercept and slope (β_0 = 192.20, 191.70, 191.90) from the null (β_0 = 192.00). The reliability of the intercept increased with an outlier in intercept, slope, and combination of intercept and slope (r = .853, .936, .803) from the no outlier data set (r = .486). There was little change in β_1 or its reliability in any of the single outlier data sets. The slopes and intercepts of the 10% outlier data sets changed across all outlier combinations with an increase in the reliability estimates ($r_{\beta 0}$ = .992, .997, .986; $r_{\beta 1}$ = .046, .465, .432). Outliers at Level One had little effect on the level two regression coefficients (γ_0 and γ_1), although the standard error estimates increased. Therefore, the Level One outliers produced "changes in the parameter estimates and their standard errors that resulted in conservative tests of significance" (1996, p. 8).

Missing Data

Missing values and the effect on multi-level models has also been investigated. Several of the current solutions to missing data, including the replacement of a missing value with a probable value, were investigated within the context of a two-level HLM model (Orsak, Mendro, & Weerasinghe, 1998). Sixth grade Iowa Test of Basic Skills (ITBS) scores in reading and mathematics for 1995 and 1996 (n = 5,197) were utilized in the analysis with student characteristics of ethnicity, socio-economic status, English proficiency status and gender included as conditioning variables. Truncated data sets were developed with 1%, 2%, 5%, 10%, and 20% missing data per school. To eliminate bias due to school size, each of the 87

schools were reduced to 30 students per school (n = 2,610). The missing test scores were estimated using HLM, an OLS procedure, and by determining the average test score for each school and compared to the actual data set. It was found that

> HLM estimates and OLS estimates are both similar to the original data up to approximately the 10% level [of missing values] whereas HLM estimates are more accurate to the original for greater percentages. This highlights the advantage of implementing HLM in educational data analysis when a greater percentage of data is missing (1998, p. 11).

Issues surrounding hierarchical linear models continue to be investigated as empirical models of school and teacher influences continue to be studied. The nested design of most educational settings (classrooms nested within schools and schools nested within districts) requires a multi-level model.

The Dallas HLM Model

The Dallas Independent School District investigated several alternative regression methodologies in the development of the current Dallas HLM model (Webster et. al, 1994). All of the alternatives used the same multiple indicators and outcomes at both the student and school level. The student level indicators included English proficiency, gender, ethnicity, and socio-economic status. The previous level of student achievement was also included in the model. School level indicators were school mobility, school overcrowding, average school socioeconomic status, and percent of minority students within a school.

It was determined that correlations among all multiple regression models and the HLM models indicated similar results across the statistical models. In fact, “the

correlation between the results produced by Dallas-FULL and HLM-FULL, two comparable models, was .970" (Webster et al., 1994, p. 25).

The study also investigated whether or not school level characteristics influenced the school rankings. Minimal correlations with school variables were identified, "the highest proportion of variance in rankings accounted for by any of these models being less than 3% (Webster et al., p. 32).

A post-hoc analysis of the first year's Classroom Effectiveness Indices investigated whether bias existed due to teacher ethnicity, gender of the teacher, students' previous levels of ability, or a teacher's number of years of experience (Bembry, Weerasinghe, & Mendro, 1997). Post-hoc analyses indicated the composite CEIs were free from bias due to a teacher's ethnicity, gender, and the academic achievement of his or her students. The only significant teacher effect found was for first year teachers who had a statistically significant lower CEI than all other groups of teachers.

The analysis also investigated bias for type of school year calendar. No bias was indicated between a traditional calendar campus and a year-round campus. Possible bias resulting from the number of students included in the CEI was identified; however, the correlation was not statistically significant at the .05 level.

Summary

Various statistical components of hierarchical models have been investigated since the process gained popularity in educational research in the late l980s. Establishing assumptions, investigating adequate sample sizes and standard error

terms, interpreting centering methods and the influence of missing data, and the influence of outliers within the two-level model have contributed to the confidence with which hierarchical modeling is currently used in educational research.

The Dallas HLM model has extended both theory and the use of hierarchical models in public education by comparing alternative hierarchical models in addition to comparing hierarchical models to other methods of regression analysis. Finally, the use of regression methodology and student achievement in assessing teacher effectiveness is still an uncommon practice. The investigation of a model used by a public school for assessing teacher effectiveness over a period of years will extend the understanding of hierarchical models in education and solidify the use of the model for teacher accountability, especially in investigating the utility of using a Classroom Effectiveness Index.

CHAPTER 3

METHODOLOGY AND PROCEDURES

This study investigated several aspects of the student residualized scores contained in the Classroom Effectiveness Index computed for each of four years at the eighth grade level. This chapter describes the subjects used in the study, type of analyses to be used for each research question, and the statistical hypotheses to be investigated.

Subjects

The research questions for this study were investigated using school and student level data for eighth-grade students who were continuously enrolled in 24 middle schools within a large urban school district. Continuously enrolled is defined as any student enrolled on a school campus by the first day of the second six-week grading period and remaining through the spring test date. Students included in this study had test scores for the Texas Assessment of Academic Skills (TAAS) in reading and/or in mathematics from the current and the previous school year. This study was limited to students enrolled in the general language arts course (Course 1100) for reading data and in the general mathematics course (Course 2550) for mathematics data. Students with excessive absences, defined as 20 absences or more per year, were removed from the sample. Both male and female students were included.

Table 1

Data Summary by Year (Mathematics Course 2550)

Year	Schools	Teachers	Students
1997	28	38	1715
1998	28	57	2717
1999	28	57	2577
2000	28	44	1790

Table 2

Data Summary by Year (Language Arts Course 1100)

Year	Schools	Teachers	Students
1997	28	32	1503
1998	28	43	2175
1999	28	43	2426
2000	28	33	1468

Dallas HLM Model

The Classroom Effectiveness Index (CEI) is a measure of student achievement that uses two years of standardized test scores and other student-level and school-level covariates. For the purposes of this study, student scores on the spring Texas Assessment of Academic Skills (TAAS) in reading and mathematics for the years of 1997, 1998, 1999, and 2000 were the outcome variables of achievement. The test scores were computed into individual student residualized gain scores for each student using a two stage (Level One and Level Two) HLM regression equation.

These residualized gain scores were grouped and assigned to teachers using student and teacher identification numbers, which resulted in a database of student residualized gain scores by teacher.

Level One: Student Level Fairness Stage:

In the first level, outcome and predictor variables were regressed against covariates called fairness variables using multiple regression. In other words, raw test scores were regressed against nine student level characteristics or covariates. Covariates included ethnicity, limited English proficiency status, gender, and variables indicating socio-economic status (SES).

Y_{1ij} = Outcome variable of interest for each student *i* in school *j*.

X_{1ij} = African American English proficient status (1 if yes, 0 if all others)
X_{2ij} = Hispanic English proficient (1 if yes, 0 if all others)
X_{3ij} = Limited English Proficient (1 if yes, 0 if all others)
X_{4ij} = Gender (1 if male, 0 if female)
X_{5ij} = Free/reduced lunch (1 if yes, 0 if not on free/reduced lunch)
X_{6ij}= Block-level average family income
X_{7ij} = Block level average family education
X_{8ij} = Block level average family poverty index
X_{9ij} = Variable *k* for the i^{th} student in school *j*

These variables with specific interactions were designated by the Accountability Task Force and must be included in the multiple regression equation as follows:

$$Y_{ij} = ß_0 + ß_1X_{1ij} + ß_2X_{2ij} + ß_3X_{3ij} + ß_4X_{4ij} + ß_5X_{5ij} + ß_6X_{6ij} + ß_7X_{7ij} + ß_8X_{8ij} +$$
$$ß_9(X_{1ij}X_{4ij}) + ß_{10}(X_{2ij}X_{4ij}) + ß_{11}(X_{3ij}X_{4ij}) + ß_{12}(X_{1ij}X_{5ij}) + ß_{13}(X_{2ij}X_{5ij}) +$$
$$ß_{14}(X_{3ij}X_{5ij}) + ß_{15}(X_{4ij}X_{5ij}) + ß_{16}(X_{1ij}X_{4ij}X_{5ij}) + ß_{17}(X_{2ij}X_{4ij}X_{5ij}) +$$
$$ß_{18}(X_{3ij}X_{4ij}X_{5ij}) + e_{ij}$$

Where $e_{ij} \sim N(0,\sigma^2)$

Level Two: School Level Equation:

A second level analysis in the hierarchical model adjusted for school level variables when regressed against the residualized student level gain scores from the first level analysis. The school level covariates were as follows:

W_{1j} = School mobility
W_{2j} = School overcrowdedness
W_{3j} = School average family education
W_{4j} = School average family income
W_{5j} = School average family poverty index
W_{6j} = School percent on free/reduced lunch
W_{7j} = School percent minority
W_{8j} = school percent African American
W_{9j} = school percent Hispanic
W_{10j} = school percent limited English proficient

These variables were included in a second multiple regression equation as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W1j + \gamma_{02} W2j + \gamma_{03} W3j + \gamma_{04} W4j + \gamma_{05} W5j + \gamma_{06} W6j + \gamma_{07} W7j + \gamma_{08} W8j + \gamma_{09} W9j + \gamma_{010} W10j + u0j$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{1j} W1j + \gamma_{2j} W2j + \gamma_{3j} W3j + \gamma_{4j} W4j + \gamma_{5j} W5j + \gamma_{6j} W6j + \gamma_{7j} W7j + \gamma_{8j} W8j + \gamma_{9j} W9j + \gamma_{10j} W10j + u_{kj}$$

Classroom Effectiveness Index

To calculate the CEI for classroom *t* in school *j* with K_{tj} students, the following formula was used:

$$CEI_{tj} = \frac{\sum_{k=1}^{Ktj} \delta^2_{ktj}}{K_{tj}}$$

where *t* is the individual classroom, *j* is the school, and *k* is the number of students. The CEI is calculated with respect to the school district regression line. Using the slope and intercept from each school equation, a districtwide grand slope and a districtwide grand intercept was computed. The Classroom Effectiveness Index is a scaled score derived from the grand intercept, the grand slope, and the individual student's previous year's test score residual. The HLM residual that is aggregated and averaged to form the CEI is derived by:

$$Y_{cei} = (\text{grand intercept}) + (\text{grand slope}) X_{ij}$$

The CEI computation places each student on the same scale of measurement across the school district by controlling for school level influences. The CEI for teacher *t* in school *j* is then scaled to a mean of 50 and a standard deviation of 10 and adjusted as follows:

$$\text{Shrinkage Adjustment} = 1/1 + (\text{Variance}/n)$$

The shrinkage adjustment is utilized to adjust for differences in class sizes.

Research Questions

The investigation of the Level Two slopes and intercepts of schools over time were analyzed according to the following research questions and procedures. The first research question was: Do CEIs produce consistent results over time? The consistency of the grand intercept and the grand slope across the four years was analyzed for eighth-grade scores in TAAS reading and TAAS mathematics according to the following statistical hypothesis:

H_0: $GI_{97} = GI_{98} = GI_{99} = GI_{00}$

H_A: $GI_{97} \neq GI_{98} \neq GI_{99} \neq GI_{00}$

where GI_{97} is the district grand intercept for 1997 and GI_{00} is the district grand intercept for 2000. The grand intercept for TAAS mathematics and the grand intercept for TAAS reading were both analyzed.

If a grand intercept is consistent across years, student scores are entering the regression equations at similar points. This may indicate that the influence of school

level variables is consistent from year to year. The statistical hypothesis for the grand slopes was as follows:

H_0: $GS_{97} = GS_{98} = GS_{99} = GS_{00}$

H_A: $GS_{97} \neq GS_{98} \neq GS_{99} \neq GS_{00}$

where GS_{97} is the district grand slope for 1997 and GS_{00} is the district grand slope for 2000. The grand slope for TAAS mathematics and the grand slope for TAAS reading were both analyzed.

If the grand slopes are similar across years, the influence of the independent variables on the dependent variables would be similar. This implies that the students are being compared similarly across the school district over the years.

The investigation of model variance for teachers within schools over time lead to the second research question: How do outliers affect CEIs? It is important to determine if extreme individual student scores affect a teacher's CEI. An analysis for the influence of outliers on CEIs by teacher for each of the four years would indicate any significant level of influence. The statistical hypothesis was stated as follows:

H_0: $CEI_{tjoutlier} = CEI_{tjnooutlier}$

H_A: $CEI_{tjoutlier} \neq CEI_{tjnooutlier}$

This analysis was conducted for each teacher within each school. All teachers who had data were included in the analysis. Separate analyses were conducted on reading and mathematics scores. The identification of significant outliers in each set

of data was established using the Mahalanobis distance =15.51 for $p<.05$ with eight degrees of freedom and Cook's distance (CD> 1) diagnostic procedures.

Multiple regression is sensitive to extreme values (outliers) that are different from the rest of the values (Stevens, 1992). In fact, the influence of just a few outliers on the regression model estimate and goodness of fit statistics has been established (Ho & Naugher, 2000).

Outliers can also influence the findings in hierarchical models. The influence of extreme cases, or outliers, on Level One student residuals was therefore investigated in this study. According to Tabachnick and Fidell (1996), "the goal (of multiple regression) is that all of the cases contribute equally to the regression solution. However, cases that are far away from the others have more impact than the others on the size of regression coefficients" (p.133). By default, outliers also have an impact on the error term, or residual.

The three most important measures of the impact of outliers on the regression solution are measures of leverage, discrepancy, and influence (Tabachnick & Fidell, 1996). Outliers with leverage are cases far from the other cases in the equation, whether they are far from the others along the same regression line or not. Mahalanobis distance is a statistical measure used to detect the existence of outliers in standardized residuals. Discrepancy is "the extent to which an extreme case is in line with the other scores" (p. 134). The measure of the impact of discrepancy is often identified with a plot of the residuals. Influence is a combination of leverage and

discrepancy. A statistical measure of influence is Cook's distance to determine which outliers are influential.

Two diagnostic assessments on Level One student-level residuals were conducted to identify outliers. Using SPSS 10.1 for Windows Regression procedures, leverage diagnostics as determined by Mahalanobis distance were conducted by teachers within schools for each of the four years. The criterion for significance was: Mahalanobis distance = 15.51 for p = .05 with eight degrees of freedom. The equation for Mahalanobis distance is:

$$\text{Mahalanobis distance} = (N - 1)\,(h_i - 1/N)$$

where N is the number of cases and h_i is the leverage value for the ith case computed from $(\mathbf{X'X})^{-1} - \mathbf{X}'$ diagonal (Ho & Naugher, 2000).

Second, influence diagnostics as determined by Cook's distance were also conducted for student residuals aggregated by teacher for each of the four years. Cook's distance is "a measure of the change in the regression coefficient that would occur if this case was omitted, thus revealing which cases are most influential in affecting the regression equation" (Stevens, 1992, p.116). Cook's distance of CD>1 is considered large (Stevens, 1992). The equation for Cook's distance is:

$$CD = \frac{1}{(k+1)} r_i^2 \frac{h_{ii}}{1 - h_{ii}}$$

where r_i is the standardized residual and h_{ii} is the hat element. Cook's distance measures the combined influence of each case on *y* and the set of predictors. A pattern of significant outliers was investigated across teachers within schools.

The investigation of CEI stability across schools over time was analyzed by the third research question: Do teachers differ in average CEIs over time? The statistical hypothesis was:

H_0: $CEI_{tY1} = CEI_{tY2} = CEI_{tY3} = CEI_{tY4}$

H_A: $CEI_{tY1} \neq CEI_{tY2} \neq CEI_{tY3} \neq CEI_{tY4}$

where *t* is a teacher of mathematics or language arts and $_{Y1\text{-}Y4}$ are the school years of 1997-2000.

A second subset of teachers who had the same assignment within the same school over three consecutive years was identified. The statistical hypothesis was:

H_0: $CEI_{tY1} = CEI_{tY2} = CEI_{tY3}$

H_A: $CEI_{tY1} \neq CEI_{tY2} \neq CEI_{tY3}$

where *t* is a teacher of mathematics or language arts and $_{Y1\text{-}Y3}$ are either the school years of 1997-1999 or the school years of 1998-2000.

A repeated measures analysis over the four years and a repeated measures analysis over three years assessed any significant differences for each teacher of mathematics and reading within the 28 schools. Separate analyses were conducted for TAAS reading and TAAS mathematics CEIs averaged across sections for each teacher within the subset.

The fourth research question was: What is the correlation between the predicted CEI for a student and the students' Level One residual? The statistical hypothesis was stated as:

H_0: $Y_0 = Y_{CEI}$

H_A: $Y_0 \neq Y_{CEI}$

where Y_0 is a residualized current student achievement score and Y_{CEI} is the predicted student achievement score.

In the Dallas HLM model, Level One student residuals, the current student achievement residual, are entered into an equation with the district grand intercept and district grand slope in order to compute an HLM residual. The equation is:

$$Y_{cei} = (\text{grand intercept}) + (\text{grand slope})X_{ij}$$

The HLM residuals for students are then aggregated and averaged into a teacher's Classroom Effectiveness Index.

Pearson Product Moment Correlations between the Level One residual, $Y_{0,}$ and the HLM residual that has been adjusted by the district grand intercept and district grand slope, Y_{cei} , were generated for all student residuals included in the mathematics and reading sample. The reading and mathematics analyses were computed separately.

The four research questions proposed in this study are presented again in Chapter 4. Results from the data analyses question are presented in tables and text.

CHAPTER FOUR

RESULTS

This study was designed to identify statistical characteristics of the Dallas HLM model that created Classroom Effectiveness Indices (CEIs) as a measure of adjusted student growth by teacher. In order to answer the four research questions in this study, three databases were constructed. The first database consisted of school-level data for the eighth grade in 26 middle schools and two academies. Academies are designated schools for identified academically advanced students. The two academies are identified as separate campuses by the state and are listed in all school listings. The standardized residuals from the Dallas HLM model are aggregated and averaged by grade and subject area for each school to produce a school level regression intercept and school level regression slope for each middle school campus. This database included school intercepts and school slopes for the years of 1997 – 2000 and an assigned code number for each school. The study was confined to eighth grade middle school students' achievement on the Texas Assessment of Academic Skills (TAAS) reading assessment and the TAAS mathematics assessment. Analysis of the differences in each school's intercept and slope was included in the study. In addition, this initial database included a district grand intercept and grand slope in reading and a district grand slope and grand intercept in mathematics for each of the four years. Repeated measures analyses of the differences among the grand slopes and the differences among the grand intercepts are included in the study.

A second database of student-level information aggregated by teacher for mathematics and language arts courses was also constructed. Data included an assigned code number for each teacher, an assigned number for each student, teacher assignment indicated by course number and individual student achievement data in the form of two residuals: a Level One residual and a final HLM residual for each student.

Teachers identified for this database constituted two subsets. One subset was comprised of teachers who remained in the same teaching assignment and on the same school campus for four consecutive years. A second subset consisted of teachers who remained in the same teaching assignment and on the same school campus for three consecutive years. Teachers included in the three year subset either taught for the consecutive years of 1997-1999 or the consecutive years of 1998-2000 with all teachers in the database teaching for the years of 1998 and 1999. The individual student residuals were aggregated by teacher, by the content area of mathematics or reading, and by the year.

Students and teachers of Mathematics Course 2550, the district-designated eighth-grade mathematics course, were identified for the database. Also, teachers whose CEIs were computed with less than 10 student residuals were excluded. Investigation of the course listing database indicated data entry errors and special course numbers unique to campuses. These course numbers had very few students assigned, sometimes just one student, creating teacher CEIs with just one student residual. Because of the unknown nature of these course groupings, they were

excluded. Also, CEIs with fewer than 10 student residuals were excluded since prior research showed HLM results are unstable below this size (Snijders & Bosker, 1993). This study investigated a two-level model of students within schools to estimate optimal sample sizes at both levels to minimum standard errors (Snijders & Bosker, 1993). With an investigation of the standard errors of the regression coefficients, the study concluded that n>10 is the minimum for classroom level investigations.

Demographics

The number of teachers and students in both subsets for mathematics are included in Table 3.

Table 3

Teachers and Students by Year (Mathematics Course 2550)

Year	Teachers		Students
1997	Four year	25	1123
	Three year	13	589
	Total	38	1712
1998	Four year	25	1229
	Three year	32	1488
	Total	57	2717
1999	Four year	25	1140
	Three year	32	1437
	Total	57	2577
2000	Four year	25	990
	Three year	19	800
	Total	44	1790

In 1998 and 1999, a total of 57 teachers were included in the study for mathematics. The two series of three consecutive years teaching the same course at the same school (1997 – 1999 and 1998 – 2000) caused the total number of teachers in 1997 and 2000 to be smaller, with a total of 38 teachers included in the study in 1997 and a total of 44 teachers included in the study in 2000. However, a total of 7806 student mathematics residuals were used in the analyses, with the smallest number in 1997 (n=1712). This is a sufficient sample of both four year and three year teachers to establish an indication of differences over time.

The total number of student residuals assigned to each teacher also differed. This distribution of differences in the number of residuals creating a teacher's CEI gave a range within each year that realistically reflected class assignments at a middle school campus. Often teachers taught one course exclusively for all five periods, although most teachers taught a combination of seventh-grade and eighth-grade courses within their subject area. The smaller class size may indicate a teacher who is only teaching one section of the course. Also, as part of the Dallas HLM Model, students needed to have test scores from both the previous year and the current year and to have fewer than 20 days of absence in order to be included in the database.

The differences in student mathematics residuals included in a teacher's CEI across years are indicated in Table 4, showing the smallest and largest aggregations with each year. The teachers' CEIs and student residuals generated for this study are not reported; rather summary data are presented in tables.

Table 4

Mathematics Residuals Aggregated by Teacher

	Residuals	
	Smallest N	Largest N
Math 1997	14	88
Math 1998	14	82
Math 1999	12	100
Math 2000	12	86

Students and teachers of Language Arts Course 1100, the district designated eighth-grade course for reading and language arts were identified for the database. The filter of a more than 10 students per teacher was also implemented. Teachers designated as three year teachers either spanned the years of 1997 – 1999 or 1998 – 2000. Again, students needed to have test scores from both the previous year and the current year and to have fewer than 20 days of absence in order to be included in the database.

The number of schools, teachers, and students included in the study for TAAS reading is indicated in Table 5.

Table 5

Teachers and Students by Year (Language Arts Course 1100)

Year	Teachers		Students
1997	Four year	22	988
	Three year	10	515
	Total	32	1503
1998	Four year	22	1115
	Three year	21	1060
	Total	43	2175
1999	Four year	22	1232
	Three year	21	1194
	Total	43	2426
2000	Four year	22	1031
	Three year	10	437
	Total	33	1468

In 1998 and 1999, a total of 43 teachers were included in the study for reading. The two series of three consecutive years teaching the same course at the same school (1997 – 1999 and 1998 – 2000) caused the total number of teachers in 1997 and 2000 to be smaller, with a total of 32 teachers included in the study in 1997 and a total of 33 teachers included in the study in 2000. However, a total of 7572 student reading residuals were used in the analyses, with the smallest number in 2000 (n=1468). This is a sufficient sample of both four year and three year teachers to establish indication of differences over time.

The differences in student reading residuals included in a teacher's CEI across years are indicated in Table 6, showing the smallest and largest aggregations with each year.

Table 6

Reading Residuals Aggregated by Teacher

	Residuals	
	Smallest N	Largest N
Read 1997	27	79
Read 1998	16	84
Read 1999	17	99
Read 2000	16	85

This chapter presents the results of the analysis for each research question proposed in the study using the two constructed databases.

Consistency Over Time

Data were analyzed indicating results over time in order to answer the first research question: Do Classroom Effectiveness Indices produce consistent results over time?

The Dallas HLM model uses the individual student residuals included in the Classroom Effectiveness Indices to produce individual school intercepts and slopes. The school level intercepts and slopes use all student residuals assigned to each

campus that remain in the analysis following the estimation of the parameters. All student residuals assigned to each campus have been included in this analysis. The statistical properties of the intercept are mean = 100, standard deviation = 1. Because of the compact nature of the distribution, the intercepts are reported to the fourth decimal place.

The individual intercepts and slopes for the 1997 – 2000 school years for the 28 middle schools and vanguards are reported in Tables 7-10. A summary of the descriptive statistics for intercepts and slopes follow.

Table 7
School-level HLM Mathematics Intercepts

School	Intercept_math97	Intercept_math98	Intercept_math99	Intercept_math00
1	100.1270	99.9678	100.0414	100.1223
2	100.0211	100.0616	100.0766	100.0798
3	99.9731	100.1088	100.0207	99.9598
4	100.0616	100.0606	100.0966	100.0262
5	100.0141	100.0580	100.0711	99.9996
6	100.0416	100.0135	100.0255	100.0876
7	100.1129	100.0012	100.0626	100.0579
8	100.1259	100.1787	100.0398	100.0734
9	100.0579	100.0026	100.0441	100.1517
10	99.9640	100.1396	100.1083	100.1236
11	100.0456	100.0723	99.9926	100.0490
12	100.0619	100.0721	100.0501	100.1263
13	100.1244	99.9960	100.1316	100.0819
14	100.0844	100.0935	100.0543	100.0118
15	100.0790	100.1364	100.1090	100.1092
16	100.0042	100.0754	99.9272	100.0172
17	100.0073	100.0397	100.0540	100.0011
18	100.0049	99.9583	99.9886	100.0238
19	100.0137	100.1356	100.1240	100.0954
20	100.0809	100.1528	100.2383	100.3200
21	100.0231	100.0176	100.2253	100.1099
22	100.0669	100.0668	100.2272	100.3697
23	100.0368	100.0661	100.1215	100.0390
24	100.0985	100.1396	100.1427	100.2590
25	100.0819	100.0916	100.1853	100.1885
26	100.0455	100.1341	100.1683	100.3280
27	99.9666	100.1933	100.1136	100.1529
28	100.1592	100.0132	100.2332	100.3685

Table 8
School-Level HLM Mathematics Slopes

School	Slope_math97	Slope_math98	Slope_math99	Slope_math00
1	0.7500	0.6776	0.7574	0.7782
2	0.8113	0.6896	0.7064	0.8099
3	0.7975	0.3681	0.7288	0.7498
4	0.7937	0.7966	0.7613	0.7752
5	0.7889	0.6104	0.7089	0.7503
6	0.7107	0.6778	0.715	0.8026
7	0.7913	0.7423	0.6753	0.7094
8	0.7167	0.5628	0.7166	0.5985
9	0.7753	0.6276	0.629	0.687
10	0.8059	0.733	0.7206	0.6402
11	0.8554	0.7028	0.8178	0.6931
12	0.7154	0.7046	0.7381	0.6896
13	0.6928	0.6000	0.6680	0.7466
14	0.8903	0.6976	0.7178	0.5553
15	0.8003	0.6349	0.6625	0.4389
16	0.7485	0.6761	0.8335	0.6473
17	0.8198	0.8027	0.7716	0.7438
18	0.8108	0.7991	0.8536	0.8188
19	0.8281	0.8469	0.7843	0.8002
20	0.7734	0.6565	0.4977	0.4487
21	0.8319	0.6451	0.635	0.7075
22	0.7106	0.7611	0.5564	0.3652
23	0.8226	0.7090	0.7459	0.8238
24	0.7438	0.7522	0.5855	0.4575
25	0.7546	0.8672	0.7325	0.7329
26	0.7076	0.8195	0.6311	0.4927
27	0.8061	0.6573	0.6894	0.5717
28	0.6522	0.7747	0.5391	0.5388

Table 9
School-level HLM Reading Intercepts

School	Intercept_read97	Intercept_read98	Intercept_read99	Intercept_read00
1	100.2118	100.1589	100.0514	100.1254
2	100.0307	100.0499	100.0262	100.0519
3	100.0309	100.0752	100.0291	100.0706
4	100.0314	99.9991	100.0673	99.9816
5	100.0261	100.0303	100.0952	100.0296
6	100.0736	99.9813	99.9746	100.1225
7	100.0311	99.9983	99.9833	100.0480
8	100.0698	100.212	100.0384	100.1248
9	100.0960	100.0518	100.0200	100.1354
10	100.0325	100.0708	100.0598	100.1014
11	99.9584	100.0053	100.0125	100.0633
12	100.0792	100.0305	100.0319	100.1023
13	100.1763	100.074	100.1349	100.0961
14	99.9377	100.0128	100.0323	100.0519
15	100.0366	100.0633	100.0633	100.1243
16	100.0231	100.0328	99.9965	100.0085
17	99.9372	99.8769	100.0797	99.9544
18	100.0093	99.9527	99.9476	99.9806
19	99.9740	100.0054	99.9257	100.0178
20	100.2121	100.2096	100.3055	100.2617
21	100.0955	100.0274	100.1656	100.1224
22	100.2023	100.2105	100.2048	100.2179
23	100.0178	100.0638	100.0256	100.0043
24	100.2054	100.2365	100.1464	100.1821
25	99.9824	99.9949	100.1363	100.0324
26	100.2216	100.215	100.1720	100.2411
27	100.1254	100.1576	100.0900	100.1284
28	100.2820	100.2279	100.1908	100.2365

Table 10
School-level HLM Reading Slopes

School	Slope_read97	Slope_read98	Slope_read99	Slope_read00
1	0.6808	0.6414	0.6648	0.8495
2	0.7192	0.6424	0.6716	0.6663
3	0.6629	0.6506	0.651	0.4983
4	0.6162	0.6827	0.6709	0.6100
5	0.6639	0.5836	0.6607	0.6078
6	0.6783	0.8558	0.5968	0.7559
7	0.4781	0.6509	0.6657	0.5811
8	0.6796	0.6176	0.6616	0.5332
9	0.6076	0.6416	0.6316	0.5812
10	0.6997	0.6501	0.6512	0.6472
11	0.7651	0.6723	0.6982	0.6159
12	0.5715	0.8145	0.5884	0.6546
13	0.4748	0.7007	0.5976	0.6232
14	0.4267	0.6688	0.6484	0.4410
15	0.3706	0.6514	0.6343	0.3663
16	0.6728	0.5866	0.7167	0.6149
17	0.6702	0.7595	0.6226	0.6568
18	0.7233	0.6353	0.7584	0.7728
19	0.7132	0.7596	0.7043	0.6814
20	0.6968	0.6456	0.4847	0.5937
21	0.6940	0.6788	0.6114	0.4852
22	0.6645	0.6288	0.5857	0.5878
23	0.6367	0.6059	0.7057	0.6662
24	0.6631	0.5778	0.5947	0.6639
25	0.5650	0.7623	0.6069	0.6208
26	0.7130	0.5803	0.5857	0.6932
27	0.6968	0.5871	0.6435	0.7013
28	0.6284	0.6985	0.5667	0.6377

Summary statistics were computed to indicate the level of differences in the school level intercepts and slopes. Table 11 indicates the range, minimum and maximum of the individual school intercepts for TAAS mathematics.

Table 11

Mathematics Intercepts: Range, Minimum and Maximum

Intercept[a]	Range	Minimum	Maximum
Math97	.1950	99.9640	100.1590
Math98	.2350	99.9583	100.1933
Math99	.3111	99.9272	100.2383
Math00	.4099	99.9598	100.3697

[a]n = 28 for each year

The range in the individual school mathematics intercepts does not vary greatly across schools within a given year, although the range increases from year to year, from .1950 in 1997 to .4099 in 2000. The increase is indicated in the maximum intercept rather than the minimum intercept each year except 1999, where the minimum intercept drops to 99.9272. The range of intercepts does not vary greatly across the four years, ranging from 99.9272 to 100.3697. This is a difference of only .2149.

Table 12 presents a summary of four years of individual mathematics slopes.

Table 12

Mathematics Slopes: Range, Minimum and Maximum

Slope[a]	Range	Minimum	Maximum
Math97	.2400	.6500	.8900
Math98	.4991	.3681	.8672
Math99	.3559	.4977	.8536
Math00	.4586	.3652	.8238

[a]n = 28 for each year

The slopes among individual schools in mathematics within a given year indicate larger differences, especially in the years 1998, with .4991, and 2000, with .4586. There is no steady increase in difference indicated across the four years. However, the range of the mathematics slopes across years indicate large differences, from .2400 in 1997 to .4991 in 1998, a .2591 difference.

Summary statistics for reading were computed to indicate the level of differences in the school level intercepts and slopes. The range, minimum and maximum of the individual school intercepts and slopes for TAAS reading are recorded in Table 13 and Table 14. Differences in the intercepts and slopes for reading and mathematics are also assessed in order to identify any subject-related characteristics. There may be indications that the range of the intercepts and slopes is different in reading than in mathematics.

Table 13

Reading Intercepts: Range, Minimum and Maximum

Intercept[a]	Range	Minimum	Maximum
Read97	.3448	99.9372	100.2820
Read98	.3596	99.8769	100.2365
Read99	.3798	99.9257	100.3055
Read00	.3073	99.9544	100.2617

[a]n = 28 for each year

Both the minimum and maximum intercepts in reading changed little from year to year. The individual school intercepts remained consistent in reading across schools and years, with a difference of .0725 across the four years. This difference compares to a difference of .2149 in the mathematics intercepts across the four years. While neither difference is large, the reading intercepts cluster more tightly.

Table 14

Reading Slopes: Range, Minimum and Maximum

Slope[a]	Range	Minimum	Maximum
Read97	.3945	.3706	.7651
Read98	.2780	.5778	.8558
Read99	.2737	.4847	.7584
Read00	.4832	.3663	.8495

[a]n = 28 for each year

The individual school slopes for reading indicated a range similar to the individual school slopes for mathematics; reading slopes indicate differences of .2737 – 4832. and mathematics slopes indicate differences of .2400 - .4991. Again, the minimum slope differs markedly from 1997 to 1998, with a difference of .2072. These fluctuating slopes in both reading and mathematics would indicate that although schools are entering the regression equation similarly as indicated by the consistent intercepts, there are differences in the degree of growth scores, student residuals, across schools. The influence of the independent variables, student and school characteristics, on the dependent variables, student achievement, may differ across schools. The influence of outlier student residuals was also possible, and this possibility was investigated in this study. Further analysis will identify the statistical importance of outliers on these differences in intercepts and slopes.

Additional computations calculated a districtwide grand intercept and slope in reading and a grand intercept and slope in mathematics for each of the four years. The grand intercept and slope for each year are simply the districtwide averages of all 28 middle schools and vanguards. The grand intercepts and grand slopes are used to compute each final HLM student residual. In this manner, all student growth is assessed in relation to the district average as indicated by the following:

$$Y_{cei} = (\text{grand intercept}) + (\text{grand slope})X_{ij}$$

Where Y_{cei} is each student's final residualized growth score and X_{ij} is each student residual from Level One regression equation.

Table 15 summarizes grand slopes and intercepts over the four years in TAAS Mathematics and Table 16 summarizes grand slopes and intercepts over four years in TAAS reading.

Table 15

TAAS Mathematics Grand Intercepts and Grand Slopes

	Grand Intercept	Grand Slope
1997	100.0530	0.7752
1998	99.9905	0.7262
1999	100.1373	0.6483
2000	100.2454	0.6585

Table 16

TAAS Reading Grand Intercepts and Grand Slopes

	Grand Intercept	Grand Slope
1997	100.0754	0.6369
1998	99.9905	0.7262
1999	100.1373	0.6483
2000	100.1810	0.7436

The grand intercepts and grand slopes were analyzed using the Repeated Measures procedures of the SPSS 10.1 for Windows program. The repeated measures analysis results are presented in Table 17 and Table 18, rounded to three decimal places.

Table 17

Mathematics Grand Intercept Repeated Measures Analysis

Source	SS	df	MS	F	p
School	.340	27	.012		
Year	.068	3	.022	5.50**	.003
Error	.360	81	.004		
Total	.768	111			

**p< .01

Table 18

Reading Grand Intercept Repeated Measures Analysis

Source	SS	df	MS	F	p
School	.700	27	.025		
Year	.008	3	.003	1.50	.229
Error	.160	81	.002		
Total	.868	111			

In repeated measures analyses, lack of statistical significance indicates no difference over time. If the intercepts across four years differ little, the analysis will not be statistically significant. In this study, the repeated measures analysis of the mathematics grand intercepts was F = 5.50, statistically significant at the .01 level. The repeated measures analysis of the reading grand intercept was not statistically significant. The intercepts in mathematics differed across the four years and the intercepts in reading did not. This means that in relation to the intercepts, or at what point students are entering the districtwide equation, the TAAS reading assessment produced consistent results across the four years and the TAAS mathematics assessment did not.

The repeated measures analyses that indicated consistency over time for the mathematics and reading grand slopes are presented in Table 19 and Table 20.

Table 19

Mathematics Grand Slope Repeated Measures Analysis

Source	SS	df	MS	F	p
School	.420	27	.015		
Year	.190	3	.063	9.00**	.000
Error	.610	81	.007		
Total	1.22	111			

**p< .01

Table 20

Reading Grand Slope Repeated Measures Analysis

Source	SS	df	MS	F	p
School	.260	27	.009		
Year	.027	3	.009	1.80	.192
Error	.460	81	.005		
Total	.747	111			

The repeated measures analysis of the mathematics grand slope yielded F = 9.00, statistically significant at the .01 level. The repeated measures analysis of the reading grand slope was not statistically significant with F = 1.80. The slopes in mathematics differed significantly across the four years, but the slopes in reading did not. The TAAS reading assessment produced consistent results across the four years while the TAAS mathematics assessment did not.

The analyses indicated consistent results over time for reading but not for mathematics. Each student was assessed against the district grand slope and intercept; if there were differences in both slope and intercept across the four years included in this study then students would have been measured against a different standard from year to year in mathematics and against a similar standard from year to year in reading.

Outlier Effects

Data were analyzed to assess outlier effects over the four years in the study in order to answer the second research question: How do outliers affect CEIs? A Level One residual was computed for each student using the initial multiple regression equation. Any outliers in these initial regression residuals would influence the final outcome computed as the Classroom Effectiveness Index for a teacher. It is important to identify the degree and dimension of outlier influence at the initial stage. Therefore, using student residuals aggregated to teacher ID, two diagnostic assessments of Level One student residuals were conducted to identify outliers. Using SPSS 10.1 for Windows Regression procedures, leverage diagnostics indicated by Mahalonobis distance were computed for each student residual across the four years in reading and across the four years in mathematics. In this study, outliers indicated by Mahalonobis distance would be statistically significant at 15.51 for $p < .05$ with eight degrees of freedom. A second influence diagnostic as determined by Cook's distance was also computed for each student residual across the four years in both reading and mathematics. An outlier is statistically significant using Cook's distance when it is indicated as $Cd > 1$. Cooks distance is a measure of "the change in the regression coefficient that would occur if this case was omitted" (Stevens, 1992, p.116). Therefore, it is possible to have an outlier that is statistically significant as indicated by Mahalonobis distance that indicates the extent to which a score is not in alignment with all other scores in the regression, and not be statistically significant as indicated by Cook's distance, which measures the degree of change in the coefficient

after removal of the outlier. A listing of the assessment of all Level One student residuals is included in the Appendix. The total number of student residuals across the four years was 8796.

In the study, 11 statistically significant outliers were identified in the student Level One residuals in mathematics as indicated by Mahalonobis distance, and eight statistically significant outliers were identified in reading as indicated by Mahalonobis distance. No outliers were identified as statistically significant using Cook's distance.

A summary of significant mathematics outliers as identified by these diagnostic assessments is included in Table 21. Also included in the table are the corresponding Level One residual and the teacher's CEI computed with the outlier, the teacher's CEI computed without the outlier, and the difference between the two CEIs. By removing the outlier from a second computation, the influence of that outlier in the final Classroom Effectiveness Index would be evident. Statistical properties of Classroom Effectiveness Indices are a mean of 50 and a standard deviation of 10.

Table 21

Level One Residual Outliers in Mathematics

	Level One Residual	Mahalonobis Distance	CEI with Outlier	CEI without Outlier	Difference[1]
M_1	85.60	191.75	39.94	48.19	+ 8.25
M_2	93.15	48.46	41.20	43.77	+ 2.57
M_3	113.71	170.92	60.03	52.28	- 7.75
M_4	113.57	167.53	54.38	45.88	- 8.50
M_5	107.33	50.46	45.65	49.51	+ 3.86
M_6	106.05	34.84	51.33	50.40	- .93
M_7	94.58	23.33	54.22	52.10	- 2.12
M_8	118.43	278.27	55.60	45.24	- 10.36
M_9	109.68	78.22	51.08	42.68	- 8.40
M_10	109.28	71.98	65.97	59.71	- 6.26
M_11	92.14	47.18	55.51	53.44	- 2.07

[1]The difference is interpreted against a criterion of CEI = 50

Only one outlier was identified by the Mahalonobis distance statistic in the 1997 student residual database, M_1, and one outlier was identified in the 1998 student residual database, M-2. Five outliers were identified in the 1999 student residual database, M_3 to M_7. The 1999 outlier M distances ranged from 23.33 to

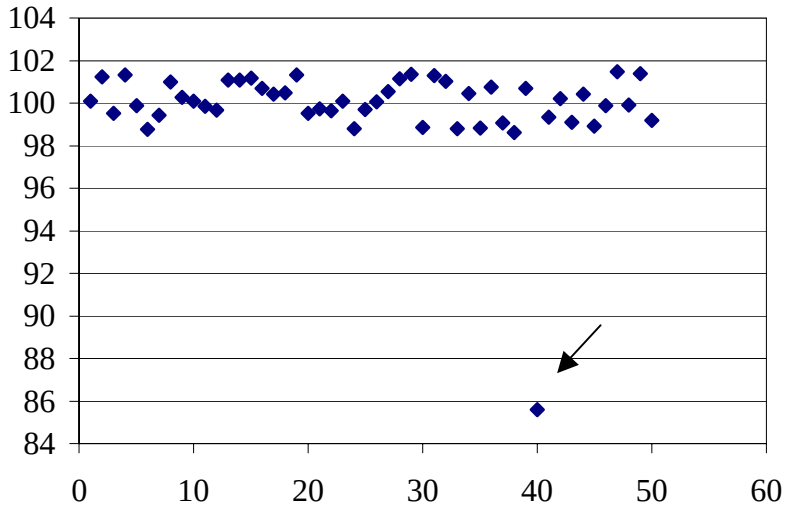
170.92. Four outliers were detected by the Mahalonobis distance statistic in the 2000 database, M_8 to M_11, ranging from 47.18 to 278.28. Cooks distance did not detect any outliers in the mathematics student residuals that were statistically significant (CD>1.0), with the highest Cooks distance statistic of .32. The CD = .32 corresponded to the M_11 outlier of 278.28. The outliers occurred for teachers in eight middle schools, with one teacher included as identified for two years, 1997 and 1999. No teacher had more than one outlier in his or her course aggregation of student residuals in one year. Therefore, the outliers were not confined to the same schools or teachers across the four years of the study.

Once the outliers were identified, they were inactivated in the databases using a code flag, and the Classroom Effectiveness Indices for all teachers were recomputed without the outliers. The recomputed CEIs will assist in identifying the influence each outlier had on the original CEI.

The influence of a single outlier on an identified teacher‘s CEI varied greatly, ranging from a difference of .93 to a difference of 10.36. Only three of the CEIs improved with the removal of the outlier, while eight CEIs decreased. Remembering that the district average for a CEI is 50, the change in CEI with the removal of one outlier may be of practical significance to a teacher. If the CEI changes enough with the removal of outliers, the perception of how effective that teacher is may also change. The degree of the outlier in three of the cases is evident in Figures 1 – 3.

Figure 1

Mathematics Residual Plot With Outlier (Teacher M_1)



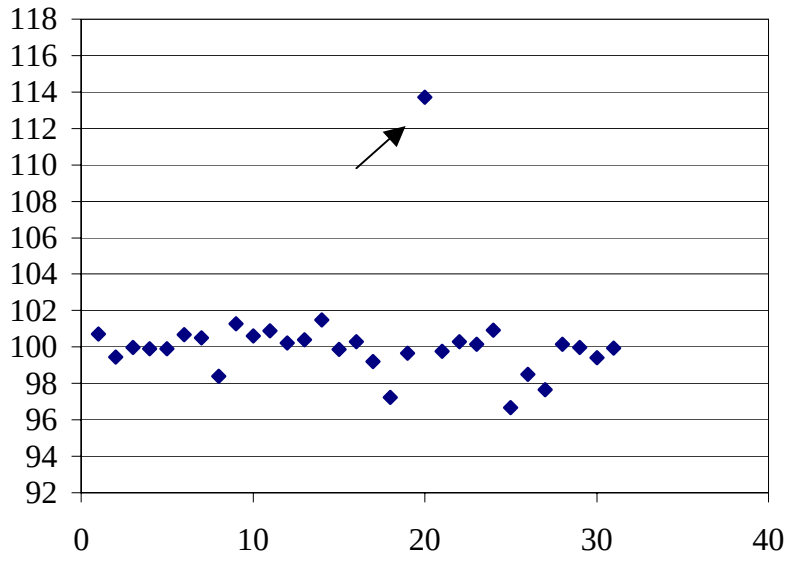
The teacher (M_1) CEI improved from a 39.94 to a 48.19 with the removal of the identified outlier. This teacher changed from being perceived as a "teacher in need of assistance," because his or her CEI was more than one standard deviation away from the district average to a teacher whose overall student growth was similar to the district average.

The removal of an outlier above the distribution has a similar influence with differing consequences. For example, another teacher (M_8) moved an entire standard deviation, from above the district average at 55.60 to below the district average at 45.65. She or he would be seen as an "above average" teacher with the outlier and a "below average" teacher without the outlier. Another example that indicates the extreme influence of an outlier above the distribution is demonstrated in

Figure 2. The teacher's CEI was 60.03 with the outlier included in the computation and a 52.28 without the outlier.

Figure 2

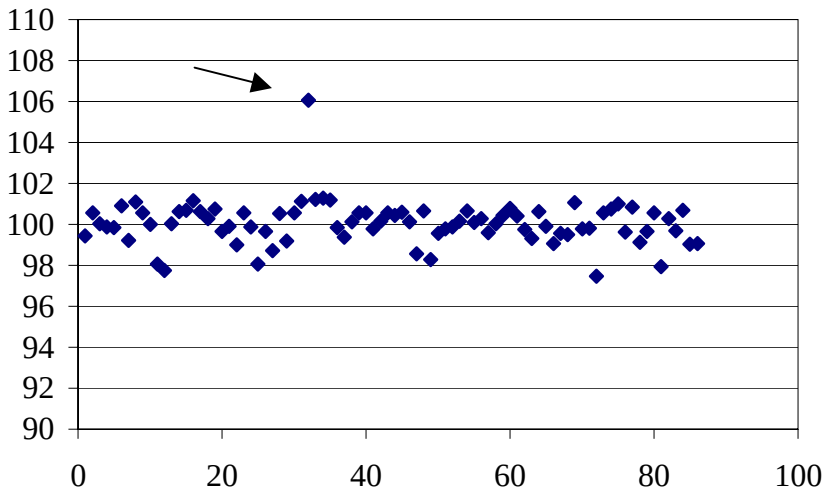
Mathematics Residual Plot With Outlier (Teacher M_3)



Four teacher's Mathematics CEIs did not change in practical significance with the removal of the outlier, since changes in the CEIs were less than 3 points. Each had a statistically significant outlier as indicated by the Mahanobis distance analysis, yet the change with the removal of the outliers was not of practical consequence. Figure 3 plots student residuals for a representative teacher. It can be concluded that the influence of outliers on the mathematics teachers' final CEIs was not consistent across all cases.

Figure 3

Mathematics Residual Plot With Outlier (Teacher M_6)



A summary of significant reading outliers as identified by these diagnostic assessments is included in Table 22. Included in the table are the corresponding Level One residual and the teacher's CEI computed with the outlier, the teacher's CEI computed without the outlier, and the difference between the two CEIs.

Table 22

Level One Residual Outliers in Reading

			CEI		
	Level One Residual	Mahalonobis Distance	Outlier	Without Outlier	Difference[1]
R_1	84.53	213.26	40.94	47.98`	+ 7.04
R_2	113.02	153.52	55.24	48.76	- 6.48
R_3	95.19	20.34	40.92	41.41	+ .49
R_4	95.54	19.43	49.23	49.85	+ .62
R_5	95.54	19.42	45.17	45.65	+ .48
R_6	95.74	17.73	47.85	48.15	- .30
R_7	111.12	125.42	53.62	51.74	- 1.88
R_8	106.26	39.94	53.87	51.52	- 2.35

[1]The difference is interpreted against a criterion of CEI = 50

No outliers were identified by the Mahalonobis distance statistic in the 1997 student residual database for reading. Three outliers were identified in the 1998 student residual database, R_1 – R_3, three outliers were identified in the 1999 student residual database, R_4 – R_6, and two outliers were identified in the 2000 database, R_7 and R_8. Cooks distance did not detect any outliers in the reading student residuals that were statistically significant (CD>1.0), with the highest Cooks distance statistic of .56. The CD = .56 corresponded to the R_1 outlier, which was the largest outlier to be detected in reading (213.26). The outliers occurred for

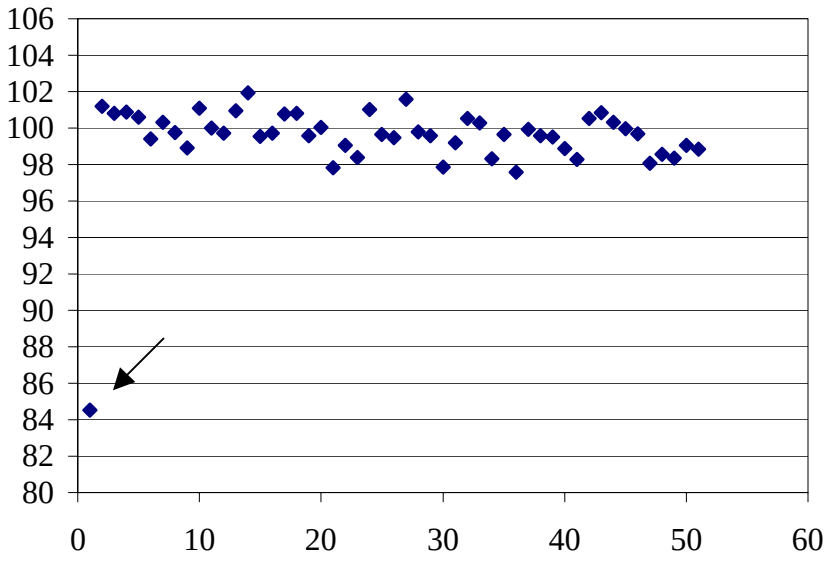
teachers in eight middle schools, with no teacher included for more than one year. No school was repeatedly identified. Finally, no teacher had more than one outlier in his or her course aggregation in one year. Therefore, the outliers were not confined to the same schools or teachers across the four years of the study.

Once the outliers were identified, they were inactivated in the databases using a code flag, and the Classroom Effectiveness Index for all teachers was recomputed without the outliers in order to identify the influence that outlier would have on the final teacher's CEI. The influence of a single outlier on an identified teacher's CEI did not vary as much as the outlier influence in mathematics. The outlier influence in reading ranged from a difference of .30 to 7.04.

Five CEIs improved with the removal of the outliers, and three CEIs were computed to a lower CEI without the outliers. However, unlike the differences in mathematics, six of the eight recomputed CEIs had differences less than three points. This small change is not a significant practical difference to a teacher in relation to the district average. However, the teacher whose CEI improved almost seven points moved from one standard deviation below the district mean to within two points of the district mean as indicated in Figure 4.

Figure 4

Reading Residual Plot With Outlier (Teacher R_1)

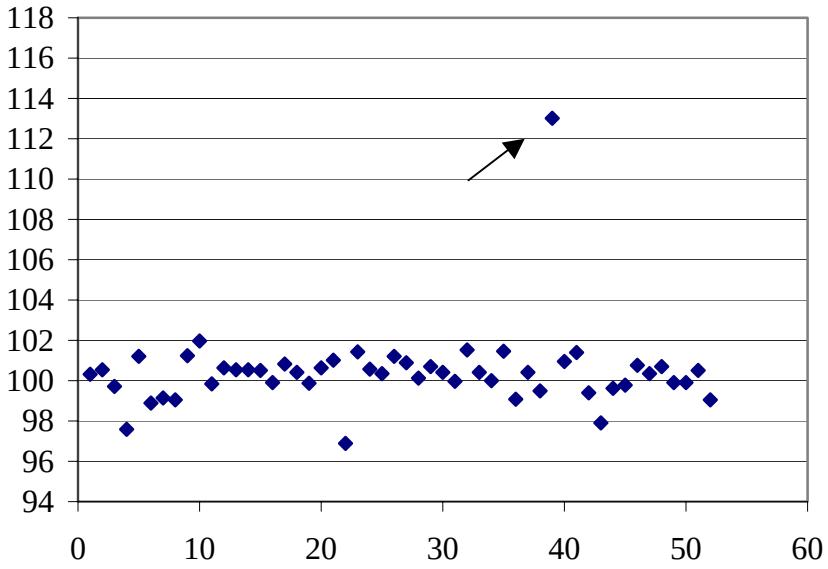


With the removal of the outlier indicated in Figure 4, the CEI for teacher R_1 changed from 40.94 to 47.98. The teacher went from being almost one standard deviation away from the district mean to within two points of the average.

Figure 5 indicates the student residual plot for a teacher with an “high score” outlier.

Figure 5

Reading Residual Plot With Outlier (Teacher R_2)

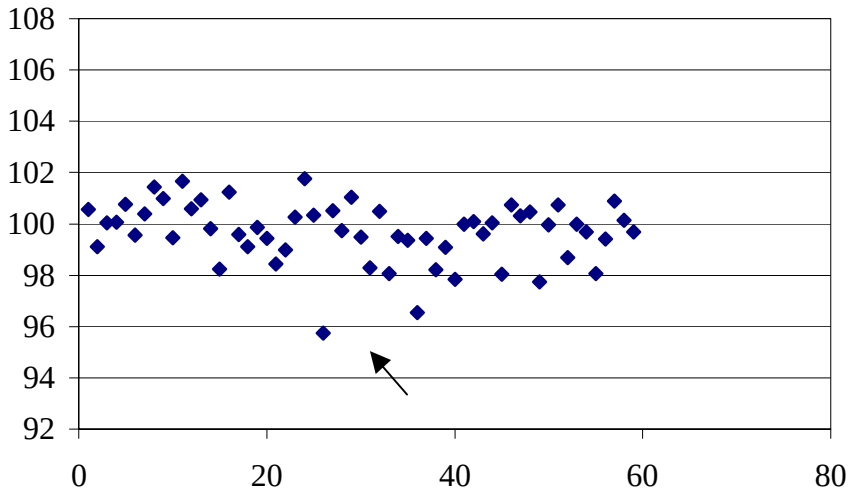


Teacher R_2 moved from a CEI of 55.24 with the outlier to a CEI of 48.76. Again, the perception of the teacher’s influence on his or her students’ achievement may change with the lower CEI.

Figure 6 indicates the student residuals for a teacher for whom a statistically significant outlier was indicated by the Mahalonobis distance but did not have an appreciably changed CEI. Teacher R_6 had a CEI of 47.85 with the outlier included and a CEI of 48.15 without the outlier.

Figure 6

Reading Residual Plot With Outlier (Teacher R_6)



Data indicated that outlier effects on an individual teacher's CEI can be statistically significant and yet have differing practical consequences. In this study, only one outlier was detected in the aggregated student residuals for each of the 19 teachers. Nevertheless, it has been established that in three of these, the practical consequences indicated the need to identify and eliminate extreme outliers, although the number of outliers identified was small in any one year. Most teachers included in the study did not have their CEIs changed because no outliers were detected, and ten of the 19 teachers with identified outliers did not have any potential practical consequences attached to changes in the CEIs because the change was so small.

Teacher Differences

Data were analyzed indicating results over time in order to answer the third research question: Do teachers differ in average CEIs over time? A subset of

teachers who remained in the same school and had the same teaching assignment over the four consecutive years and another subset of teachers who remained in the same school and had the same teaching assignment for three consecutive years were identified. The teachers designated as three year teachers taught in the same course in the same school for the years of 1997 – 1999 or 1998 – 2000. The average CEI for each teacher is an average of all student residuals in all sections of Course 1100 (eighth grade reading) or all sections of Course 2550 (eighth grade mathematics) assigned to a teacher. Tables of Mathematics CEIs by teacher by year for the three and four year teachers and tables of Reading CEIs by teacher by year for the three and four year teachers are included in the Appendix. The range, minimum and maximum CEI for each year in math are summarized in Table 23.

Table 23

TAAS Mathematics: Range, Minimum and Maximum CEI

	Subset	Range	Minimum	Maximum
1997	4 year	30.43	32.29	62.74
	3 year	28.75	40.64	69.40
1998	4 year	23.14	38.15	61.30
	3 year	25.77	37.68	63.45
1999	4 year	23.50	36.18	59.68
	3 year	25.99	39.67	65.66
2000	4 year	22.36	37.70	60.06
	3 year	32.51	33.46	65.97

CEIs for 2000 teachers who taught the 2550 mathematics course for all four years of the study had the smallest range of 22.36, just over two standard deviations. The range of CEIs for teachers who taught the 2550 mathematics course for three consecutive years in 2000 was the largest at over three standard deviations. There was a difference of approximately ten points between the highest and lowest maximum CEI, and approximately seven points between the highest and lowest minimum CEI.

The range, minimum and maximum CEI for each year in reading are summarized in Table 24.

Table 24

TAAS Reading: Range, Minimum and Maximum CEI

	Subset	Range	Minimum	Maximum
1997	4 year	28.88	36.50	65.37
	3 year	21.50	37.57	59.07
1998	4 year	16.59	39.70	56.29
	3 year	20.07	40.37	60.44
1999	4 year	16.26	36.38	52.26
	3 year	17.64	37.04	54.68
2000	4 year	14.67	41.71	56.38
	3 year	7.38	49.42	56.81

CEIs for 2000 teachers who taught the 1100 language arts course for three years had the smallest range of 7.38, less than one standard deviation. The range of 1997 CEIs for teachers who taught the 1100 language arts course for the four years of the study was the largest at 28.88. Overall, the ranges for reading were smaller than the ranges for mathematics. There was a difference of approximately thirteen points between the highest and lowest maximum CEI, and between the highest and lowest minimum CEI.

Analyses using SPSS 10.1 for Windows Repeated Measures procedures for the three year teachers and four year teachers produced the results in Tables 25 and Table 26. The three year teacher database was collapsed into CEIs by year one, year two, and year three, and all teachers were included in the same repeated measures analysis. Previously, it was necessary in the research design for the CEIs to remain within the calendar year. It was not necessary for this analysis that we know the year of the CEI. Teaching the same course in the same school over three years was the criterion.

Table 25

Repeated Measures Analysis of Four Year Mathematics CEIs

	SS	df	MS	F	p
Teacher	1548.58	24	64.52		
Year	38.84	3	12.95	.44	.73
Error	2137.14	72	29.68		
Total	3724.56	99			

Table 26

Repeated Measures Analysis of Three Year Mathematics CEIs

Source	SS	df	MS	F	p
Teacher	2983.31	31	96.24		
Year	85.42	2	42.71	1.79	.18
Error	1478.50	62	23.85		
Total	4547.23	95			

No difference was found for mathematics teachers across three and four years of CEIs. The repeated measures analyses indicated that teacher CEIs are consistent across consecutive years.

The repeated measures analyses for four year reading teachers and three year reading teachers are presented in Tables 27-28.

Table 27

Repeated Measures Analysis of Four Year Reading CEIs

Source	SS	df	MS	F	p
Teacher	945.60	21	45.03		
Year	43.53	3	14.51	.52	.67
Error	1743.05	63	27.67		
Total	2732.18	87			

Table 28

Repeated Measures Analysis of Three Year Reading CEIs

Source	SS	df	MS	F	p
Teacher	608.93	19	32.05		
Year	15.53	2	7.77	.40	.68
Error	742.67	38	19.54		
Total	1367.13	59			

No difference was found for reading teachers across three and four years of CEIs. The repeated measures analyses indicated that the average CEIs for teachers in the study are consistent across consecutive years in both reading and mathematics.

Residual Correlations

Data were analyzed indicating the relationship between the Level One student residual and the HLM student residual in order to answer the fourth research question: What is the correlation between the predicted CEI and the Level One residual? The HLM model adjusts the initial student residuals for school level variables. A strong positive correlation between the Level One residuals and the final student-level HLM residual that are aggregated to compute a teacher's CEI would indicate similar outcomes from the first and second stage of the Classroom Effectiveness Indices; as the value of the Level One residual increases, the value of the HLM residual would similarly increase.

The Pearson correlations between the Level One residual, $Y_{O,}$ and the final student residual, $Y_{cei,}$ that has been adjusted by the district intercept and slope was reported for all students in mathematics and all students in reading for the two subsets of teachers identified as three year and four year teachers. Table 29 includes summaries of the number of student residuals and the correlations for TAAS mathematics for each year.

Table 29

Correlations for HLM and Level One Mathematics Residuals by Year

Year	*r*	N
1997	.71	1715
1998	.69	2717
1999	.74	2577
2000	.76	1790

Note: $r = .25$, df = 100, $p<.01$

All four correlations between the Level One residuals and the HLM residuals for mathematics are positively correlated. The correlations exceeded the critical tabled value of $r = .25$. Therefore, all four correlations are statistically significant. The correlations ranged from .69 to .76. The correlations indicate that the initial residualized growth score from the multiple regression process were similar to the final student residual. Additionally, the significant correlations indicated that the school level variables within the second stage of the model do not change with the initial student outcome.

Summaries of the number of student residuals and the correlations for TAAS reading for each year are in Table 30.

Table 30

Correlations for HLM and Level One Reading Residuals by Year

Year	*r*	N
1997	.75	1503
1998	.77	2187
1999	.75	2427
2000	.73	1487

Note: $r = .25$, df = 100, $p<.01$

All four correlations between the Level One residuals and the HLM residuals for reading are positively correlated. The correlations all exceeded the critical tabled value of $r = .25$. Therefore, all correlations between the HLM and Level One residuals in reading were statistically significant for each of the four years. The correlations ranged from .73 to .77.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

Utility of A Classroom Effectiveness Index

The definition of Utility as used in this study was "the level of consistency and integrity sufficient for using the model to identify effective and ineffective teachers." The four questions investigated in the study concerning the statistical properties of the Classroom Effectiveness Index sought to identify whether the Dallas HLM Model had a sufficient level of consistency and statistical integrity for the CEIs to be used for teacher accountability.

The first question addressed whether or not the final HLM student residuals were consistently computed over the four years of the study. This analysis sought to determine if students were being measured by similar standards from year to year. In order to determine this consistent standard across time, the grand intercepts and grand slopes for reading and mathematics were analyzed, since each student's final score was computed from the individual student's residualized score and the grand intercept and grand slope developed for the district. The differences identified by the analyses were confined to the mathematics intercept and slope.

The intercepts in mathematics across the four years were statistically significantly different while the intercepts in reading were not. This means that in relation to the intercepts, which represent the point at which students are entering the districtwide regression equation, the TAAS reading assessment produced consistent results across the four years, while the TAAS mathematics assessment did not.

The slopes in mathematics were also statistically significantly different across the four years while the slopes in reading were not. This indicated that the impact of independent variables on the dependent variables changed from one year to the next in mathematics but did not change significantly in reading. It was determined that the Classroom Effectiveness Index using the TAAS reading assessment produced consistent results across the four years while the Classroom Effectiveness Index using the TAAS mathematics assessment did not.

Further investigation indicated that mathematics scores across the four years improved at a much greater rate than the reading scores. This may account for the differences in slopes and intercepts in mathematics. It is reasonable that within a system based on improving student achievement that teachers and student may be measured against differing standards when achievement increases rapidly. However, this information must be clearly understood if the information is to be included in a teacher accountability system.

How outliers influenced a teacher's CEI was also assessed. It was determined that there were 19 outliers out of 15,378 identified as statistically significant. The dramatic influence of even a small number of outliers has been demonstrated (Fox, 1991; Ho & Naugher, 2000). Unusual data points in a least squares regression analysis are problematic because the data points are outside the rest of the distribution and influence the regression analysis.

Ho and Naugher (2000) demonstrated that outliers can easily be detected either visually through plotting the data or through statistical tools that included

Mahalonobis distance and Cook's distance. These two measures give additional information about the nature of the outliers. Mahanobis distance measures how far the outlier is from the center of the multidimensional distribution (Tabachnik & Fidell, 1996). The Mahalanobis distance attributed to each of the identified outliers in this study indicated the degree to which the Classroom Effectiveness Index was affected.

Cook's distance assessed the degree of change in the regression coefficient when an outlier was deleted (Ho & Naugher, 2000; Tabachnik & Fidell, 1996). No statistically significant outliers using Cook's distance were identified in this study. However, the two outliers with the highest influence scores of CD =.32 and CD =.52 corresponded with the two largest Mahalonobis distance scores, indicating that these two outliers were possibly exerting a small degree of change on the two regression coefficients.

Although 19 statistically significant outliers were detected in this study, only eight were of practical significance. When the CEIs were recomputed for these eight teachers, the placement of the newly computed CEIs within the overall distribution of CEIs changed to the extent that the teacher's effect on student achievement would be viewed differently. The change in removing one outlier in each case moved the teacher toward the district average CEI, and the teacher would be interpreted as a teacher whose students had achievement growth similar to the district's average.

The influence of outliers within the Dallas HLM Model is small when viewed from an overall perspective; 19 statistically significant outliers in 15,378 student

residuals. However, the impact on individual teachers can be extreme. The findings in this study indicated a need to address outlier influence if the CEI is to be viewed as a consistent measure of student achievement for teacher accountability. Since the Classroom Effectiveness Index is an indicator of the average student achievement, outliers that exert undue influence must be addressed. How outliers influence CEIs must be established before exploring ways to adjust their influence.

Classroom Effectiveness Indices across consecutive years were also analyzed to indicate whether or not CEIs changed from year to year. CEIs that fluctuated from year to year would indicate a possible unstable assessment of a teacher's influence on student achievement. In all instances, the analysis indicated that CEIs did not change significantly from year to year, whether they were assessed across four years of teaching or three years of teaching and whether they were generated from a reading assessment or a mathematics assessment. This lack of significance indicates that teachers' CEIs were relatively stable across years. There were no statistically significant differences identified in either mathematics or reading CEIs, indicating that the CEIs across years in either subject area were stable.

Finally, the relationship between the initial student residual, the Level One residual and the final student score, the HLM residual, was investigated to see whether or not there was a significant correlation between the two. In all instances, the correlations in both reading and mathematics for each of the four years were above +.69. Therefore, there was a statistically significant positive correlation when comparing the CEIs from the Level One residual and the final HLM residual. School

level variables, controlled for in the HLM stage of the procedure, did not influence student residuals differently. This implies that the model was consistent for all students within each yearly computation.

Relationship of Findings to Review of Literature

The possibility of establishing teacher differences using student achievement has been established (Bingham, Heywood & White, 1991; Holt & Collins, 2001; Sanders & Horn, 1993; Webster, Mendro, Orsak & Weerasinghi, 1997). Other statistical properties of the Classroom Effectiveness Index developed by the Dallas Model have been investigated and found to be unbiased across teacher gender, years of teaching, and previous student achievement (Bembry, Weerasinghe, & Mendro, 1997). This study adds to the understanding of the model since it establishes that the CEIs were also consistent over time.

The school district investigated several alternative regression models during the development of the current Dallas HLM Model (Webster et. al., 1994). The correlations among the varying multiple regression models and the HLM models being investigated indicated similar results across all statistical models. This study extends the information concerning the relationship between the first and second stage of the model. The significant correlations between the Level One residual and the HLM residual indicated consistency in the outcomes from the two stages. This additional information also contributed to the information concerning the influence of the school level characteristics on school rankings (Webster et. al., 1994). The influence of school characteristics on the final school rankings in the earlier study

was identified as less than three percent of the variance. Information concerning the relationship between school characteristics and the initial student residual indicated that there is consistency in the outcomes of the two stages.

The influence of outliers in two-level models continues to be investigated (Ho & Naugher, 2000; Sheehan & Han, 1996). Mahalanobis distance and Cook's distance have been found to indicate how outliers influence the estimates of regression coefficients. The findings in this study confirm the dramatic influence of outliers on the final outcomes of HLM, with the unique opportunity for additional information: the identification of only one outlier for each teacher within the subset of 19 teachers allowed for the investigation of the influence of one outlier within groups of varying numbers of residuals. The study indicated that outliers do influence a small portion of the teachers' final CEIs, with important practical consequences.

In addition to providing information concerning HLM models in schools, information about the Dallas HLM model, and the influence of outliers in regression, this study also added to the discussion concerning teacher accountability. Student achievement is the single most important concern for public schools, and the teacher is the most important factor in how well students learn. Attempts to identify appropriate measures of teacher accountability using student achievement data is of paramount concern. This study included information concerning how well a value-added assessment using student achievement performed over time.

Educational Importance

At present, 49 states have established standards for student learning (Linn, 2001). Communities continue to ask for evidence that students are learning in the public school systems. In addition, the testing and research community have expressed concern over using student achievement data in high stakes situations, according to the AERA Position Statement Concerning High-Stakes Testing in PreK-12 Education. Concern is expressed that tests and assessments are being used in ways other than those for which the tests and assessments were intended.

> Within this context, the Dallas HLM Model established a process for developing adjusted student growth scores in the form of residuals from achievement tests. The model was developed as a fair measure of school and teacher effectiveness across very different instructional circumstances. The information at the teacher level has been used as additional information teachers and principals may use in assessing the success of their instructional efforts. Any change in the use of the CEIs from planning data to an accountability measure would necessitate additional understanding of their characteristics, especially in the context of personnel evaluation.

Recommendations for Future Research

The results of this study indicated that the Classroom Effectiveness Index overall yields reliable results over time. More study is needed to better understand the impact of influential outliers in the student residuals and to identify causes for the instability of the mathematics grand intercept and slope from year to year.

The investigations in this study should be applied to other K-12 grades in order to further validate findings. CEIs computed from both elementary and high school test scores in addition to eighth grade scores should be investigated in a similar study. Second, applying the Dallas HLM model to populations other than a large urban district may yield further information concerning the utility of CEIs. Third, various procedures for identifying and isolating outliers in student residuals should be investigated. The nature of the outliers should indicate how their influences can be adjusted. Finally, other multi-level models for assessing teacher effectiveness on student achievement need to be investigated and compared to the current model.

Two other characteristics of the Dallas HLM model need to be investigated. Teachers are measured against the average overall gain within the district in that subject area for that year. Thus, by its very nature, half of the teachers will be above the district average and half will be below the district average. This may be a sufficient model in a large school district with great differences in student achievement across teachers and schools, but it may not hold in other districts

Second, the model is built on the concept that residuals from a process that begins with an achievement score and controls for student and school characteristics represents student achievement to such a degree that it is an indicator of student growth. Althugh there is precedent in the literature, Bryk and Raudenbush have only extended the research on school effectiveness using a multi-level model and residuals since 1989. Since then, both school districts and states have initiated some form of value-added assessment in order to evaluate student achievement. The underlying

question still needs to be answered: How much of the residual is student achievement and how much is error? In order to clarify this issue, other valid measures of teacher effectiveness must be identified and their relationships to the CEIs established.

What has passed for variables of teacher effectiveness in the past are variables often not related to student achievement. How well a teacher adheres to a paradigm established by a model of effective teaching does not indicate whether or not students in that teacher's classroom learn. The current situation in public education requires that any assessment of an effective teacher be related to how well his or her students learn. To this end, a set of variables for identifying teacher effectiveness as indicated by student achievement needs to be identified. A possible future investigation might be to use the CEI as a dependent measure to predict teacher effectiveness rather than using it to infer teacher effectiveness, thereby adding to the understanding of effective teaching.

Finally, an accountability system at the teacher level has an added layer of responsibility that a school level accountability system does not have. Any assessment of a teacher would need to adhere to The Personnel Evaluation Standards developed by the Joint Committee on Standards for Educational Evaluation (1988). An explanation of the Reliable Measurement Standard states that "consistency should be sought across different indicators of the same criterion" (p.104). The guidelines for this same standard recommends that evidence of reliability of any measure is collected prior to using that measure in an evaluation system. Therefore, any teacher accountability system that does not use multiple measures of success and does not

publish the reliability of any included measure would not be in compliance with the standards.

Much is still unknown about teacher effectiveness on student achievement, especially when confined to teacher activity in one school year. While investigations into the Dallas HLM model continue and especially with the inconclusive results from this study, it is recommended that several years of CEIs should be reported for an individual teacher in assessing overall teacher effectiveness on student achievement, especially in the context of a teacher accountability system.

APPENDIX

(Mainframe Program – Establishing the Student Database)

```
FILE DESCRIPTION.                                                          00000100
KAREN/MASTER                                                               00000200
                                                                           00000300
FILE MAXRECSIZE=XXX  BLOCKSIZE=XXXX                                        00000400
                                                                           00000500
*Fld Strt Stop Fmt  Mnemonic          Description                          00000600
  1   1    6 I6   STUDENT-ID                EOY 1-6                       00000700
                          REFERENCES TO EOY ARE TO                         00000800
                          INDB9900DEMOEOY                                  00000900
  2   7    9 I3   REPORT-LOC                EOY 10-12                     00001000
  3  10   11 A2   GRADE                     EOY 15-16                     00001100
  4  12   12 I1   GENDER                    EOY 23  1=M 2=F               00001200
  5  13   13 I1   ETHNICITY                 EOY 24  1-5,b                 00001300
  6  14   14 I1   LUNCH-CODE                EOY 161  1,2,3,7,8=1          00001400
  7  15   15 A1  LIMITED-ENG-DATA           EOY 178  Y,N  Y=1             00001500
  8  16   16 I1   ETHCODE           1=OTHER 2=B 3=HEP 4=LEP                00001600
  9  17   17 I1   BLACK-EP          BLACK CODE                             00001700
 10  18   18 I1   HISPANIC-EP        HISPANIC EP                           00001800
 11  19   19 I1   LEP                ALL LEP                               00001900
 12  20   20 I1   BEP-BY-GEND         BLACK GEND INTERACTION               00002000
 13  21   21 I1   HEP-BY-GEND         HEP GEND INTERACTION                 00002100
 14  22   22 I1   LEP-BY-GEND         LEP GEND INTERACTION                 00002200
 15  23   23 I1   BEP-BY-LUNCH        BEP LUNCH INTERACTION                00002300
 16  24   24 I1   HEP-BY-LUNCH        HEP LUNCH INTERACTION                00002400
 17  25   25 I1   LEP-BY-LUNCH        LEP BY LUNCH INTERACTION             00002500
 18  26   26 I1   GEND-BY-LUNCH       GEND LUNCH INTERACTION               00002600
 19  27   27 I1   B-X-GEND-X-LUNCH    B X G X L 3WAY INTERACTION           00002700
 20  28   28 I1   H-X-GEND-X-LUNCH    H X G X L 3WAY INTERACTION           00002800
 21  29   29 I1   LEP-X-GND-X-LNCH    LEP X G X L 3WAY INTERACTION         00002900
 22  30   38 F9.4 CENSUS-INC           CENSUS INCOME                        00003000
 23  39   43 F5.3 CENSUS-POV           CENSUS POVERTY                       00003100
 24  44   48 F5.3 CENSUS-COL           CENSUS COLLEGE                       00003200
 25  49   49 A1   TAAS-M-FLAG-99       TAASEXT M-FLAG 27  S=SCORE           00003300
 26  50   51 I2   TAAS-M-TOT-RAW-99    TAASEXT M-TOT-RAW 30-31              00003400
 27  52   52 A1   TAAS-R-FLAG-99       TAASEXT R-FLAG 51  S=SCORE           00003500
 28  53   54 I2   TAAS-R-TOT-RAW-99    TAASEXT R-TOT-RAW 54-55              00003600
 29  55   55 A1   TAAS-M-FLAG-00       TAASEXT M-FLAG 30  S=SCORE           00003700
 30  56   57 I2   TAAS-M-TOT-RAW-00    TAASEXT M-TOT-RAW 31-32              00003800
 31  58   58 A1   TAAS-R-FLAG-00       TAASSEXT R-FLAG 55  S=SCORE          00003900
 32  59   60 I2   TAAS-R-TOT-RAW-00    TAASEXT R-TOT-RAW 56-57              00004000
 33  61   69 F9.6 RTAAS-M-TOT-99                                            00004100
 34  70   78 F9.6 RTAAS-R-TOT-99                                            00004200
 35  79   87 F9.6 RTAAS-M-TOT-00                                            00004300
 36  88   96 F9.6 RTAAS-R-TOT-00                                            00004400
 37  97  120 X24  FIELD-37                                                 00004500
```

(Mainframe Program – Student Residuals - Math)

```
FILE DESCRIPTION.                                                    00000100
KAREN/TAAS/GR8/MATH.                                                 00000200
*FLD STRT STOP FMT    MNEMONIC          DESCRIPTION                  00000300
  1    1    6 I6   STUDENT-ID                                       00000400
  2    7    9 I3   REPORT-LOC                                       00000500
  3   10   11 I2   GRADE                                           00000600
  4   12   12 I1   GENDER          1=MALE, 2=FEMALE                 00000700
  5   13   13 I1   MAGNET-FLAG      1=IN MAGNET, 2=NOT               00000800
  6   14   14 I1   LUNCH-CODE       1=FREE-LUNCH, 2=NOT              00000900
  7   15   15 I1   BLACK-EP                                         00001000
  8   16   16 I1   HISPANIC-EP                                      00001100
  9   17   17 I1   LEP                                             00001200
 10   18   18 I1   BEP-BY-GEND                                     00001300
 11   19   19 I1   HEP-BY-GEND                                     00001400
 12   20   20 I1   LEP-BY-GEND                                     00001500
 13   21   21 I1   BEP-BY-LUNCH                                    00001600
 14   22   22 I1   HEP-BY-LUNCH                                    00001700
 15   23   23 I1   LEP-BY-LUNCH                                    00001800
 16   24   24 I1   GEND-BY-LUNCH                                   00001900
 17   25   25 I1   B-X-GEND-X-LUNCH                                00002000
 18   26   26 I1   H-X-GEND-X-LUNCH                                00002100
 19   27   27 I1   LEP-X-GND-X-LNCH                                00002200
 20   28   36 F9.4 CENSUS-INC                                      00002300
 21   37   41 F5.3 CENSUS-POV                                      00002400
 22   42   46 F5.3 CENSUS-COL                                      00002500
 23   47   48 I2   TAAS-M-TOT-RAW-99                               00002600
 24   49   50 I2   TAAS-M-TOT-RAW-00                               00002700
 25   51   59 F9.6 RTAAS-M-TOT-99                                  00002800
 26   60   68 F9.6 RTAAS-M-TOT-00                                  00002900
```

(Mainframe Program – Student Residuals – Reading)

```
FILE DESCRIPTION.                                                    00000100
KAREN/TAAS/GR8/READ.                                                 00000200
*FLD STRT STOP FMT    MNEMONIC          DESCRIPTION                  00000300
  1    1    6 I6   STUDENT-ID                                       00000400
  2    7    9 I3   REPORT-LOC                                       00000500
  3   10   11 I2   GRADE                                           00000600
  4   12   12 I1   GENDER           1=MALE, 2=FEMALE                00000700
  5   13   13 I1   MAGNET-FLAG       1=IN MAGNET, 2=NOT             00000800
  6   14   14 I1   LUNCH-CODE        1=FREE-LUNCH, 2=NOT            00000900
  7   15   15 I1   BLACK-EP                                         00001000
  8   16   16 I1   HISPANIC-EP                                      00001100
  9   17   17 I1   LEP                                             00001200
 10   18   18 I1   BEP-BY-GEND                                     00001300
 11   19   19 I1   HEP-BY-GEND                                     00001400
 12   20   20 I1   LEP-BY-GEND                                     00001500
 13   21   21 I1   BEP-BY-LUNCH                                    00001600
 14   22   22 I1   HEP-BY-LUNCH                                    00001700
 15   23   23 I1   LEP-BY-LUNCH                                    00001800
 16   24   24 I1   GEND-BY-LUNCH                                   00001900
 17   25   25 I1   B-X-GEND-X-LUNCH                                00002000
 18   26   26 I1   H-X-GEND-X-LUNCH                                00002100
 19   27   27 I1   LEP-X-GND-X-LNCH                                00002200
 20   28   36 F9.4 CENSUS-INC                                      00002300
 21   37   41 F5.3 CENSUS-POV                                      00002400
 22   42   46 F5.3 CENSUS-COL                                      00002500
 23   47   48 I2   TAAS-M-TOT-RAW-99                               00002600
 24   49   50 I2   TAAS-R-TOT-RAW-99                               00002700
 25   51   52 I2   TAAS-R-TOT-RAW-00                               00002800
 26   53   61 F9.6 RTAAS-M-TOT-99                                  00002900
 27   62   70 F9.6 RTAAS-R-TOT-99                                  00003000
 28   71   79 F9.6 RTAAS-R-TOT-00                                  00003100
```

File Description for HLM program

For Down Load For HLM	Format	Variable Name	Notes
Student Level Files	A3	SLN	
	A6	Student ID	
	F10.6	Outcome	Include decimal point
	F10.6	Predictor 1	Include decimal point
	F10.6	Predictor 2 (if needed)	Include decimal point
School Level Files	A3	SLN	
	A2	Grade	
	F5.1	% Mobility	Include decimal point
	F5.1	% Over Crowding	Include decimal point
	F7.3	Census Income * 1000	Include decimal point
	F5.1	Census Poverty	Include decimal point
	F5.1	Census College	Include decimal point
	F5.1	% Free/Reduces Lunch	Include decimal point
	F5.1	% LEP	Include decimal point
	F5.1	% Black	Include decimal point
	F5.1	% Hispanic	Include decimal point
	F5.1	% Minority	Include decimal point
For Upload to mainframe			
Student Residuals	F8.0	SLN	
Not Sorted by Student ID	F8.0	Student ID	
	F8.6	Residuals	No decimal point
School EBs	F8.0	SLN	
Sorted By SLN	F8.0	Number Of students	
	F8.6	Ranking	No decimal point

January 31, 2001

Karen Bembry
[address removed]

Dear Karen,

I have reviewed and approved your proposal to conduct the study, *Establishing the Utility of a Classroom Effectiveness Index As a Teacher Accountability Measure*, using the Dallas Independent School District databases.

This approval is with the understanding that you have agreed to the procedures and policies for conducting research in the Dallas Independent School District. I will serve as your contact person regarding any additional information you will need.

Best wishes on your study.

Sincerely,

Robert L. Mendro
Chief Evaluation Officer
Evaluation, Accountability and Information Systems

Classroom Effectiveness Indices and Number of Student Residuals for 4 Year Reading Teachers

Teacher	N	1997 CEI	N	1998 CEI	N	1999 CEI	N	2000 CEI
R-4-1	44	60.16517022	52	55.24052749	38	46.35068908	29	47.71888290
R-4-2	61	50.19146596	55	56.28947935	61	48.12616155	44	53.62165089
R-4-3	48	42.20234169	31	50.26236828	29	52.21694504	13	56.38133222
R-4-4	29	40.05386947	49	44.77426138	48	48.58612215	47	41.70872634
R-4-5	53	37.21475638	61	41.35924608	67	51.42774806	62	53.75988587
R-4-6	36	36.49754735	51	49.38147393	48	52.43896498	61	53.08448968
R-4-7	31	36.97095958	36	47.51626329	65	47.31278160	25	52.09414861
R-4-8	37	42.31583528	39	48.78409761	62	46.32925052	63	53.86635824
R-4-9	54	48.05119034	65	43.00711006	69	47.47119581	62	46.21627829
R-4-10	49	65.37301482	49	53.46211018	44	45.14567465	26	48.18983179
R-4-11	27	47.57751056	65	49.66855411	29	51.25716642	47	43.90621314
R-4-12	66	38.98537291	44	39.82629601	62	48.10518765	74	47.43600043
R-4-13	46	58.90971619	55	52.05918650	99	50.70383682	68	45.58067319
R-4-14	56	50.35033598	47	48.19511286	53	48.43992884	64	47.66983779
R-4-15	46	48.80048170	43	48.76106298	69	52.64204402	54	46.26650972
R-4-16	51	41.67082127	51	40.93508028	60	50.24070242	29	45.92057531
R-4-17	56	56.85089153	72	55.95845626	51	49.72009928	35	53.97382340
R-4-18	41	46.45750158	47	46.05340583	64	44.22339714	55	42.72819610
R-4-19	53	47.69578095	69	39.70420313	73	45.16653629	62	45.00870842
R-4-20	47	42.55933090	48	43.25145945	39	36.37911138	37	50.70492163
R-4-21	46	39.58230492	57	48.12295043	47	42.71071978	50	47.44523791
R-4-22	36	55.58417784	34	42.80934507	49	40.48794253	31	52.45322006
R-4-23	38	50.42598131	63	48.57289836	47	49.01541251	19	53.60864549

Classroom Effectiveness Indices and Number of Student Residuals for 4 Year Math Teachers

Teacher	N	1997 CEI	N	1998 CEI	N	1999 CEI	N	2000 CEI
M-4-1	22	48.37441132	23	40.59934027	12	45.34935970	14	54.19299155
M-4-2	63	44.99341712	76	45.39921547	46	47.31113015	18	38.77735550
M-4-3	53	54.97027188	58	53.67784243	66	46.50143866	60	47.83311159
M-4-4	52	41.71168079	82	47.77138967	56	51.98601651	43	48.70108847
M-4-5	28	48.97443000	42	48.51668259	13	46.98606081	48	40.68090976
M-4-6	28	44.33101908	56	50.84772665	43	54.34780290	24	51.08076676
M-4-7	60	50.21498777	63	51.96518846	58	48.55356865	12	46.38819238
M-4-8	50	39.94493295	69	49.70946556	20	45.65189349	31	48.51567150
M-4-9	53	49.76044727	65	44.46188030	24	44.07536722	46	41.64468818
M-4-10	51	47.26659341	19	46.67091551	25	44.92586102	28	40.96794118
M-4-11	88	47.68133323	79	44.16231429	73	42.72209149	48	37.70442611
M-4-12	57	54.12782783	47	49.62685625	88	46.68863725	57	40.46229853
M-4-13	21	61.96950924	31	53.65468210	65	53.00472381	58	48.70154380
M-4-14	69	56.63074449	23	56.46786744	30	45.44429022	41	39.77026084
M-4-15	48	37.44704002	49	45.12457182	15	49.78596320	38	46.68043881
M-4-16	51	41.98162286	65	41.69421736	25	54.05879062	36	52.60855854
M-4-17	35	60.38853039	47	45.61445241	54	36.17767515	44	48.67858475
M-4-18	56	54.43274272	44	53.60011097	60	51.62042622	86	52.59632976
M-4-19	36	32.29237765	14	48.59488952	34	50.34085810	32	52.97636353
M-4-20	22	62.73154193	33	61.29924590	23	59.68170035	16	55.94986677
M-4-21	67	40.14702728	71	49.48903997	97	49.99155833	75	60.06145168
M-4-22	18	42.99627282	18	38.15439680	17	40.84422243	13	38.78949799
M-4-23	49	47.89975637	45	48.54105608	86	51.32587761	48	52.84422312
M-4-24	32	45.05502331	32	41.21979619	61	48.17119147	49	44.75838323
M-4-25	14	51.34275979	66	39.33652055	27	54.22126066	22	39.23237246

Classroom Effectiveness Indices and Number of Student Residuals for 3 Year Reading Teachers

Teacher	N	1997 CEI	N	1998 CEI	N	1999 CEI	N	2000 CEI
R-3-1	46	48.83286362	51	40.92422710	40	46.86568347		
R-3-2	79	43.45453388	81	47.17050567	75	46.75569748		
R-3-3			55	44.07332891	67	45.05292465	55	50.03467637
R-3-4			39	53.09142761	98	47.99803841	85	49.42494868
R-3-5	30	48.19684908	43	51.80517143	31	54.39870871		
R-3-6	32	59.07168728	49	46.61596710	75	49.22642584		
R-3-7	59	56.07864655	78	60.43956873	47	37.03835139		
R-3-8			77	42.23009587	61	53.30291265		
R-3-9			26	53.03183505	60	53.06424119	63	53.24115161
R-3-10			52	51.55139648	53	51.85069698	44	55.63743248
R-3-11			16	51.06190033	30	50.57034232	22	50.99802262
R-3-12	54	37.56722394	52	43.50543785	59	47.84641501		
R-3-13			66	40.36540955	68	46.62463771	36	50.92881809
R-3-14	63	38.12042040	84	47.62623246	77	47.06651882		
R-3-15			47	43.13940566	37	48.33389430	36	50.64343985
R-3-16			20	47.34756466	17	54.67527207	17	50.87906503
R-3-17			37	54.73281769	56	51.91356171	47	56.80742181
R-3-18	47	51.31169903	35	54.79454411	77	51.59436345		
R-3-19	67	48.50149438	45	52.19835942	77	49.41630260		
R-3-20			29	51.13239642	30	49.13208166	33	51.84690604

Classroom Effectiveness Indices and Number of Student Residuals for 3 Year Math Teachers

Teacher	N	1997 CEI	N	1998 CEI	N	1999 CEI	N	2000 CEI
M-3-1			15	43.62079874	30	39.67094252	13	43.96636970
M-3-2			47	49.29127510	62	41.89420671	82	41.54073017
M-3-3			87	37.79623127	66	41.68177202	65	36.22005430
M-3-4	43	49.33752453	40	38.70039712	45	41.27706267		
M-3-5	52	47.49144250	67	48.16357076	59	44.93871283		
M-3-6	43	53.54723864	30	52.12217876	22	53.45072780		
M-3-7			26	50.04290976	46	42.78470470	37	42.88249670
M-3-8			61	41.19506388	23	41.43573963	29	40.59862950
M-3-9			69	44.15090077	31	56.46895513	29	49.92972311
M-3-10	59	50.38554533	63	38.03264626	61	45.76048824		
M-3-11			29	47.27728629	36	47.23630894	38	56.05141426
M-3-12			55	52.00760147	63	47.83045187	41	44.96770892
M-3-13	57	51.70757059	54	47.54848027	60	48.27783808		
M-3-14			54	52.68092955	52	57.49199819	65	56.63523728
M-3-15			13	49.01583467	19	56.10503763	16	43.71792789
M-3-16			24	49.31560554	22	52.20211418	19	41.14819744
M-3-17			66	45.95300021	61	48.17149512	55	53.07345193
M-3-18	67	48.20606459	72	47.33203703	36	40.94896349		
M-3-19	34	40.64469341	32	42.31952038	74	45.08704048		
M-3-20	32	60.52171113	35	42.16789956	40	45.61626915		
M-3-21	49	49.39840064	53	47.18586153	66	45.07394121		
M-3-22	18	48.52536755	19	58.09497395	22	55.82346602		
M-3-23	74	53.33753354	60	47.81768798	49	39.72275093		
M-3-24			54	37.67861852	61	40.79223759	59	34.92968120
M-3-25			42	43.05259093	45	48.24622120	30	55.60156111
M-3-26			77	49.34126999	59	41.45030093	72	40.78863727
M-3-27	49	69.39609785	48	47.88073215	43	46.64195017		

M-3-28			42	60.11221344	31	60.02544430	32	55.10085904
M-3-29			56	42.21905555	58	45.18158954	40	46.06650405
M-3-30	15	50.41920476	18	45.09061561	16	40.70954589		
M-3-31			20	47.69134661	33	54.91014343	45	55.50822216
M-3-32			35	62.44261661	23	65.66112718	31	65.97013877

REFERENCES

Adcock, E. P. & Phillips, G. W. (1997). Measuring school effects with hierarchical linear modeling: Data handling and modeling issues. Multiple Linear Regression Viewpoints, 24, 1-10.

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

Bembry, K. L., Weerasinghe, D. & Mendro, R. L. (1997, March). Classroom effectiveness indices: Statistical methodology, post-hoc analysis, and practical applications. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Bingham R. D., Heywood J. S., & White, S. B. (1991). Evaluating schools and teachers based on student performance. Evaluation Review, 15, 191-218.

Bock, R. D. (Ed.). (1989). Multilevel analysis of educational data. San Diego: Academic Press.

Bryk, A. S. & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In Bock, R. D. (Ed). Multilevel Analysis of Educational Data. (pp.159-199). San Diego: Academic Press.

Bryk, A. S. & Raudenbush, S. W. (1992). Hierarchical linear models: Applications, and data analysis methods. Newbury Park: Sage Publications.

Commission for educational excellence: Final report. (1991). Dallas, Texas: Dallas Independent School District.

Darling-Hammond, L. (1997). The right to learn: A blueprint for creating schools that work. San Francisco: Jossey-Bass Publications.

Farkas, G., Sheehan, D., Grobe, R. P. (1990). Coursework mastery and school success: gender, ethnicity, and poverty groups within an urban school district. American Educational Research Journal, 27, 807-827.

Fox, J. (1991). Regression diagnostics. Newbury Park: Sage Publications.

Jordan, H. R., Mendro, R. L. & Weerasinghe, D. (1997, July). Teacher effects on longitudinal student achievement: A report on research in progress. Paper presented at the meeting of CREATE, Indianapolis, IN.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. Educational Evaluation and Policy Analysis, 8, 45-60.

Ho, K. & Naugher, J. R. (2000). Outlier Lies: An Illustrative Example of Identifying Outliers and Applying Robust Models. Multiple Linear Regression Viewpoints, 26, 2-6.

Holt, J. K. & Collins, V. L. (2001). Dynamic Accountability Systems: Multilevel Modeling of Educational Growth. Multiple Linear Regression Viewpoints, 27, 46-52.

Holt, J. K. (2001). Guest Editor's Introduction: Hierarchical Linear Models. Multiple Linear Regression Viewpoints, 27, 1-2.

Joint Committee on Standards for Educational Evaluation. (1988). Personnel Evaluation Standards: How To Assess Sytems for Evaluating Educators. Newbury Park: Sage Publications.

Linn, R. L. (2001). The design and evaluation of educational assessment and accountability aystems. CSE Technical Report 539. Center for the Study of Evaluation, Los Angeles, CA.

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

Millman, J. (1997). Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Mundfrom, D. J. & Schultz, M. R. (2001) A Comparison Between Hierarchical Linear Modeling and Multiple Linear Regression in Selected Datasets. Multiple Linear Regression Viewpoints, 27, 3-11.

Orsak, T. O., Mendro, R. L., & Weerasinghe, D. (1998). Calculating missing student data in hierarchical linear modeling: Uses and their effects on school rankings. Multiple Linear Regression Viewpoints 25, 3-12.

Phillips, G. W. & Adcock, E. P. (1997). Practical applications of hierarchical linear models to district evaluations. Multiple Linear Regression Viewpoints, 23, 25-34.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. Journal of Educational Statistics, 13, 85-116.

Sanders, W. L. & Horn, S. P. (1993). The Tennessee value-added assessment system: Mixed model methodology in educational assessment. University of Tennessee Value-Added Research and Assessment Center: Knoxville, TN.

Sanders, W. L. & Rivers, June C. (1996). Cumulative and residual effects of teachers on future student academic achievement. University of Tennessee Value-Added Research and Assessment Center: Knoxville, TN.

Schumacker, R. E. & Bembry, K. L. (1995). Empirical characteristics of centering methods for level-1 predictor variables in HLM. Multiple Linear Regression Viewpoints 23, 1-8.

Sheehan J. K. & Han, T. (1996, April). How do extreme schools change the interpretation of results in school effectiveness research? Effects of outlying second-level variables in HLM. Paper presented at the meeting of the American Educational Research Association, New York City, NY.

Sheehan, J. K. & Han, T. (1997, March). Detection and modeling of outliers in 3-level hierarchical models: An examination of unusual growth rates from LSAY. Paper presented at the meeting of the American Educational Research Association, March, Chicago, IL.

Snijders, T. A. B. & Bosker, R. L. (1993). Standard errors and sample sizes for two-level research. Journal of Educational Statistics, 18, 237-259.

Stevens, J. (1992). Applied multivariate statistics for the social sciences, (2nd ed.) Hillsdale, NJ : Lawrence Erlbaum.

Stevens, J. (1999). Intermediate statistics: A modern approach. London: Lawrence Erlbaum.

Tabachnick, B. G. & Fidell, L. S. (1996). Using multivariate statistics, (3rd ed.) Northridge: HarperCollins Publishers, Inc.

Thomas, S. L. & Heck, R. H. (2001). Multilevel Models: Thoughts About the Future. Multiple Linear Regression Viewpoints, 27, 57-61.

Webster, W. J., & Olson, G. H. (1988). A quantitative procedure for the identification of effective schools. Journal of Experimental Education, 56, 213-219.

Webster, W. J., Mendro, R. L., and Almaguer, T. O. (1993). Effectiveness indices: The major component of an equitable accountability system, ERIC TM 019 913.

Webster, W. J., Mendro, R. L., Bembry, K. L., & Orsak, T. H. (1995, April). Alternative Methodologies for Identifying Effective Schools. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.