DISSERTATION

USING HIERARCHICAL LINEAR MODELING TO MEASURE SCHOOL EFFECTS ON THE COLORADO STUDENT ASSESSMENT PROGRAM

Submitted by

Marc A. Winokur

School of Education

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

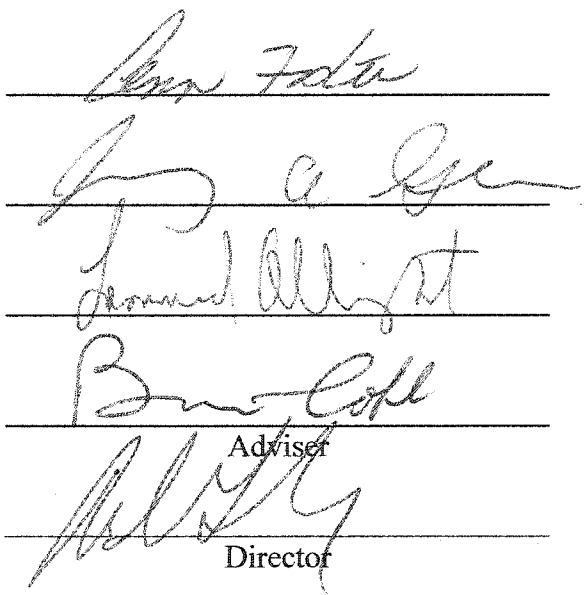
Summer 2004

COLORADO STATE UNIVERSITY

April 27, 2004

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY MARC A. WINOKUR ENTITLED *USING HIERARCHICAL LINEAR MODELING TO MEASURE SCHOOL EFFECTS ON THE COLORADO STUDENT ASSESSMENT PROGRAM* BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work



Adviser

Director

ABSTRACT OF DISSERTATION

USING HIERARCHICAL LINEAR MODELING TO MEASURE SCHOOL EFFECTS ON THE COLORADO STUDENT ASSESSMENT PROGRAM

The purpose of this dissertation was to contribute a theoretically and empirically sound approach for analyzing and interpreting results from the Colorado Student Assessment Program (CSAP). This study was grounded in school effectiveness research, as the primary objective was to isolate the impact of school practice by controlling for student characteristics and school context. Furthermore, this investigation was designed to build on contemporary studies that have employed promising statistical models to analyze high-stakes tests. Thus, secondary databases were analyzed using hierarchical linear modeling (HLM) to identify the most consistent and powerful predictors of student achievement on the CSAP fourth-grade reading test. A value-added analysis also was implemented to estimate school effects and to compare school performance on this standards-based statewide assessment. The significance of this research is that interpretations of high-stakes testing (HST) can be made fair and accurate for all educational stakeholders.

The two-level HLM analysis generated a predictive model based on the specification of significant student- and school-level variables. Specifically, prior achievement, special education status, socioeconomic status (SES), ethnicity, and continuous enrollment accounted for 64% of the within-school variance in student achievement. School SES and teacher experience accounted for 77% of the between-

school variance. The analysis of school effects estimates produced the most interesting finding, as performance differences and school rankings yielded contradictory results. For example, schools that outperformed expectations tended to have smaller value-added residuals than did schools that scored below their predicted value. Furthermore, there was a differential effect of school practice on students within a school based on prior achievement.

The main implication of this study is that school accountability systems will be inequitable unless they control for the effect of educational inputs on HST results. The primary recommendation is that researchers, educators, and policymakers should utilize appropriate quantitative and qualitative methods to better understand and explain the complex dynamic between school practice and student achievement.

Marc A. Winokur
School of Education
Colorado State University
Fort Collins, CO 80523
Summer 2004

ACKNOWLEDGEMENTS

I want to first thank Dr. Brian Cobb, Dr. Ann Foster, Dr. Jeff Gliner, and Dr. Len Albright for their steadfast support during the writing of this dissertation. As a doctoral committee and as friends, they offered me the encouragement, expertise, and guidance I needed to make this an extremely rewarding learning experience. I would like to especially recognize Brian for providing me with so many exceptional opportunities for educational research and program evaluation during my time at CSU.

I also want to acknowledge the faculty, staff, and students in the School of Education for their kindness and professionalism. I owe a special debt of gratitude to Dr. Carole Makela for her extensive feedback on my proposal and to Dr. Don Quick for his timely help in formatting this document.

Many thanks to Chrys Dougherty from Just for the Kids, Dr. Robert Reichardt formally of McREL, Dr. Susan Schafer of the Colorado Department of Education, and Joan Clay from the Research and Development Center for their assistance in obtaining the data used in this study.

I wish to express my deepest gratitude to my wife, Heather, whose patience, love, and understanding truly kept me going during the many highs and lows of being a graduate student. Finally, thank you to my parents, Shelly and Steve, and to all of my family and friends for always believing in me.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The expanded use of high-stakes testing in the United States has heralded a new age for school accountability. The future holds even greater emphasis for this educational reform with the passage of Public Law 107-110, more commonly known as the No Child Left Behind Act of 2001 (NCLB). Coming on the heels of state legislation that requires schools to assess students on numerous academic standards, this federal mandate contains provisions for the annual testing of all students in grades 3-8. Since high-stakes testing (HST) is here to stay, there is a need for conceptual and applied research that critically investigates the impact of these assessments on students, teachers, and schools.

Problem Statement

During the past ten years, high-stakes testing has been thrust to the forefront of educational policy, practice, and research (Hess & Brigham, 2000; Ryan, 2002). For example, educational researchers have been asked to assess the psychometric and statistical properties of standardized tests and statewide assessments that serve as school accountability measures (Raudenbush & Willms, 1995). According to Bernstein (1990), past studies "have often arrived at conflicting conclusions due to the use of divergent sampling procedures, different units of analysis, disparate data analysis techniques, and varying operational definitions of the input variables and the outcome measures" (p. 1).

Traditional statistical analyses also have ranked schools on uncorrected proficiency levels to determine the highest and lowest performers (Schagen, 1991). This has led to a metaphorical comparison of "apples to oranges" that forces politicians and

educators to draw unsupported and illogical inferences from HST results (Ediger, 2001; Haladyna, Nolen, & Haas, 1991). Thus, most current interpretations of school accountability measures are unfair for certain student populations (Haladyna, Haas, & Allison, 1998). Furthermore, the persistent "achievement gap" on high-stakes tests between students of different ethnic and socioeconomic backgrounds exacerbates this inequity (Nye, Hedges, & Konstantopoulos, 2002).

Conceptual Framework

According to Raudenbush and Willms (1995), "theory and appropriate measures, though difficult to collect, supply a foundation for studying the contributions of school practice" (p. 332). The present study is grounded in school effectiveness research (SER), in that school practice is "conceived to be the unobservable part of the school effect that remains after removing the contributions of student background and context" (Raudenbush & Willms, 1995, p. 321). As articulated in the equation, $Y_{ij} = \mu + P_{ij} + C_{ij} + S_{ij} + e_{ij}$, the conceptual framework is based on the work of Raudenbush and Willms (1995).

The outcome variable (Y) for student i in school j is a function of the grand mean (μ), school practice (P), school context (C), student background (S), and random error (e). School practice is commonly defined as "administrative leadership, curricular content, utilization of resources, and classroom instruction" (Raudenbush & Willms, 1995, p. 310). School context typically includes the demographic makeup of the student sample, the quality of teachers in a school, and other factors outside the control of administrators. Finally, student background is usually represented by measurable attributes (e.g., SES) and prior achievement.

The conceptual framework is based on several key caveats. First, student background, school practice, and school context all contribute to between- and within-school variation in the outcome variable (Raudenbush & Willms, 1995). Second, school context and school practice must be orthogonal if estimates of school effects are to be unbiased and consistent (Raudenbush & Willms, 1995). By addressing these methodological and statistical challenges, multilevel modeling is considered the appropriate method for estimating the effect of school practice on student achievement on high-stakes tests (Goldstein, 1997).

Rationale

This dissertation is designed to build on contemporary studies that have employed promising statistical models to analyze student performance on high-stakes tests. The rationale is that the application of multilevel modeling in SER can be enhanced through replicable equation building and efficient variable selection in an appropriate educational setting. To explore this notion, the procedures, findings, and limitations from three recent studies are discussed to generate a strong case for the need of this investigation. However, acknowledging only three studies is somewhat arbitrary, in that many prestigious researchers have contributed to the SER canon.

The Ma and Klinger (2000) study is very thorough in its description of the HLM process and unique in its inclusion of effect sizes and reliability data. The researchers utilized two-level HLM to explore the influence of student- and school-level variables on the performance of Canadian sixth-grade students on provincial assessments in mathematics, science, reading, and writing. They found that almost all of the variation in

mathematics (.89), science (.91), reading (.95), and writing (.91) achievement was within-schools and could be attributed to student-level variables (e.g., SES).

The present study extends the work of Ma and Klinger in a few ways. Rather than collecting data on less available inputs such as academic emphasis and parental involvement, more accessible school-level predictors (e.g., teacher quality) were considered to make the study replicable. Second, this study focused on American schools and students to account for the unique contextual influences that differentiate the U.S. educational system.

Similar to this dissertation, Yu and White (2002) used HLM to analyze the impact of school- and student-level variables on statewide assessment results for elementary school students. The researchers examined the effect of prior achievement, student and school SES, and teacher educational attainment on the achievement of sixth-grade students on mathematics, science, reading, and writing high-stakes tests. They found that 77% of the variance in student achievement was due to within-school variability and that individual SES and prior achievement accounted for most of this variance. The results also showed that unmeasured level-2 variables were responsible for some of the unexplained between-school variance in test scores.

The present study adds to Yu and White's research in several aspects. First, it assessed actual school performance relative to predicted test scores in addition to ranking schools on school effects estimates. Second, additional student-level predictors were specified to allow for more precise value-added models that can help mitigate the HST achievement gap (Coe & Fitz-Gibbon, 1998).

Wenglinsky (2002) employed structural equation modeling to examine the relationship between teacher quality and student achievement on the National Assessment of Educational Progress (NAEP) mathematics test. He found that some teacher quality inputs (e.g., teacher major, professional development) were significantly related to student test scores in eighth grade. Furthermore, Wenglinsky offered a template for the specification of school practice in hierarchical models by collecting and analyzing data on classroom instruction and assessment.

To build upon this research, data from statewide assessments were utilized rather than standardized tests to more closely approximate the relationship between school effects and student achievement. Furthermore, fourth-grade test scores were analyzed because focusing on the elementary school level allows for greater control of organizational factors (Wright, Horn, & Sanders, 1997).

Purpose

The purpose of the present study was to contribute a theoretically and empirically sound approach for analyzing and interpreting the results of high-stakes tests in Colorado. The main objective was to identify the most consistent, powerful, and accessible indicators of student achievement on the Colorado Student Assessment Program (CSAP) fourth-grade reading test. Specifically, hierarchical linear modeling (HLM) was employed to develop a replicable statistical model that predicts HST scores from school- and student-level inputs. A secondary goal was to investigate the underlying hypothesis for SER. Namely, that schools make a difference in student achievement because school practice (i.e., instruction, curriculum, organization) is responsible for the unexplained between-school variance on high-stakes test scores (McGee, 1997).

Research Question

The research question for this dissertation was complex associational in nature.

1. What combination of student- and school-level variables best predicts student achievement on the Colorado Student Assessment Program fourth-grade reading test?

Independent Variables

The present study attempted to measure school effects by controlling for exogenous demographic and contextual factors (Raudenbush & Willms, 1995). Thus, the most powerful, consistent, and accessible predictors of student achievement on high-stakes tests functioned as the independent variables for the study. These educational inputs were separated into student (level-1) and school (level-2) predictors. At the student level, the independent variables were socioeconomic status (SES), English as a Second Language (ESL) status, special education status, ethnicity, mobility, and prior achievement.

The most studied school-level predictor is teacher quality, which is typically measured by a combination of teacher certification, education, and experience (Laczko-Kerr & Berliner, 2002). In addition to these teacher characteristics, the level-2 independent variables were school SES and class size. The operational definitions for the selected student- and school-level predictors are described in Chapter 3.

Dependent Variables

To focus the research effort, HST results from Colorado elementary school students served as the educational output of interest. Specifically, the dependent variable was the total scale score from the 2002 CSAP fourth-grade reading test. The total performance level (i.e., unsatisfactory, partially proficient, proficient, advanced) also was

collected to allow for additional interpretations of the results. The CSAP fourth-grade reading test is administered in three 50-minute sessions during the spring semester and consists of multiple-choice and constructed-response questions.

Definitions

The following definitions are for the most commonly used terms in this dissertation.

Colorado Department of Education (CDE): State entity that is responsible for the administration, analysis, and reporting of the Colorado Student Assessment Program.

Colorado Student Assessment Program (CSAP): Statewide high-stakes tests in reading, writing, mathematics, and science that are aligned with content standards.

Hierarchical Linear Modeling (HLM): A complex associational research method that accounts for the nested and multilevel nature of educational data.

High-Stakes Testing (HST): The euphemism for large-scale assessments that hold serious consequences for schools, teachers, and students based on participation and performance.

Just for the Kids (JFTK): Under the auspices of the National Center for Educational Accountability, this project features an interactive website that presents adjusted HST results from numerous states including Colorado.

School Accountability Report (SAR): Annual report card available on the CDE website containing data on district expenditures, school characteristics, and student outcomes for every K-12 public school in Colorado.

School Effectiveness Research (SER): The domain of educational research that investigates whether school practice has an effect on student achievement after controlling for student characteristics and school context.

Delimitations

The major delimitation of the present study is that the secondary data are cross-sectional and do not have the stability or scope of data gathered in a longitudinal manner (Willms & Raudenbush, 1989). For example, new advances in longitudinal data collection have allowed researchers to estimate the effects of more intangible teacher characteristics including content knowledge and pedagogical skill (Goldhaber, 2002).

As for external validity, the sample includes a small percentage of districts in the Colorado public school system and is skewed toward urban and large suburban communities. Furthermore, the findings may have limited applicability for other state accountability systems, in that "other locations may have access to different variables, making it impossible to replicate [the] predicting equation" (Bingham, Heywood, & White, 1991, p. 201). It also is inappropriate to generalize the results from the high-stakes reading test to other content areas, as unique performance interactions might exist (Pituch, 1999). In addition, validity is threatened by focusing on only one grade, as school effects estimates may be different for other student levels.

Significance

The present study offers several important contributions to school effectiveness research. From a methodological perspective, the use of HLM to analyze secondary databases should encourage educational researchers to integrate appropriate statistical techniques with accessible and efficient sources of HST data. The practical significance

is that policy decisions based on HST can be made equitable for all educational stakeholders (Ma & Klinger, 2000; Yu & White, 2002). Specifically, this dissertation presents a predictive model that can be used to gauge student progress, identify improving schools, and highlight areas of concern. For example, by making the persistent inequalities in student achievement more transparent, additional resources may be targeted to the schools that need the most assistance (Ryan, 2002). Finally, this dissertation is particularly pertinent and timely in light of new federal guidelines advocating "scientifically-based" research.

Organization

This dissertation is presented in five chapters. Chapter 1 defines the research problem, describes the conceptual framework, and offers a rationale for the need and significance of the study. Chapter 1 also specifies the research question, variables, definitions, and delimitations that inform the research design. Chapter 2 features a review of the literature on school accountability, high-stakes testing, school effectiveness research, and educational inputs and outputs. Chapter 3 describes the research approach, sampling design, data collection, variable selection, methodology, and data analysis. Chapter 4 displays findings from the descriptive, assumption, and HLM analyses along with results for the school effects estimates. Finally, Chapter 5 discusses the conclusions, interpretations, implications, limitations, and recommendations from the study.

CHAPTER 2: LITERATURE REVIEW

This dissertation builds a case for improving the analysis and interpretation of high-stakes tests through a synthesis of the research on school accountability policy and practice. The Educational Resources Information Center (ERIC) database was accessed to locate relevant and timely journal articles, unpublished reports, and conference papers on high-stakes testing, school effectiveness research, and educational inputs and outputs. Based on the abundance of pertinent references, it was decided to consider only studies from 1990 to the present. However, some articles published before 1990 were incorporated to provide greater depth to the literature review. The search was further bounded by only including empirical research conducted with elementary school students. The selected studies represent the "best evidence" from SER based on the variety, quality, and suitability of achievement predictors, research designs, and statistical methods.

High-Stakes Testing

High-stakes tests are commonly defined as "direct measures of accountability for students, educators, schools, or school districts, with significant sanctions or rewards attached to test results" (Gordon & Reese, 1997, p. 345). Tests have high stakes for students when they determine grade advancement, graduation, or college admission (Darling-Hammond, 1991; Domenech, 2000). For teachers and administrators, high-stakes tests may impact job security, salary, and reputation (Haladyna et al., 1991; Popham, 2000). Schools and districts face a loss of autonomy or funding if they fail to

show adequate yearly progress in student achievement (Lindjord, 2001). While public opinion is generally favorable to the rigorous assessment of students and schools, there is little agreement on how this should be accomplished (Rose & Gallup, 2000).

Context

Standardized testing has been part of the U.S. educational system since the mid-1800s and has been an integral component of American public schools for the past 50 years (Haladyna et al., 1998; Walker, 2000). Historically, standardized testing was used to sort students by academic ability, place students into special education or gifted programs, and modify curricula and instructional planning (Madaus, 1991; Walker, 2000). This changed in the early 1970s when states began to attach high stakes (e.g., student graduation) to the results of standardized tests (Langenfeld, Thurlow, & Scott, 1996).

The 1980s witnessed a sharp rise in the use of high-stakes tests, as policymakers and educators responded to criticism of public schooling by raising academic standards and holding schools more accountable for student outcomes (Darling-Hammond, 1991; Walker, 2000). This trend continued unabated in the 1990s, in that HST became the primary vehicle for enacting educational reform (Gordon & Reese, 1997). The unprecedented growth of HST persists in the new millennium, as nearly every state utilizes these assessments for school accountability purposes (Hoffman, Assaf, & Paris, 2001). Furthermore, federal legislation now requires states to better align high-stakes tests with standards in specific content areas (Linn, Baker, & Betebenner, 2002).

Criticism

While politicians are quick to support high-stakes testing, many educational stakeholders are reticent or downright hostile about these assessments (Hoffman et al., 2001; Kohn, 2000). For example, a majority of teachers believe that HST results do not accurately reflect the skills or knowledge of students (Barksdale-Ladd & Thomas, 2000). Specifically, teachers argue that high-stakes tests are biased against disadvantaged students because they measure what a student brings to school rather than what they learn there (Coleman, 2000; Hess & Brigham, 2000; Jones et al., 1999). Teachers also indict HST for causing "frustration, burnout, fatigue, physical illness, misbehavior, and psychological distress" for students and educators (Smith & Rottenberg, 1991, p. 10).

Another common teacher criticism is that HST narrows the school curriculum (Domenech, 2000; Dounay, 2000) because of the enormous time and money being spent on test preparation in an already busy instructional schedule (Lattimore, 2001; Paul, 2002; Waite, Boone, & McGhee, 2001). As a result, teachers believe that students are not provided with the appropriate classroom experiences to be successful on these tests (Coleman, 2000). Perhaps the most cogent critique is that HST results do not provide sufficient guidance to educators, policymakers, or parents on how to improve student learning (Riddle Buly & Valencia, 2002; Ryan, 2002).

Although teachers articulate the most stinging criticisms of HST, parents and students are more vocal in their contempt for test-based accountability (Kohn, 2000). Many parents question the validity of HST results that do not match student grades, while others doubt the reliability of the proficiency designations for these tests (Dounay, 2000). The research indicates that parents believe HST places undue pressure and stress on

young children without improving their learning or achievement (Barksdale-Ladd & Thomas, 2000; Domenech, 2000). In a survey conducted in 2001, Public Agenda found that 60% of parents felt too much emphasis was being placed on one test (Paul, 2002).

This sentiment is shared by a majority of students who think it is unfair that one test can be used to determine graduation and promotion decisions (Madaus, 1991; Walker, 2000). Perhaps the most disturbing criticism is that students have little input into the development, implementation, and interpretation of school accountability measures (Waite et al., 2001). As a response to being disenfranchised, some students have taken the lead in speaking out against the unintended consequences of high-stakes testing (Kohn, 2000). This dissent has led to legal challenges that have forced politicians to reconsider assumptions underlying school accountability systems (Coleman, 2000; Walker, 2000).

Support

Despite persistent criticism and controversy, HST supporters in the private sector and the political realm have articulated strong arguments for test-based accountability. While the arguments originate from different agendas, the rhetoric of both camps has been similar. The most popular claim is that HST improves academic achievement because students rise to higher expectations and try harder when tests really count (Hess & Brigham, 2000). Another common refrain is that educators will improve their teaching if they are held more accountable for the results (Shepard, 1991). A corollary to these two assertions is that HST focuses classroom instruction because teachers and students are presented with specific and attainable goals (Madaus, 1991). A related contention is that high-stakes testing leads to a better alignment between curricula and academic standards (Hess & Brigham, 2000).

In addition, it is maintained that HST serves a diagnostic function so schools can identify instructional "best practices" (Riddle Buly & Valencia, 2002) and target the appropriate resources where needed (Walker, 2000). Furthermore, it is posited that HST promotes the impartial and equitable distribution of educational resources because it places smaller and larger districts on the same playing field (Madaus, 1991; Walker, 2000). However, school accountability policies that rely on high-stakes tests have resulted in implications that run contrary to the intentions of those who support this educational reform (Noble & Smith, 1994).

Consequences

According to Gordon and Reese (1997), "high-stakes testing has become the object rather than the measure of teaching and learning, with negative side-effects on curriculum, teacher decision making, instruction, student learning, [and] school climate" (p. 366). As a result, many teachers feel compelled to "teach to the test," which minimizes content areas like the arts (Dounay, 2000; Hess & Brigham, 2000), reduces opportunities for experiential instruction, and hampers the development of critical and analytical thinking skills in the classroom (Domenech, 2000; Smith & Rottenberg, 1991). These unintended consequences fall excessively upon certain populations of students, teachers, and schools (Waite et al., 2001). For example, the narrowing of instructional and assessment practices limits the opportunities for students with special needs to explore academic interests (Langenfeld, et al., 1996; Shepard, 1991).

As a result of increased administrative oversight and widespread media coverage, teachers often feel pressured into doing whatever is necessary to raise school test scores (Herman & Golan, 1993; Smith & Rottenberg, 1991). For example, there have been

numerous examples of egregious security breaches (Dounay, 2000) and blatant educator cheating (Ryan, 2002; Yettick, 2002). Additionally, some administrators have attempted to improve school achievement on high-stakes tests by exempting low-performing students through spurious special education placements (Darling-Hammond, 1991; Hess & Brigham, 2000).

School Effectiveness Research

Since the influential Coleman Report (Coleman et al., 1966), one of the most virulent educational debates has centered on whether schools have a significant effect on student achievement (Ascher & Fruchter, 2001; Mandeville & Heidari, 1988). This complex question has defined SER as the strand of educational inquiry that examines school differences on HST outcomes after statistically adjusting for input variables (Goldstein, 1997). School effectiveness research is based on the idea that "school effects" represent the impact of an educational practice or policy on student outcomes or the level to which attending a specific school influences student achievement (Raudenbush & Willms, 1995).

Proponents of SER believe that schools do make a difference in the academic achievement of students because schools with similar input variables produce different results (Strand, 1997). The foremost criticism is that SER has failed to demonstrate that school effects are responsible for the variance in HST results (Coe & Fitz-Gibbon, 1998). This ongoing debate is played out in studies that feature value-added, input-output, and multilevel modeling research designs.

Value-Added

When HST scores are regressed on one or more demographic variables, the "adjusted residuals are often known as 'value added' school estimates" (Goldstein, 1997, p. 383) because they measure the "boost" that a school contributes to student achievement after controlling for past performance and background characteristics (Yu & White, 2002). The most publicized example of value-added research is the Tennessee Value-Added Assessment System (Sanders, 1998). Critics of this approach note that there is little agreement on the most powerful and consistent predictors of school effectiveness (Holdaway & Johnson, 1993). Serious questions also have been raised about the integrity, stability, and consistency of variables utilized in value-added research (e.g., Crone, Lang, Franklin, & Halbrook, 1994).

Input-Output

This SER approach employs production functions to measure changes in a school's outputs based on changes in inputs (Glasman & Biniaminov, 1981; Wenglinsky, 2002). According to Hanushek (1989), "if the production function for schools is known, it is then possible to predict what would happen if resources were added or subtracted" (p. 45). Input-output analyses carry policy implications, as most educational initiatives are based on presumed relationships between school resources and student outcomes (Darling-Hammond, 2000). In addition, "econometric studies are probably more externally valid than small scale experiments since they arise from data on operating schools" (Nye et al., 2002, p. 202).

The main weakness of production functions is that the cumulative nature of education precludes measuring the "true" impact of educational inputs on student

achievement (Greenwald, Hedges, & Laine, 1996). Furthermore, "input-output analyses do not deal with characteristics of the dynamic and ongoing interrelationships between students and teachers or those among students themselves" (Glasman & Biniaminov, 1981, p. 509). Lastly, numerous meta-analyses of production function studies have been inconclusive because of differences in the definitions, quantity, and quality of educational input variables (Bernstein, 1990; Wenglinsky, 2002).

Multilevel Modeling

Previous SER studies have employed ordinary least squares (OLS) regression to estimate the effect of educational inputs on student achievement (Greenwald et al., 1996). However, OLS analyses typically suffer from aggregation bias and the inability to measure interrelationships among predictor variables (Wenglinsky, 2002). Adjusting for these factors has been extremely complicated because school and student data often are in multilevel form (Bingham et al., 1991; de Leeuw & Kreft, 1995). Instead of treating educational data as single level, multilevel modeling accounts for its hierarchical and nested structure (Ma & Klinger, 2000; Raudenbush & Bryk, 1986; Wenglinsky, 2002).

The most common multilevel modeling approach is HLM, which is designed to control for variables that might confound school-level interpretations of student achievement data (Schafer, Yen, & Rahman, 2000). Specifically, HLM "allows researchers to adjust for multiple sources of bias (i.e., pupil and school differences) in their comparisons and avoid problems (i.e., smaller than warranted standard errors) associated with statistical inference for fixed effects that affects traditional approaches" (Pituch, 1999, p. 191). Hierarchical linear modeling has the potential to be the basic paradigm for future research in education, sociology, and psychology because it provides

a conceptual framework for understanding the dynamics of student learning (Bryk & Raudenbush, 1988). Ma and Klinger (2000) add that HLM facilitates the identification of exemplary practices and policies by providing more accurate statistical estimations.

Educational Inputs

Educational inputs are student- and school-level variables that are hypothesized to have a significant relationship with high-stakes test scores. Educational inputs also are referred to as predictors, indicators, and independent variables. Even though the SER evidence base is filled with hundreds of variables that are associated with student achievement on high-stakes tests (Greenwald et al., 1996; Wright et al., 1997), there is no formula for selecting educational inputs (Centra & Potter, 1980). The key challenge is to identify the most relevant indicators that vary among schools (Goldstein, 1997). However, predictors often are selected based on the availability of data and not on the specific research question being asked (Hanushek, 1986).

Despite the recommendation from Willms and Raudenbush (1989) that SER studies should include community-level variables, district-level inputs were not included because of a lack of accessibility and consistency. For example, per-pupil expenditures (PPE) have been the most studied district-level input. Although some researchers have found PPE to be a significant predictor of HST scores (e.g., Greenwald et al., 1996), others have found no relationship between educational spending and student achievement (e.g., Hanushek, 1989).

Student-level

The most popular student-level predictors in SER are socioeconomic status, ethnicity, and prior achievement. According to Ma and Klinger (2000), individual SES

has a significant absolute effect on student achievement and a greater relative influence than all other level-1 variables. Caldas (1993) found that student ethnicity was a powerful and consistent predictor of academic achievement in the primary grades. Many researchers have found prior achievement to be a strong and unfailing predictor of student performance on high-stakes tests (e.g., Bernstein, 1990; Goldstein, 1997; Mandeville & Heidari, 1988; Yu & White, 2002). Ryan (2002) argues that prior achievement can serve as the primary student-level predictor because it "provides some control for differences in SES while leaving factors that teachers and schools can influence" (p. 457).

The research is less supportive of student mobility as a reliable level-1 predictor of academic achievement on high-stakes tests. Johnson and Lindblad (1991) found that students who moved within a school district performed significantly worse on standardized achievement tests than did nonmobile students. However, other researchers have concluded that student mobility does not significantly correlate with student achievement (Heinlein & Shinn, 2000; Jennings, Kovalski, & Behrens, 2000).

Special education status and ESL status rarely are used in SER because the variables often are unavailable at the student level. This lack of accessibility is compounded because these groups of students have traditionally been left out of high-stakes testing programs (Hess & Brigham, 2000). As a result, there is an absence of research on how HST affects students with special needs (Langenfeld et al., 1997). Furthermore, students receiving special education or ESL services typically take high-stakes tests with accommodations, which complicates statistical comparisons with other student groups (American Educational Research Association [AERA], 2000).

School-level

Unlike student-level data, it is easier to access reliable information on level-2 inputs (Bernstein, 1990; Raudenbush & Willms, 1995). For example, there has been a spate of recent studies on the relationship between teacher quality and student achievement (Darling-Hammond & Youngs, 2002; Goldhaber, 2002; Sanders, 1998). Other popular school-level variables include school SES, parental involvement, academic emphasis, teacher attendance, and class size.

Overall, the research supports the specification of teacher quality indicators as level-2 predictors of student achievement on high-stakes tests (e.g., Ascher & Fruchter, 2001; Wright et al., 1997). According to Wenglinsky (2002), the effects of "teacher characteristics are comparable in size to those of student background, suggesting that teachers can contribute as much to student learning as the students themselves" (p. 1). Darling-Hammond (2000) argues that teacher certification and teaching in field are the most powerful school-level predictors of student achievement on high-stakes tests. Laczko-Kerr and Berliner (2002) found that students of certified and more experienced teachers significantly outperformed students with undercertified and less experienced teachers. However, some researchers argue that teacher effects are actually quite weak (e.g., Hanushek, 1986; Lankford, Loeb, & Wyckoff, 2002). For instance, Goldhaber (2002) found that some widely studied teacher quality indicators (e.g., education level) were not powerful predictors of student achievement on high-stakes tests.

School SES is perhaps the most consistent level-2 input, with higher rates of poverty predicting lower student achievement (Battistich, Solomon, Kim, Watson, & Schaps, 1995; Ma & Klinger, 2000; Yu & White, 2002). In a study conducted by the

Piton Foundation, students from low-SES backgrounds performed significantly better on high-stakes tests when placed in high-SES schools (Gottlieb, 2002). However, Sanders (1998) reported that cumulative gains on statewide assessments in Tennessee were unrelated to the mean SES of elementary schools.

As defined by participation in school activities, greater parental involvement in the primary grades is significantly correlated with higher student test scores even after controlling for student-level inputs (Griffith, 1996; Oakes, 1989). The academic emphasis of a school, as measured by teacher perceptions of the importance of academics, is a significant predictor of student achievement (Goddard et al., 2000).

According to the research, teacher attendance and class size are inconsistent predictors of HST scores. Although Harter (1999) found that substitute teacher pay had a strong negative relationship with student achievement, teacher absenteeism has failed to produce statistically significant effects on student test performance (Gottlieb, 2002; Norton, 1998). In a meta-analysis of class size research, Hanushek (1997) found that less than 15% of studies showed statistically significant effects on student achievement. However, Darling-Hammond (2000) found that smaller student-teacher ratios were associated with higher student achievement on fourth-grade reading tests.

Educational Outputs

Educational outputs are student academic outcomes that are commonly represented by scores from standardized achievement tests and statewide assessments. With the multitude of potential accountability measures, it is no surprise that a lack of agreement exists regarding the most appropriate educational outputs for SER. While the public supports the use of performance assessments and classroom portfolios to measure

student achievement, economic, logistical, and political constraints compel most states to rely solely on high-stakes tests (Rose & Gallup, 2000; Ryan, 2002). As a result, SER studies typically consider evidence from norm- and criterion-referenced standardized achievement tests and standards-based statewide assessments (Shepherd, 1991).

Standardized Achievement Tests

Standardized achievement tests measure student aptitude or achievement and are administered and scored in a systematic and a priori manner (Popham, 1999). Policymakers favor standardized tests because they are thought to be predictors of future student success in the academic and vocational domains (Hanushek, 1989). Unlike statewide assessments that hold high-stakes for teachers and schools, standardized achievement tests often are used for decisions about student tracking, retention, and graduation (Smith & Fey, 2000). However, standardized tests are criticized for psychometric and interpretative weaknesses (Coleman, 2000) that undermine their appropriateness for evaluating school quality (Popham, 1999).

Psychometrics. Persistent concerns over the reliability and validity of standardized tests have diminished their usefulness in measuring student achievement (Smith & Fey, 2000). This is especially true for students with disabilities, in that standardized tests measure characteristics extraneous to the intended construct (AERA, 2000). Furthermore, students from lower socioeconomic backgrounds are less likely to have the resources needed to succeed on standardized tests that often tap nonacademic learning (Popham, 1999). In addition, the generalizability of standardized tests has been compromised because results for students with special needs are frequently excluded in school accountability reports (Langenfeld, et al., 1997).

Interpretations. The interpretation of standardized achievement test scores is perhaps the most contentious issue in the educational evaluation and measurement field (Domenench, 2000). According to Haladyna et al. (1998), it is appropriate to interpret standardized test results for curriculum evaluation and instructional grouping purposes. Some researchers argue, however, that these tests should not serve as the sole basis for high-stakes decisions about students or schools (e.g., Thompson, 2001). Educational researchers also must do a better job of educating parents and politicians about the proper contextual interpretation of standardized test scores because of the complexity inherent to student achievement (Gordon & Reese, 1997; Haladyna et al., 1991).

Statewide Assessments

Although similar in format, statewide assessments differ from standardized achievement tests because they are aligned with state content standards and are designed primarily for school improvement purposes (Yen & Ferrara, 1997). For example, statewide assessments are used to raise public awareness about educational issues and to motivate educators and students to improve their performance (Haertel, 1999).

Psychometrics. According to Haladyna at al. (1991), the primary psychometric concern is test score pollution arising from the high-stakes nature of statewide assessments. Specifically, comparing districts through the media (Nolen, Haladyna, & Haas, 1992), connecting student results to funding decisions (Domenech, 2000), and teaching to the test (Haladyna at al., 1991) contribute to test score pollution. Similar to standardized tests, statewide assessments have psychometric weaknesses that disproportionately impact students of color, English language learners, and students with disabilities (Haladyna et al., 1991; Hoffman et al., 2001).

Interpretations. Because statewide assessments rarely have convergent validity with other student achievement measures, conclusions about HST data often are inaccurate (Linn et al., 2002). Specifically, interpreting statewide assessment results for accountability purposes is problematic because testing proficiency does not always generalize to academic success (Haertel, 1999). In addition, comparing student achievement on statewide assessments without controlling for school context yields biased interpretations (Gordon & Reese, 1997). As a result of these limitations, Burns (1998) argues that statewide assessments may not be appropriate for high-stakes decisions about students or schools.

Colorado Student Assessment Program

Since 1997, the Colorado Student Assessment Program has been the centerpiece of a statewide accountability system that requires the annual testing of students in grades 3-11 (Linn & Haug, 2002). The CSAP is designed to assess student achievement of the Colorado Model Content Standards through paper-and-pencil tests comprised of multiple-choice and constructed-response questions (Linn & Haug, 2002). Colorado students took more than 1.2 million of these exams in 2002, as the number of CSAP tests increased to 26 with the addition of new assessments in mathematics and science (Whaley, 2002). The CSAP was designed to provide instructional and curricular feedback to participating schools and educators (Kiplinger, 1998). However, the CSAP is primarily used as a trigger for high-stakes consequences for schools if student achievement declines over a period of time (Hubler, 2002).

Psychometrics. Kiplinger (1998) found that the CSAP fourth-grade reading test had sufficient construct validity, as there was alignment between the assessment and the

benchmarks for three of the four reading standards. Although scores for all students are on the rise, the biggest psychometric concern is the persistent and large achievement gap for Hispanic/Latino and African-American students (Nick, 2002; Whaley & Hubler, 2002). Another threat to validity arises from students prevented from completing CSAP tests due to "extreme frustration" or parental refusal (Whaley & Hubler, 2002).

CSAP tests have demonstrated high internal consistency, as evidenced by an alpha of .93 for the fourth-grade reading exam (Kiplinger, 1998; Linn & Haug, 2002). The reliability of CSAP data is threatened by the increase in "no scores" from chronic student absenteeism, teacher cheating, and administrative miscues (Yettick, 2002). Longitudinal changes in CSAP test scores also are unreliable because of measurement error (e.g., regression to the mean), sampling error (e.g., greater variability in small schools), changes in student demographics, and nonpersistent factors that affect student cohorts (Linn & Haug, 2002). To illustrate, Linn and Haug (2002) found that schools with relatively high baseline scores showed smaller gains than did schools with lower initial student proficiency levels. Furthermore, schools with strong gains from year one to year two showed a decline in year three while the opposite was true for schools that had an initial decrease in student scores.

Interpretations. The main interpretative weakness of the CSAP is that it relies on narrowly focused performance indicators to compare schools rather than identifying strategies for explaining school differences (Goldstein, 1997). Interpretations are further muddied when so-called high and low achieving schools implement the same types of instruction and teacher training while receiving similar funding (Hubler, 2002). Consequently, "it is likely to be a mistake to assume that the practices of the schools

recognized as outstanding are ones that should be adopted by other schools" (Linn & Haug, 2002, p. 34).

The interpretations of CSAP results are further compromised by the reliance on cross-sectional and successive group designs for comparisons of student achievement (Linn & Haug, 2002). The lack of consistency in the annual performance of schools results in a foggy picture in which schools improve one year and regress the next (Hubler, 2002). This volatility supports Linn and Haug's (2002) contention that some schools are "being recognized as outstanding and other schools identified as in need of improvement simply as the result of random fluctuations" (p. 35).

Summary

With the recent tidal wave of federal and state school accountability policies, the pressure is on all educational stakeholders to raise the bar on student achievement. Supporters argue that HST positively impacts academic achievement because teaching and learning is improved when tests actually mean something. Detractors claim that HST eliminates deeper student learning because teachers are forced to focus on basic skills and test taking techniques. On balance, the original conception of HST as a way to align classroom instruction with high academic standards has been transformed into a "one size fits all" assessment regime. A major implication from a system in which low test scores reflect poorly on schools is that educators fear reprisals on wages, job security, and autonomy. As a result, some teachers are forced to forsake the curriculum in favor of test preparation while others resort to illegal or unethical actions to protect themselves and their students. There also are unintended consequences for students of color and students with disabilities because of psychometric and interpretative weaknesses inherent to HST.

In response to the HST movement, school effectiveness research has taken center stage in the debate over whether schools really make a difference in the academic achievement of students. Supporters contend that schools with similar educational inputs produce different student outcomes, while opponents believe SER is severely undermined by the instability of school effects estimates and the unreliability of high-stakes tests. To answer this complex question, researchers have used value-added and input-output analyses to study the relationship between educational inputs and student achievement on high-stakes tests. However, the use of OLS regression to account for variables at multiple levels has resulted in sampling, aggregation, and measurement biases. The most promising statistical methodology is hierarchical linear modeling because it can account for the nested structure of educational data.

To have theoretical and practical significance, SER studies must incorporate the most consistent and accessible predictors of student achievement on high-stakes tests. At the student-level, the most powerful educational inputs are socioeconomic status, prior achievement, and ethnicity. The most popular and controversial school-level variable is teacher quality. Other significant level-2 predictors include school SES, parental involvement, and academic emphasis.

As for educational outputs, SER has relied on standardized achievement tests and statewide assessments to provide evidence of student learning. Ongoing concerns regarding the analysis and reporting of these outcome measures have limited their effectiveness in measuring and comparing school performance. For example, the Colorado Student Assessment Program has been criticized for ranking schools based on raw proficiency levels without controlling for student characteristics and school context.

CHAPTER 3: METHODS

Educational research has been hampered by an inability to account for the multilevel nature of school data. This has resulted in ambiguous and inconsistent evidence regarding factors that influence student achievement (Bernstein, 1990). Because educational systems have a hierarchical structure (i.e., students nested within schools), a well-designed SER study must use appropriate statistical techniques to analyze the results of high-stakes tests (Ma & Klinger, 2000; Raudenbush & Willms, 1995). For the present study, HLM was selected as the appropriate methodology to simultaneously estimate the effects of student- and school-level variables on student achievement (Yu & White, 2002). This chapter details the research approach, sampling design, data collection, variable selection, and hierarchical linear modeling analysis.

Research Approach

This investigation was based on similar SER studies (e.g., Yu & White, 2002), as an associational approach based on quantitative methods and inferential statistics was used to make predictions about HST outcomes (Gliner & Morgan, 2000). The approach was appropriate for the research question because of the difficulty in conducting an experimental design when investigating the relationship between input variables and school outputs (Bernstein, 1990; Raudenbush & Bryk, 1986). As described by Raudenbush and Willms (1995), "studies of school effects are quasi-experiments, and estimation requires some attempt to identify and control for exogenous covariates that are confounded with the 'treatment' provided by the school" (p. 310).

Sampling Design

Educational outcomes represent both the individual and combined efforts of students and teachers, so test scores were analyzed at the student level but interpreted at the school level. Ultimately, students were the participants because their test scores were used as the dependent variable in the HLM analysis.

Sampling Frame

The theoretical population is elementary schools in the U.S. who are participating in high-stakes testing. It is appropriate to focus on the primary grades to control for organizational issues that may influence HST scores (Laczko-Kerr & Berliner, 2002) and to allow for more consistent measurement of educational inputs (Goddard, Sweetland, & Hoy, 2000). Elementary schools also were chosen because demographic and HST data often are more complete at this level (Gottlieb, 2002). The lack of comparability between school accountability systems coupled with differing regional, state, and local influences on HST precluded sampling from the entire theoretical population.

The researcher's geographic locale constrained the accessible population to be Colorado public elementary schools. Specifically, the sampling frame consisted of schools that participated in the Colorado Student Assessment Program during the 2001-2002 school year. These schools were accessible because their districts participated in the JFTK project and expressed a willingness to participate in related studies. Out of the 17 school districts in the sampling frame, three districts encompassing 168 elementary schools and 8,400 fourth-grade students did not consent to participate in the study before data collection began.

Final Sample

The selected sample consisted of 14,699 students from 334 elementary schools representing 14 Colorado school districts. To accurately measure school effects, only students who attended the same elementary school for at least two years were included in the sample. During the data cleaning process, all 37 schools from one school district were eliminated because of missing student-level data. In addition, six schools from a second district and 16 from a third were deleted because of data quality concerns. Fourteen charter and newly opened schools were removed from the sample because of insufficient outcome data. Seven schools with less than 10 students each were eliminated because of the unreliability of such a small sample (Schafer et al., 2000). Finally, 21 students who received the lowest possible score on either the CSAP third-grade or fourth-grade reading test were deleted from the sample for statistical reasons.

The final sample contained 12,936 students from 254 elementary schools representing 11 school districts in Colorado. The elimination of three districts, 80 schools, and 1763 students was not problematic for the two-level HLM analysis because there was sufficient power at the student and school levels. Although convenience samples are common for this type of research design (e.g., Laczko-Kerr & Berliner, 2002), the resultant threat to external validity cannot be overlooked. Specifically, the lack of random selection limits the generalizability of the predictive model for the approximately 1700 elementary schools and 180 school districts in Colorado.

Data Collection

To access the appropriate student- and school-level variables, data were collected from the JFTK database, a secondary dataset on teacher quality, and the school

accountability reports available on the CDE website. The use of secondary sources allowed for the efficient collection of relevant educational input and output data. Furthermore, collecting actual HST data was essential for any internal validity claims regarding the effect of school practice on student achievement.

Secondary Databases

As an alternative to more time consuming, costly, and invasive primary data collection techniques (e.g., administrator interviews), secondary databases allow researchers to utilize previous high quality research in a new context. For this study, the secondary databases were understood to be complete and accurate in their representation of student test scores and educational input variables. The JFTK database provided the CSAP fourth-grade reading test scores, school SES, and all of the level-1 inputs (i.e., prior achievement, continuous enrollment, ethnicity, SES, ESL status, and special education status). Teacher certification and teacher education data were collected from a database complied by a researcher at Mid-Continent Research for Education and Learning (McREL) during a recent study on the supply and demand of teachers in Colorado. The secondary data from the JFTK and McREL databases were transmitted online in Microsoft Excel file attachments.

These secondary databases were supplemented with information contained in the Colorado school accountability report. This data collection approach is similar to McGee's (1997) study, in which the Illinois state report card provided the independent variables used in the analysis. Among other level-2 variables, the SAR presents data on teacher experience and class size. These data were accessed from Excel spreadsheets posted on the CDE website (http://www.cde.state.co.us/index_assess.htm).

Archival data are rarely collected for research purposes, so investigators often are compelled to accept variables as defined and to analyze databases with missing information (Jennings et al., 2000). Fortunately, the secondary databases were accessed from credible sources and contained little missing or problematic data. Another concern is that limited researcher control during the original data collection could potentially impact the results or interpretations derived from secondary sources. Specifically, the reliability of secondary data depends on having consistent definitions for variables that are collected at the same level and over the same time period.

Human Subjects Research

The human subjects research approval process was simplified because there was no primary data collection and student and school identifiers were removed from the secondary data. As displayed in Appendix A, consent forms were sent in February 2003 asking school districts for permission to release student- and school-level data from the JFTK database. Superintendents from 14 of the 17 selected districts signed and returned consent forms authorizing the release of these data. Upon receipt in June 2003, the 14 consent forms were sent to Just for the Kids. As shown in Appendix B, letters of cooperation were secured from JFTK granting access to the data for the participating districts and from CDE and McREL providing access to the teacher quality database. The regulatory compliance office at Colorado State University granted final approval after the human subjects research application was submitted (see Appendix C).

Variable Selection

Although variable selection has been a problematic component of SER, there is strong agreement that a predictive model must incorporate input variables that meet both

statistical and theoretical criteria (Coe & Fitz-Gibbon, 1998; Goldstein, 1997). There also is support for selecting variables that increase explicatory power without contributing to multicollinearity (Darling-Hammond, 2000). As displayed in Table 1, the most popular, powerful, consistent, and accessible level-1 and level-2 predictors of student achievement on high-stakes tests in the primary grades were considered.

Table 1

Predictors of Student Achievement on High-Stakes Tests in the Primary Grades

	Student-level	School-level
Popular	Family income	Teacher certification
	Free/reduced lunch status	Teacher education
	Parental education	Teacher experience
	Family size	Teacher attendance
	Parental employment	Teacher turnover
	Ethnicity	School SES
	Gender	Percentage of special ed. students
	Special education status	Percentage of ESL students
	ESL status	Class size
	Mobility	Academic emphasis
	Prior achievement	Parental involvement
Powerful And Consistent	Family income	Teacher certification
	Free/reduced lunch status	Teacher experience
	Parental education	School SES
	Ethnicity	Academic emphasis
	Prior achievement	Parental involvement
Accessible	Free/reduced lunch status	Teacher certification
	Ethnicity	Teacher experience
	Prior achievement	Teacher education
	Special education status	School SES
	ESL status	Class size
	Mobility	

Level-1

Student-level attributes such as SES, ethnicity, and prior achievement are the norm in SER because of their predictive power regarding student achievement on high-stakes tests (Nye et al., 2002). The strongest indicator is individual SES, which is usually measured by some combination of family income, free/reduced lunch status, parental education and employment, and family size (Gottlieb, 2002; McGee, 1997; Stringfield & Herman, 1996). In the present study, student SES was operationalized as a dichotomous variable based on eligibility for free/reduced lunch benefits.

Originally collected as a nominal variable with five levels, ethnicity was recoded as a dichotomous variable (i.e., white, minority) to allow for easier interpretation of the results. Prior achievement, which is usually measured by HST scores attained previous to the school year under consideration (Yu & White, 2002), was included as a continuous level-1 predictor. Specifically, CSAP third-grade reading test results from 2000-2001 served as the prior achievement variable. Special education status and ESL status were nominal variables with two levels based on student eligibility for these services. Student mobility was measured by continuous enrollment on campus and was operationalized as a dichotomous variable (i.e., two years, three years or more). Although it was collected from the JFTK database, gender was not included because of its inconsistent relationship with HST scores (Ma & Klinger, 2000).

Level-2

The most popular, powerful, and consistent predictors at the school-level are teacher quality (Greenwald et al., 1996), school SES (Harter, 1999), parental involvement (Griffith, 1996), and academic emphasis (Goddard et al., 2000). However, data

measuring parental involvement and academic emphasis are unavailable from most secondary data sources. Thus, the most accessible level-2 predictors of student achievement are teacher quality and school SES.

Teacher quality is defined as a composite of teaching experience, educational attainment, and certification status (Ascher & Fruchter, 2001; Laczko-Kerr & Berliner, 2002). In the present study, teaching experience was operationalized as the average years of experience for teachers in a school. Teacher education was measured by the percentage of teachers in a school with a graduate degree. Teacher certification was defined as the percentage of teachers in a school who were fully certified or had a master teacher designation. Although the SAR includes other teacher quality variables (e.g., attendance, turnover), these indicators are weak predictors of student achievement in the primary grades. Additionally, the percentage of teachers teaching in field was not appropriate at the elementary level.

Data on the percentage of students qualified for free/reduced lunch benefits in a school were collected from the JFTK database. Many researchers have specified school SES as the strongest predictor of student achievement in the primary grades (e.g., Caldas, 1993). Although considered to be a less consistent predictor, class size was defined in the SAR as a ratio between the number of fourth-grade students and teachers in a school.

Hierarchical Linear Modeling

Contemporary SER typically employs multilevel modeling to determine the relative performance of schools on high-stakes tests (Nye et al., 2002). Specifically, HLM has demonstrated sufficient efficiency and stability to answer questions about the effects of student- and school-level variables on HST results (Ma & Klinger, 2000; Schafer et

al., 2000). Hierarchical linear modeling offers solutions to some of the most vexing problems inherent to multilevel educational data (Bryk & Raudenbush, 1988). For example, HLM has resolved misestimated standard errors by incorporating unique school effects to account for the dependence of outcomes among students in the same school (Bryk & Raudenbush, 1988). As for heterogeneity of regression, HLM accounts for the between-school variance on student characteristics and outcomes by estimating a separate regression coefficient for each school (Bryk & Raudenbush, 1988).

Theory

The central tenet of HLM is that "individuals grouped together in some way (e.g., within the same school) experience a similar environment and are more likely to have things in common than children in different groups" (Schagen, 1991, p. 216). On the other hand, HLM assumes that educational interventions do not have a constant effect on all students, and thus the within-school variability of test scores must be considered (Centra & Potter, 1980). Perhaps the most controversial notion of HLM is that "school-level residual terms are assumed to represent the effects of school practice on the student outcome" (Pituch, 1999, p. 194) after controlling for student attributes and school context variables that influence academic achievement (Bingham et al., 1991). A major theoretical assumption is that input variables can be measured at a higher level of aggregation than outcome measures (Raudenbush & Bryk, 1986).

de Leeuw and Kreft (1995) challenge many of these assumptions by raising questions about the theoretical framework of HLM. The residual view is countered by the assertion that causality between school effects and HST scores cannot be demonstrated because of unmeasured interactions (Goldstein, 1997; Pituch, 1999). Although it never

can be adequately proven that residuals are solely attributable to school effects (Bingham et al., 1991), HLM provides both theoretical and practical insight into the complex dynamics of student achievement on high-stakes tests.

Process

The application of HLM in school effectiveness research is accomplished through the creation of within- and between-group models that allow for simultaneous exploration of the structural relationships of schools (Battistich et al., 1995; Bernstein, 1990). The first step is to create an unconditional model (i.e., a model with no predictors or control variables) to determine the within- and between-school variation on the dependent variable (Battistich et al., 1995; Goddard et al., 2000; Ma & Klinger, 2000). The next step is to develop a within-school equation that describes the relationship between test scores and student characteristics within each school (Raudenbush & Bryk, 1986; Willms & Raudenbush, 1989). In effect, a distinct regression equation is estimated for each school in the sample (Raudenbush & Bryk, 1986). After creating the level-1 model, a school-level model that represents the influence of school context is defined. In this between-school equation, the within-school parameters are now the outcomes that are predicted by the level-2 inputs and random error intercepts (Willms & Raudenbush, 1989).

Data Analysis

The data analysis procedures included selecting software, cleaning secondary databases, creating required files, generating descriptive statistics, testing assumptions, specifying models, and analyzing school effects results.

Software

Based on expert recommendations, two hierarchical linear modeling computer programs, HLM5 and MLwiN 1.10, were considered. To select the appropriate software, the Kreft, de Leeuw, and van der Leeden (1994) matrix was examined. Furthermore, the websites for HLM (www.ssicentral.com) and MLwiN (http://multilevel.ioe.ac.uk) were reviewed. The decision to purchase HLM5 was based on its ease of use, interactive interface, abundant significance tests, thorough documentation, and reasonable cost. Furthermore, the researcher received extensive training during two workshops on the application of HLM5 in school effectiveness research. Although Pituch (1999) criticized past HLM versions for not providing standard error estimates of level-2 residuals, the current version offers results with and without robust standard errors. In addition, the software allows for model estimation from incomplete data, faster and more efficient importing of raw data, and latent variable analysis.

Procedures

The secondary data files were imported into the Statistical Package for the Social Sciences (SPSS), which is the standard data analysis and management program for this type of study. The files were checked for missing and illegal values before school and student records were removed because of these issues. Several new variables were computed from the original variables contained in the secondary databases. A level-1 SPSS file was created from the JFTK data while a level-2 SPSS file was formed by merging variables from the McREL, SAR, and JFTK databases.

To protect confidentiality, the identification numbers of participating schools and districts were changed, and the linked list containing the original identifiers was

destroyed. The resultant data files were saved on secured email servers with password-protected access and CD-RW disks, which were stored in a locked location at the researcher's home office. The codebook included in Appendix D was generated for both SPSS files, so that variable descriptions, labels, and values would be recorded. After creating the two SPSS files, preliminary descriptive analyses were conducted and univariate and multivariate assumptions were tested at the student and school levels.

The next procedure was to construct a Sufficient Statistics Matrices (SSM) file in HLM5 based on the student- and school-level SPSS files. The SSM file enables faster computation by merging and summarizing all of the data into one file (Raudenbush, Bryk, Cheong, & Congdon, 2001). Command files then were created in HLM5 for the one-way ANOVA model, means-as-outcomes model, random coefficient model, and intercepts-and-slopes-as outcomes model. The command files contain the level-1 and level-2 equations that specify the hierarchical linear model to be tested.

After running the HLM analysis for the intercepts-and-slopes-as outcomes command file, a residual file was created and formatted into a standard SPSS file. Data from the residual file were used to analyze cross-level assumptions. The final procedure was the formation of a school effects file, which was comprised of level-2 data from the SPSS file, residual terms and fitted values from the residual file, and new variables and ranks generated from the school effects equation employed in the analysis.

Application

For this study, a two-level hierarchical linear model was specified. The predictive equations at the student and school levels represented the hypothesized relationships between the input variables and HST outcome. A district level was not specified because

of the statistical complexity of adding a third level in HLM and the lack of power with only 11 participating school districts. The model specification was designed as a "step up" approach, in which the level-1 model is created before considering the influence of school-level predictors (Raudenbush & Bryk, 2002).

According to Raudenbush and Bryk (2002), the "early phases of model building involve an interplay of theoretical and empirical considerations. The substantive theory under study should suggest a relatively small number of predictors for possible consideration in the level-1 model" (p. 256). Thus, the most powerful and consistent student-level variables in the SER evidence base were selected for inclusion in the model. The main empirical decisions were whether to specify level-1 variables as random, fixed, or nonrandomly varying and whether to delete a predictor (Raudenbush & Bryk, 2002).

The specification of the level-2 model was again theory-based, as school-level predictors were selected for their hypothesized relationship with student achievement (Raudenbush & Bryk, 2002). From an empirical standpoint, Raudenbush and Bryk (2002) argue against a backward approach, in which all possible level-2 predictors are entered before the nonsignificant variables are removed. As with a level-1 predictor, the inclusion of a school-level input is based on the strength and significance of its fixed effect (Raudenbush & Bryk, 2002). Thus, level-2 predictors with both theoretical and empirical support were specified in the school-level equation for the hierarchical linear model.

CHAPTER 4: RESULTS

Before the findings from the HLM analysis and the school effects estimation are presented, results from preliminary analyses conducted with the level-1 (student), level-2 (school), and residual SPSS data files are described. Subsequent to cleaning the data, variables at both levels were explored to determine the extent of missing data. There was only one missing value at level 1, which resulted in the deletion of the student record that was lacking ethnicity information. The total absence of missing data at level 2 was essential because the HLM5 software requires complete school-level data.

Descriptive Statistics

Descriptive statistics were generated for student characteristics, school context, and the outcome measure after the data were checked for potential outliers. Box plots and extreme value tables indicated the presence of outliers for the prior achievement variable and the outcome measure. There were 11 students who recorded the lowest possible score on the third-grade test (i.e., 150) and ten students with the lowest possible score on the fourth-grade test (i.e., 180). It was assumed that these students had just signed their names without taking the test because the 11 third-graders scored an average of 388 on the fourth-grade test and the ten fourth-graders averaged 375 on the third-grade test. The elimination of these 21 cases removed potentially confounding outliers from the HLM analysis.

Student Characteristics

The final sample consisted of 12,936 fourth-grade students from 254 elementary schools in 11 Colorado school districts. The number of fourth-grade students in each school varied from a minimum of 11 to a maximum of 112 with an average of 50.93. As for gender, there were 6608 female students (51.1%) and 6328 male students (48.9%). The ethnic composition was 59.9% White (n = 7744), 27.0% Hispanic (n = 3496), 8.7% African-American (n = 1128), 3.3% Asian/Pacific Islander (n = 421) and 1.1% Native American/Alaskan Native (n = 147). The frequencies for the other student-level variables are displayed in Table 2.

Table 2

Demographic Characteristics of Participants (N = 12,936)

Characteristic	*n*	%
Ethnicity		
White	7744	59.9
Minority	5192	40.1
Free/Reduced Lunch Status		
Not eligible	8423	63.7
Eligible	4693	36.3
English as Second Language Status		
Not eligible	12047	93.1
Eligible	889	6.9
Special Education Status		
Not Eligible	11436	88.4
Eligible	1500	11.6
Continuous enrollment on campus		
3 years or more	10566	81.7
2 years	2370	18.3

A cursory check on the generalizability of this sample was made through comparisons with demographic data for the Colorado fourth-grade test-taking population from the 2001-2002 school year. According to the CDE website, a total of 55,446 students took the CSAP fourth-grade reading test, so the sample represented 23.3% of all test takers from that year. The population was comprised of 66% white and 34% minority students, as compared with 60% and 40% respectively in the sample. There were slightly more males in the statewide population (51.1%) than were in the sample (48.9%). The sample was underrepresented in students eligible for special education (11.6% compared to 15.4%) and ESL services (6.9% compared to 10.2%).

School Context

As displayed in Table 3, there was a relatively large disparity in the number of schools (3 to 87) and students (112 to 3461) per participating Colorado school district.

Table 3

District-Level Distribution of Schools (N = 254) and Students (N = 12,936)

District	Schools		Students	
	n	%	*n*	%
A1	28	11.0	1728	13.4
B2	7	2.8	300	2.3
C3	5	2.0	298	2.3
D4	31	12.2	2372	18.3
E5	87	34.3	3461	26.8
F6	13	5.1	577	4.5
G7	25	9.8	1273	9.8
H8	18	7.1	874	6.8
I9	22	8.7	1027	7.9
J10	3	1.2	112	.9
K11	15	5.9	914	7.1

Several level-2 variables not specified in the final HLM model (e.g., school ESL) were included in the preliminary analyses. As shown in Table 4, the descriptive statistics indicate a good deal of variability between schools. For example, schools range from 0% of students eligible for free/reduced lunch benefits, ESL services, and special education services to 98%, 86% and 40% of students eligible, respectively. There was much less variability for teacher certification, which contributed to its exclusion from the HLM analysis (Schafer et al., 2000). Descriptive statistics for these level-2 inputs were not available for comparison at the state level.

Table 4

Descriptive Statistics for School-Level Variables

Variable	Min	Max	*M*	*SD*
School SES	.000	.980	.453	.304
School ESL	.00	.864	.182	.207
School Special Education	.00	.402	.120	.056
School Prior Achievement	467.24	633.85	556.14	35.78
Teacher Experience	4	21	11.05	3.33
Teacher Education	.115	.824	.440	.128
Teacher Certification	.500	1.00	.930	.083
Student-teacher Ratio	7	25.2	16.52	3.19

Note. All means and standard deviations are unweighted.

Outcome Measure

Descriptive statistics for the total scale scores from the CSAP third- and fourth-grade reading tests are reported in Table 5. Independent samples *t*-tests were conducted to compare the performance of student groups on the outcome measure. Although the same tests were run for the prior achievement measure, findings are not reported because they

were wholly consistent with the results for the fourth-grade test (i.e., statistically significant differences across all level-1 variables).

Table 5

Descriptive Statistics for Prior Achievement and Outcome Measure

Variable	Min	Max	*M*	*SD*
Grade 3 Reading Total Scale Score	179	756	561.39	75.07
Grade 4 Reading Total Scale Score	240	856	583.06	63.82

Specifically, white students averaged 602.69 on the CSAP fourth-grade reading test while minority students averaged 553.79 ($t = 45.45, p < .001$). Students ineligible for free/reduced lunch benefits scored 602.94 while students eligible for this benefit scored 548.15 ($t = 50.47, p < .001$). Students not eligible for ESL services averaged 587.64 on the outcome measure while eligible students averaged 521.12 ($t = 34.12, p < .001$). Students not eligible for special education services averaged 591.73 on the test while eligible students averaged 517.00 ($t = 38.72, p < .001$). Finally, students with three years or more of continuous enrollment on campus scored 585.86, on average, while those with two years of continuous enrollment scored 570.61 ($t = 10.55, p < .001$).

The CSAP results also are reported as performance levels in Table 6 to simplify interpretation. For the third-grade reading test, the range of possible scores was 150 to 795, with 150-465 being unsatisfactory, 466-525 being partially proficient, 526-655 being proficient, and 656-795 being advanced. For the fourth-grade reading test, the range was 180 to 940, with 180-516 being unsatisfactory, 517-571 being partially proficient, 572-670 being proficient, and 671-940 being advanced.

Table 6

Frequency Counts for Prior Achievement and Outcome Measure

Performance Level	*n*	%
Grade 3 Reading Test		
Unsatisfactory/Partially Proficient	3817	29.5
Proficient	7939	61.4
Advanced	1180	9.1
Grade 4 Reading Test		
Unsatisfactory/Partially Proficient	5119	39.6
Proficient	6984	54.0
Advanced	833	6.4

As for the generalizability of the achievement data, the performance levels for the test-taking population were 37.1% unsatisfactory or partially proficient, 54.9% proficient, and 6.5% advanced, as compared with 39.6%, 54%, and 6.4% respectively for the sample. Although the sample was not randomly selected, the external validity seems adequate based on comparisons between demographic data and test scores.

Assumptions

To make legitimate inferences from a hierarchical linear model, assumptions regarding the structural and random components must be defensible (Raudenbush & Bryk, 2002). Thus, the following key assumptions underlying the HLM methodology were tested before commencing with the two-level analysis (Raudenbush & Bryk, 2002).

1. Conditional on student characteristics, the within-school error terms are normally distributed with a mean of zero in each school and equal variances between schools.
2. Level-1 predictors that are excluded from the model and consigned to the error term are independent of student characteristics and school-level predictors.

3. The residual school effects are normally distributed.
4. Level-2 predictors that are excluded from the model and relegated to the random effects are independent of school-level predictors and student characteristics.
5. The student-level error term is independent of the residual school effects.

Specifically, assumptions two and four "focus on the relationship between the variables included in the structural portion of the model and those factors relegated to the error terms. They pertain to the adequacy of model specification" (Raudenbush & Bryk, 2002, p. 255). As for the random components, the tenability of assumptions one, three, and five affects the reliability and precision of regression coefficients, variance estimates, hypothesis tests, and confidence intervals (Raudenbush & Bryk, 2002). Before these assumptions were tested, traditional univariate and multivariate assumptions were assessed at each level of the hierarchical linear model.

Univariate

A key assumption of HLM is that the predictors and outcome measure must be normally distributed (Raudenbush & Bryk, 1986). Although "non-normality of the errors at level 1 will not bias estimation of the level-2 effects, it will introduce bias into standard errors at both levels and therefore into the computation of confidence intervals and hypothesis tests" (Raudenbush & Bryk, 2002, p. 266). The continuous level-1 variables were assessed for univariate normality, and the nominal school-level inputs were tested for multivariate normality. To analyze these assumptions, skewness and kurtosis statistics, Kolmogorov-Smirnov (K-S) tests, histograms, and Q-Q plots were examined.

Outputs. Skewness statistics for the CSAP third- and fourth-grade reading test scores were well within +/- 1.0 (-.087 and -.503 respectively). Kurtosis statistics also

were within +/- 1.0 (.809 and .936 respectively), with histograms confirming normal distributions. Although the K-S tests were statistically significant ($p < .001$), visual inspection of the Q-Q plots confirmed the judgment of univariate normality for the prior achievement variable and outcome measure.

Inputs. The continuous school-level variables were tested for univariate normality. Although the "estimation of fixed effects will not be biased by a failure of the normality assumption at level 2 . . . failure of the normality assumption will affect the validity of the confidence intervals and hypothesis tests" (Raudenbush & Bryk, 2002, p. 274). As displayed in Table 7, skewness, kurtosis, and K-S test statistics were generated for each level-2 predictor. Histograms and Q-Q plots also were inspected for these inputs.

Table 7

Univariate Normality Statistics for School-Level Variables

Variable	Skewness	Kurtosis	K-S Test
School SES	.261	-1.29	.000
School ESL	1.34	.739	.000
School Special Education	1.58	4.24	.000
School Prior Achievement	-.175	-.651	.011
Teacher Education	.127	-.337	.038
Teacher Experience	-.555	-.190	.007
Teacher Certification	-1.69	3.65	.000
Student-Teacher Ratio	-.133	.061	.200[a]

[a]This is a lower bound of the true significance.

For teacher education, teacher experience, and school prior achievement, skewness and kurtosis statistics were within +/- 1.0, the K-S tests were not significant ($p > .001$), and visual inspection of the histograms and Q-Q plots confirmed the judgment

of univariate normality. For student-teacher ratio, skewness and kurtosis were within +/- 1.0, the K-S test was not interpretable, and visual inspection of the histogram and Q-Q plot confirmed a judgment of univariate normality. For school SES, skewness was within +/- 1.0, and kurtosis was within the +/- 2 range when the statistic is divided by the standard error and the square root is taken. Although the K-S test was statistically significant, a visual inspection of the histogram and Q-Q plot confirmed the judgment of univariate normality.

For teacher certification and school special education, skewness and kurtosis were outside of the +/- 2 range, the K-S tests were statistically significant, and a visual inspection of the histograms and Q-Q plots confirmed a judgment of univariate non-normality. For school ESL, kurtosis was within +/- 1.0, but skewness was outside of the +/- 2.0 range. Furthermore, the K-S test was significant and a visual inspection of the histogram and Q-Q plot confirmed the judgment of univariate non-normality. Buttressed by these results and the lack of empirical support in the HLM analysis, school ESL, school special education, and teacher certification were dropped from consideration as level-2 predictors.

Multivariate

As displayed in Table 8, multivariate statistics were generated for level-1 variables to test assumptions related to normality and multicollinearity.

Normality. For special education status, skewness and kurtosis were within +/- 1.0 for both groups, the K-S test was significant only for the "not eligible" group, and a visual inspection of the histograms and Q-Q plots confirmed a judgment of multivariate normality. For continuous enrollment on campus, skewness and kurtosis were within +/-

1.0 for both groups. Although the K-S tests were significant ($p < .001$), visual inspection of the histograms and Q-Q plots supported a judgment of multivariate normality.

For ethnicity, skewness was within +/- 1.0 for both groups, but kurtosis was slightly greater than 1.0 and outside of the +/- 2 range. Although the K-S tests were statistically significant, a visual inspection of the histograms and Q-Q plots indicated multivariate normality. For student SES and ESL status, skewness was within +/- 1.0, but kurtosis was outside of the +/- 2 range at both levels. Although the K-S tests were statistically significant and the Q-Q plots displayed some deviations, a visual inspection of the histograms supported a judgment of multivariate normality for both variables.

Table 8

Multivariate Normality Statistics for Student-Level Variables

Variable	Skewness	Kurtosis	K-S Test
Special Education Status			
Not Eligible	-.266	.674	.000
Eligible	-.245	.198	.044
Continuous enrollment			
3 years or more	-.501	.968	.000
2 years	-.510	.844	.000
Ethnicity			
White	-.557	1.460	.000
Minority	-.470	1.018	.000
SES Status			
Not eligible	-.469	1.240	.000
Eligible	-.561	1.063	.000
ESL Status			
Not eligible	-.516	1.053	.000
Eligible	-.805	1.557	.000

Multicollinearity. To test for multicollinearity at level 1, tolerance and variance inflation factor (VIF) statistics were calculated for the student-level predictors. For ethnicity, continuous enrollment, SES, special education, and ESL, tolerance ranged from .729 to .990 and VIF ranged from 1.01to 1.37. The tolerance statistics were well over the .10 benchmark and the VIF statistics well under the 10.0 standard, which indicated a low degree of multicollinearity.

Multicollinearity at level 2 was assessed according to the bivariate correlation matrix presented in Table 9. Statistically significant intercorrelations ranged from .14 to .85, with only school SES and school ESL (.81) and school SES and school prior achievement (.85) having a high level of multicollinearity.

Table 9

Intercorrelations for School-Level Variables

Variable	1	2	3	4	5	6	7	8
1. School SES	--							
2. School ESL	.81**	--						
3. School Special Education	.31**	.07	--					
4. School Prior Achievement	-.85**	-.70**	-.25**	--				
5. Teacher Education	-.41**	-.34**	-.04	.40**	--			
6. Teacher Experience	-.46**	-.41**	-.09	.47**	.54**	--		
7. Student-Teacher Ratio	-.29**	-.14*	-.31**	.23**	.11	.14*	--	
8. Teacher Certification	-.43**	-.42*	-.03	.45**	.46**	.59*	.05	--

* $p < .05$, 2 tailed. ** $p < .01$, 2 tailed.

Cross-level

Equation building in HLM requires the examination of cross-level relationships to identify potential violations to homogeneity of variance, normality, and linearity assumptions (Raudenbush & Bryk, 2002). For example, scatter plots can expose nonlinear bivariate relationships (Raudenbush & Bryk, 2002). To facilitate the cross-level analysis, a residual file created during the HLM analysis was formatted in SPSS. The file contained Mahalanobis distances, estimates of level-1 variability, Empirical Bayes (EB) residuals, OLS residuals, fitted values for intercepts and slopes, and posterior variances and covariances for the intercept and slope estimates (Raudenbush et al., 2001).

Homogeneity of Variance. In most HLM analyses, researchers start with the assumption that errors from the level-1 model have equal variance. According to Raudenbush and Bryk (2002), "heterogeneity may indicate a possible misspecification of the level-1 model. In particular, unidentified slope heterogeneity at level 1 would appear as heterogeneity of level-1 error variance" (p. 264). To account for the nested nature of the level-1 data, a histogram based on the residual distributions from the final HLM model is displayed in Figure 1.

There were four schools in which the within-school standard deviation was smaller than expected if there was homogeneity of variance at the student level (Raudenbush et al., 2001). As three of the four schools had zero variability on the SES variable (i.e., no eligible students), it appears these extraordinarily uniform schools were responsible for most of the heterogeneity in level-1 variance (Raudenbush & Bryk, 2002). Although one option is to remove these schools from the sample, this would compromise external validity because there are many schools in Colorado with comparable

demographics. A second alternative is to include additional variables in the level-1model to remove some of the residual heterogeneity (Raudenbush & Bryk, 2002). However, there must be theoretical and empirical support for such an approach. In addition, "a violation of the homogeneity assumption is not per se a serious problem for estimating either the level-2 coefficients or their standard errors" (Raudenbush & Bryk, 2002, p. 264).

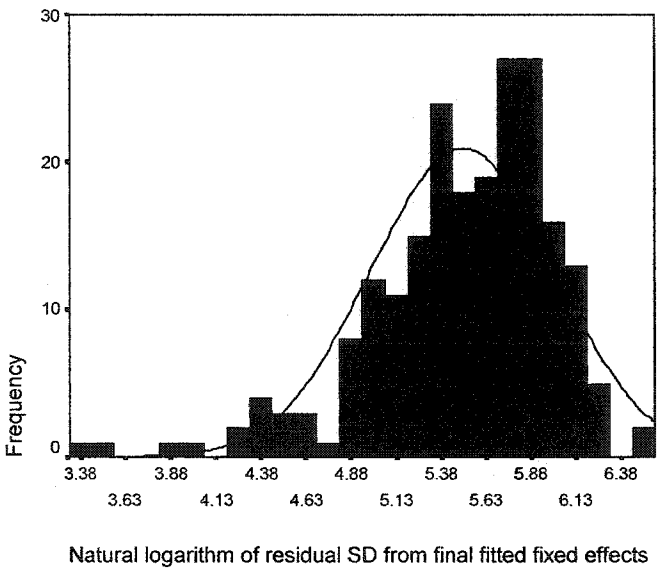


Figure 1. Histogram plot of within-school residual dispersions ($N = 214$).

Normality. The Mahalanobis distance statistic for each school is used to "measure the distance between the residual estimates for each group relative to the expected distance based on the model" (Raudenbush & Bryk, 2002, p. 274). As displayed in Figure 2, a Q-Q plot of the Mahalanobis distances and the expected values of the order statistics was generated to test the cross-level normality assumption. The Q-Q plot is reasonably close to a 45-degree line, which indicates that the level-2 random effects are normally distributed (Raudenbush et al., 2001).

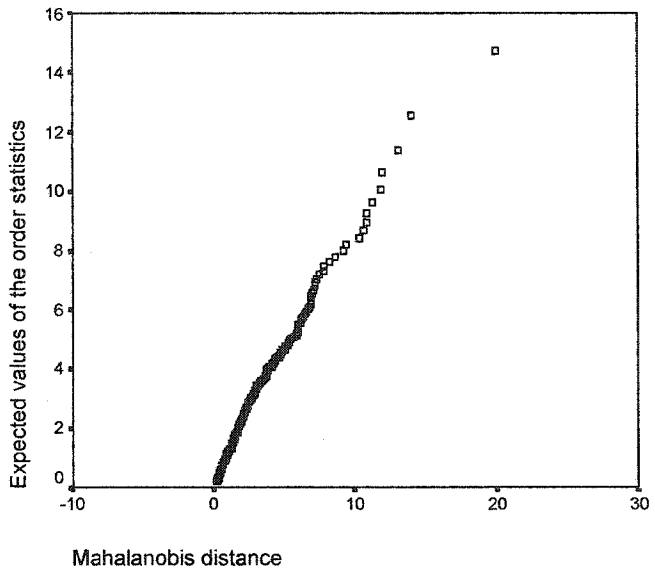


Figure 2. Plot of Mahalanobis distances ($N = 249$).

Linearity. To test the cross-level assumption of linearity, the residuals from the predictors in the final school-level model were plotted against the EB residuals for the appropriate slopes and intercepts from the level-1 model. The expectation is that the residuals are randomly distributed around the zero line without regard to the values of the level-2 variable (Raudenbush et al., 2001).

The scatter plot in Figure 3 shows a linear relationship between school SES and the prior achievement slope, as the residuals are randomly distributed around the zero line. This pattern of linearity held for the relationship between school SES and the intercept, school SES and the special education slope, and teacher experience and the intercept. As the SES-achievement slope was fixed, EB residuals were not estimated to allow testing of the linearity assumption.

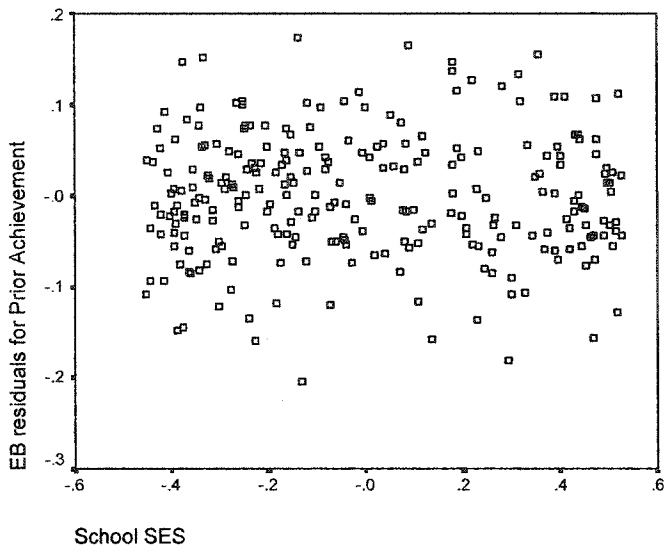


Figure 3. Residual plot for prior achievement and school SES ($N = 254$).

Conditional shrinkage. The residual file also contains evidence of conditional shrinkage for the Empirical Bayes intercept and slope residuals. To increase the accuracy in parameter estimation, HLM pulls OLS regression lines toward a predicted value based on the level-2 model (Raudenbush & Bryk, 2000). Thus, EB residuals are expected to be more compressed than OLS residuals (Raudenbush et al., 2001).

As displayed in Figure 4, the range for the special education slope EB residuals was –20 to 20, where the range for the OLS residuals was much larger (-40, 40). This plot of OLS and EB residuals is representative of the residual plot for the prior achievement slope. The expectation for the intercepts is that, although EB residuals are "shrunk" as compared with OLS residuals, the ranges are more comparable than for the slopes (Raudenbush et al., 2001). As displayed in Figure 5, the range of EB residuals (-25, 25) was slightly smaller than the range of OLS residuals for the intercepts (-35, 35).

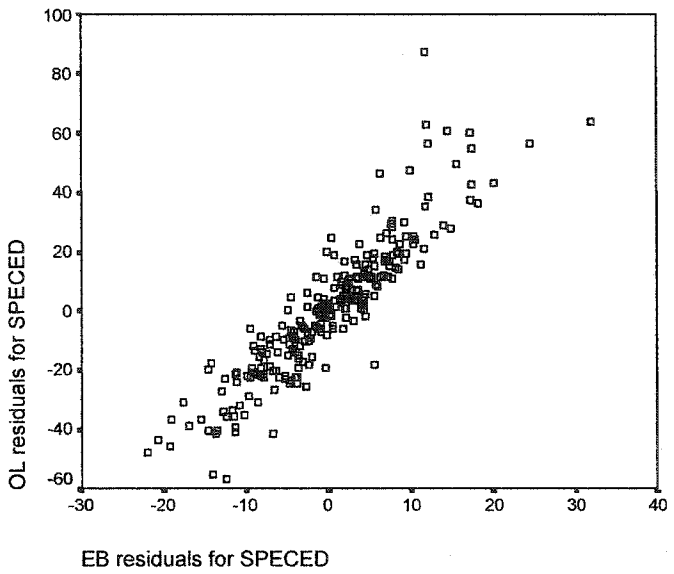


Figure 4. Residual plot for special education slopes ($N = 249$).

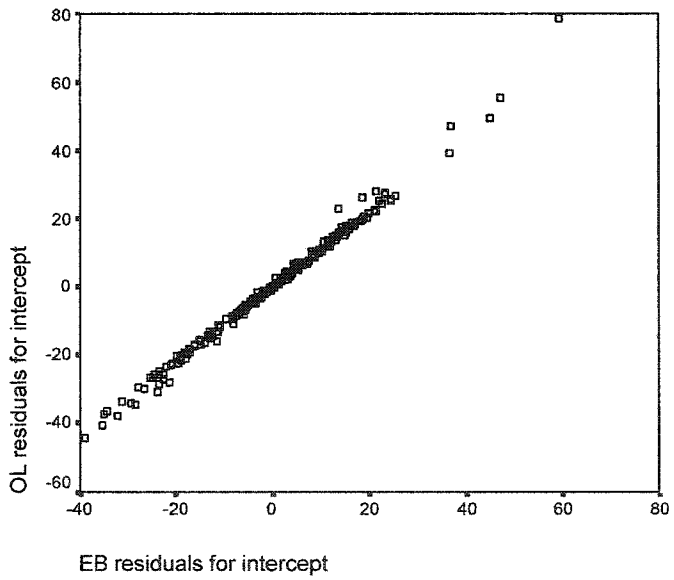


Figure 5. Residual plot for intercepts ($N = 249$).

Model Building

The present study employed two-level hierarchical linear modeling with students at level-1 and schools at level-2. The HLM analysis was based on a four-step model

building process: one-way ANOVA, means-as-outcomes, random coefficient, and intercepts- and slopes-as-outcomes (Raudenbush & Bryk, 2002).

One-way ANOVA

As there are no level-1 or level-2 predictors, the unconditional model is viewed as a one-way ANOVA with random effects because it separates variability into within-school and between-school components (Raudenbush & Bryk, 2002). This preliminary step provides an indication of whether there is enough variance to be explained at the school-level to support the use of multilevel modeling (Yu & White, 2002). The equations for the one-way ANOVA model are

Level 1 $Y_{ij} = ß_{0j} + r_{ij}$

Level 2 $ß_{0j} = \gamma_{00} + u_{0j}$

where Y_{ij} is the outcome for student i in school j; $ß_{0j}$ is the intercept or mean school achievement; r_{ij} is the student-level residual or the difference between the score of student i from the average score of school j; γ_{00} is the grand mean achievement across all schools; and u_{0j} is the level-2 residual or the difference between the mean score of school j and the grand mean (Raudenbush & Bryk, 2002). The results from the unconditional model are presented in Table 10.

Fixed Effects. For all HLM analyses, the final estimation of fixed effects with robust standard errors was used. For the fixed effects, the weighted least squares estimate for the grand mean of reading achievement was $\hat{\gamma}_{00} = 578.61$.

Random effects. The within-school variance was $\hat{\sigma}^2 = 3127.50$ and the between-school variance was $\hat{\tau}_{00} = 1017.24$. According to Raudenbush and Bryk (2002), the "intraclass correlation, which represents the proportion of variance in Y between schools,

is estimated by substituting variance components for their respective parameters in the equation: $\hat{\rho} = \hat{\tau}_{00}/(\hat{\tau}_{00}+\hat{\sigma}^2)$" (p. 71). In this case, 1017.24 / (1017.24 + 3127.50) = .245 which indicates that 25% of the variance in student achievement was between-schools and 75% was within-schools. The rule of thumb is that HLM is appropriate to estimate school effects if the intraclass correlation exceeds .10 (Yu & White, 2002).

Table 10

Results from the One-way ANOVA Model

Fixed Effect		Coefficient	*SE*	
Average school mean, γ_{00}		578.61	2.07	
Random Effect	Variance Component	*df*	χ^2	*p* value
School mean, u_{0j}	1017.24	253	4244.98	.000
Level-1 effect, r_{ij}	3127.50			

Supplementary statistics. An overall reliability estimate of the sample school means was generated from the one-way ANOVA model. The reliability estimate of $\hat{\lambda}$ = .93 suggests that the sample means were very reliable as gauges of the true school means on the outcome measure (Raudenbush & Bryk, 2002). Another auxiliary statistic is to test the null hypothesis that all schools have the same mean by determining whether the estimated value of level-2 variance ($\hat{\tau}_{00}$) is significantly different than zero (Raudenbush & Bryk, 2002). The test statistic (χ^2) had a value of 4244.98 with 253 degrees of freedom, which was statistically significant at $p < .001$. Thus, there was significant variation in mean reading achievement across schools in the sample.

The final analysis for an unconditional model is to estimate the variation in test scores across schools by calculating a plausible range for the means (Raudenbush & Bryk, 2002). The equation, $\hat{\gamma}_{00} \pm 1.96\ (\hat{\tau}_{00})^{1/2}$, is used to calculate a range in which 95% of the school means are included (Raudenbush & Bryk, 2002). For this sample, $578.61 \pm 1.96\ (1017.24)^{1/2} = (516.10, 641.12)$, which represents considerable breadth in mean reading achievement between schools.

Means-as-Outcomes

The second step in the model building process is to analyze a regression with means-as-outcomes model. Specifically, the level-1 model from the one-way ANOVA remains the same, but school-level predictors of each school's mean are entered into the level-2 model (Raudenbush & Bryk, 2002). Thus, the equations are

Level 1 $\quad Y_{ij} = \beta_{0j} + r_{ij}$

Level 2 $\quad \beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{SCHSES}) + \gamma_{02}\,(\text{TEACHEXP}) + u_{0j}$

where γ_{00} is the intercept; γ_{01} and γ_{02} are the effects of school socioeconomic status (SCHSES) and teacher experience (TEACHEXP) on β_{0j}; and τ_{00} is the residual level-2 variance after controlling for school-level variables (Raudenbush & Bryk, 2002).

Fixed effects. The results from the means-as-outcomes model are presented in Table 11. There was a statistically significant association between SCHSES and mean achievement ($\hat{\gamma}_{01} = -85.60$, $t = -21.74$) and between TEACHEXP and mean achievement ($\hat{\gamma}_{02} = 1.29$, $t = 3.61$). Specifically, schools with less affluent students scored substantially lower on the CSAP fourth-grade reading test, while schools with more experienced teachers scored slightly higher on the assessment.

Table 11

Results from the Means-as-Outcomes Model

Fixed Effect	Coefficient	*SE*	*t* ratio
Model for school means			
Intercept, γ_{00}	577.99	1.03	560.40
SCHSES γ_{01}	-85.60	3.94	-21.74
TEACHEXP, γ_{02}	1.29	.36	3.61

Random Effect	Variance	*df*	χ^2	*p* value
School mean, u_{0j}	206.56	251	1066.00	.000
Level-1 effect, r_{ij}	3129.50			

Random effects. The between-school residual variance was $\hat{\tau}_{00} = 206.56$, which is much smaller than the original value of $\hat{\tau}_{00} = 1017.24$ from the one-way ANOVA model. This reduction is further illustrated by the range of plausible values for school means, (549.82, 606.16), which is considerably smaller than the range when level-2 characteristics were not held constant (516.10, 641.12). However, the chi-square statistic ($\chi^2 = 1066.00, p < .001$) indicates that there was still significant variation between schools on mean reading achievement even after controlling for level-2 predictors.

Supplementary statistics. After controlling for the level-2 predictors, the relationship between scores in the same school was reduced from .245 to .062 based on the equation, $\hat{\rho} = \hat{\tau}_{00}/(\hat{\tau}_{00} + \hat{\sigma}^2) = 206.56 / (206.56 + 3129.52)$. According to Raudenbush and Bryk (2002), "the estimated ρ is now a conditional intraclass correlation and measures the degree of dependence among observations within schools" (p. 75) that are of the same school SES and teacher experience level. Thus, without the commonality

of school SES and teacher experience, the already small association between scores in the same school was reduced even further.

Random Coefficient

The random coefficient model regresses the outcome measure on student-level predictors to provide an estimate of variability in the intercepts and slopes across schools and to offer guidance on the specification of level-1 coefficients as random, fixed, or nonrandomly varying (Yu & White, 2002). Furthermore, no level-2 variables are included in this model and regression coefficients are assumed to randomly vary across the population of schools as a function of the grand mean and random error (Raudenbush & Bryk, 2002). The equations for the random coefficient model are

Level 1

$$Y_{ij} = ß_{0j} + ß_{1j}(\text{ETHNIC}) + ß_{2j}(\text{CEC}) + ß_{3j}(\text{SES}) + ß_{4j}(\text{SPECED}) + ß_{5j}(\text{PRIOR}) + r_{ij}$$

Level 2

$$ß_{0j} = \gamma_{00} + u_{0j}$$
$$ß_{1j} = \gamma_{10} + u_{1j}$$
$$ß_{2j} = \gamma_{20} + u_{2j}$$
$$ß_{3j} = \gamma_{30} + u_{3j}$$
$$ß_{4j} = \gamma_{40} + u_{4j}$$
$$ß_{5j} = \gamma_{50} + u_{5j}$$

where the distribution of each school's reading achievement is represented by the intercept $ß_{0j}$ and the slopes, $ß_{1j}$, $ß_{2j}$, $ß_{3j}$, $ß_{4j}$, $ß_{5j}$. The slopes characterize the expected change in test scores for a unit change in ethnicity (ETHNIC), continuous enrollment on campus (CEC), free/reduced lunch status (SES), special education status (SPECED), and prior achievement (PRIOR), respectively. According to Raudenbush and Bryk (2002), γ_{00} is the average of school mean achievement across the population of schools; γ_{10}, γ_{20}, γ_{30}, γ_{40}, γ_{50} are the average ETHNIC-achievement, CEC-achievement, SES-achievement, SPECED-achievement, and PRIOR-achievement regression slopes across schools; u_{0j} is

the unique increment to the intercept associated with school j; and u_{1j}, u_{2j}, u_{3j}, u_{4j}, u_{5j} are the unique increments to the slopes associated with school j.

Fixed effects. Table 12 provides parameter estimates derived from the random coefficient model. The average school mean on the CSAP fourth-grade reading test was $\hat{\gamma}_{00} = 578.44$, the average ETHNIC-achievement slope was $\hat{\gamma}_{10} = -5.44$, the average CEC-achievement slope was $\hat{\gamma}_{20} = -2.54$, the average SES-achievement slope was $\hat{\gamma}_{30} = -7.26$, the average SPECED-achievement slope was $\hat{\gamma}_{40} = -21.44$, and the average PRIOR-achievement slope was $\hat{\gamma}_{50} = .598$.

Table 12

Results from the Random Coefficient Model

Fixed Effect	Coefficient	*SE*	*t* ratio
Overall mean achievement, γ_{00}	578.44	2.08	277.94
Mean ETHNIC-achievement slope, γ_{10}	-5.44	.819	-6.65
Mean CEC-achievement slope, γ_{20}	-2.54	.906	-2.80
Mean SES-achievement slope, γ_{30}	-7.26	.93	-7.83
Mean SPECED-achievement slope, γ_{40}	-21.44	1.37	-15.66
Mean PRIOR-achievement slope, γ_{50}	.598	.008	74.47

Random Effect	Variance	*df*	χ^2	*p* value
School mean, u_{0j}	1077.41	218	9796.61	.000
ETHNIC-achievement slope, u_{1j}	15.60	218	206.23	> .500
CEC-achievement slope, u_{2j}	28.94	218	251.59	.059
SES-achievement slope, u_{3j}	35.22	218	257.35	.035
SPECED-achievement slope, u_{4j}	181.42	218	366.30	.000
PRIOR-achievement slope, u_{5j}	.0095	218	491.82	.000
Level-1 effect, r_{ij}	1115.69			

Note. The chi-square statistics are based on only 219 of 254 units that had sufficient data for computation. Fixed effects and variance components are based on all of the data.

According to the associated *t* ratios, special education status, SES, ethnicity, continuous enrollment, and prior achievement were all significant level-1 predictors of fourth-grade reading achievement. Specifically, students eligible for special education services scored an average of 21 points lower on the CSAP test than did ineligible students when other level-1 variables were held constant. Students eligible for free/reduced lunch benefits scored, on average, 7 points lower than did ineligible students with similar characteristics. Minority students scored, on average, 5 points lower than did white students holding other level-1 variables constant. Students with only two years of continuous enrollment scored 3 points lower, on average, than did students with three or more years of continuous enrollment when other student-level predictors were held constant. Finally, students' fourth-grade reading scale score increased an average of .60 for every one-point increase on their third-grade reading scale score holding other level-1 inputs constant.

Random effects. The estimated variance among school means was $\hat{\tau}_{00} = 1077.41$ with a chi-square value of 9796.61. This indicates that highly significant differences still existed among the 254 schools (Raudenbush & Bryk, 2002). Plausible value ranges that represent 95% confidence intervals were generated for the school means and slopes. For the school means, the equation $\hat{\gamma}_{q0} \pm 1.96\ (\hat{\tau}_{qq})^{1/2}$ yielded $578.44 \pm 1.96\ (1077.41)^{1/2} =$ (514.11, 642.78), which is very similar to the results from the unconditional model (516.10, 641.12).

As for the slopes, the equation for u_{1j} was $-5.44 \pm 1.96\ (15.60)^{1/2} =$ (-13.18, 2.30) and $\chi^2 = 206.23$ with 218 degrees of freedom, which indicates minimal variability among schools on the ETHNIC-achievement slope. For u_{2j}, $-2.54 \pm 1.96\ (28.94)^{1/2} =$ (-13.08,

8.00) and $\chi^2 = 251.59$, which indicates little variability across schools on the CEC-achievement slope. For u_{3j}, $-7.26 \pm 1.96\ (35.22)^{1/2} = (-18.89, 4.37)$ and $\chi^2 = 257.35$, which indicates little variability between schools on the SES-achievement slope. For u_4, $-21.44 \pm 1.96\ (181.42)^{1/2} = (-47.84, 4.96)$ and $\chi^2 = 366.30$, which indicates significant variability among schools on the SPECED-achievement slope. For u_{5j}, $.598 \pm 1.96\ (.0095)^{1/2} = (.407, .789)$ and $\chi^2 = 491.82$, which indicates significant variability across schools on the PRIOR-achievement slope.

Supplementary statistics. The reliability estimates for each school's intercept and slope depend on the amount of variation between schools in the underlying parameters and the precision of school regression equations (Raudenbush & Bryk, 2002). Based on an average of 51 students per school, the intercepts were highly reliable ($\hat{\beta}_0 = .98$). The slopes were much less reliable for ETHNIC ($\hat{\beta}_1 = .08$), CEC ($\hat{\beta}_2 = .14$), SES ($\hat{\beta}_3 = .16$), SPECED ($\hat{\beta}_4 = .37$) and PRIOR ($\hat{\beta}_5 = .55$). The slope estimates are unreliable because the "true slope variance across schools is much smaller than the variance of the true means. Also, the slopes are estimated with less precision that are the means because many schools are relatively homogeneous" (Raudenbush & Bryk, 2002, p. 79).

The variance explained at the student level was determined by comparing the level-1 variance estimates from the random coefficient and unconditional models (Raudenbush & Bryk, 2002). Thus, the proportion of variance explained at level 1 equals $\hat{\sigma}^2$ (random ANOVA) - $\hat{\sigma}^2$ (random coefficient)/ $\hat{\sigma}^2$ (random ANOVA) = 3127.50 – 1115.69 / 3127.50 = .643. This result suggests that ethnicity, continuous enrollment, SES, special education status, and prior achievement accounted for 64% of the variance in reading achievement at the student level. Although 36% of the within-school variance

remains unexplained, this model fit is consistent with other SER studies (e.g., Yu & White, 2002).

Model Specification

The fourth step in the HLM analysis is to build an explanatory model that accounts for the variability in the school regression equations (Raudenbush & Bryk, 2002). Thus, an intercepts- and slopes-as-outcomes model was specified to investigate whether schools had different mean reading achievement and varying slope relationships (Raudenbush & Bryk, 2002). Before presenting the results, the variable selection and centering process is detailed to provide further support for the final HLM model.

Variable Inclusion

Although Raudenbush and Bryk (2002) caution against specifying a "saturated" final model, the HLM process requires the inclusion of all hypothesized student-level predictors in the original intercepts- and-slopes-as-outcomes model. In this case, SES, prior achievement, special education status, ESL status, ethnicity, and continuous enrollment on campus were entered into the level-1 equation.

A student-level predictor should be removed only if there is a lack of slope heterogeneity and no evidence of a fixed effect (Raudenbush & Bryk, 2002). For example, the ETHNIC-achievement and CEC-achievement slopes did not vary significantly across schools. However, both ethnicity and continuous enrollment had a statistically significant effect on student achievement and thus were retained in the level-1 model. The specification of these variables also reduced the deviance statistic without adding additional parameters.

Although there was a significant fixed effect and slope heterogeneity for ESL, it was removed from the student-level equation for a technical reason. Specifically, the lack of within-school variation on student ESL resulted in insufficient data for the estimation of OLS level-1 coefficients. This was problematic from a generalizability standpoint because cases were then omitted for the univariate chi-square tests, reliability statistics, and least-squares estimates.

The specification of the level-2 model was based on statistical assumptions and a recommendation from Raudenbush and Bryk (2002):

> Although it is possible to enter a different set of level-2 predictors for each outcome in the level-2 model, this flexibility should be used judiciously and with caution. The safest way to proceed is to introduce a common set of level-2 predictors in all of the level-2 equations. (p. 151)

Thus, a common set of school-level predictors was entered into all of the level-2 equations. School SES was retained in the school-level model because it was a significant predictor of the intercept, SES-achievement slope, SPECED-achievement slope, and PRIOR-achievement slope. As a significant predictor of the intercept and the SES-achievement slope, teacher experience also was specified in the level-2 model.

As a statistically significant predictor of the SES-achievement slope, teacher education was retained in the school-level model. Teacher education also was a significant predictor when entered alone into the intercept equation. Although this level-2 variable did not significantly predict the intercept in the final model, it is important to account for "between-school context differences, even if the effects of school context do not meet the 'technical' criterion of statistical significance" (Pituch, 1999, p. 203).

Variables that did not significantly predict the intercept or slopes were dropped from the analysis. For example, schools with varying student-teacher ratios and teacher

certification levels did not differ on school mean reading achievement holding the other school-level predictors constant. Furthermore, the RATIO-achievement and TEACHCRT-achievement slopes did not significantly vary across schools. As a result, both level-2 predictors were dropped from the final model.

Raudenbush and Bryk (2002) recommend that, "the aggregate of a level-1 predictor should also be entered in the intercept model. This allows representation for each level-1 predictor of the two separate relationships, β_w and β_b that might exist" (p. 152). Thus, school mean prior achievement and special education status were entered into the intercept equation as level-2 predictors. Special education was not a significant predictor of the intercept and was deleted from the school-level model. Although significantly predictive of the intercept and slopes, prior achievement was deleted from the model because of a high level of multicollinearity with school SES. School ethnicity and continuous enrollment were not added as level-2 predictors because the variables were not available from the secondary databases. As the other level-2 variables were based on data from the entire school (except for prior achievement), it was inappropriate to compute new variables based solely on data from fourth-grade students.

Variable Centering

The centering of level-1 variables directly impacts the interpretations of the HLM analysis (Raudenbush & Bryk, 2002). One option is to center student-level predictors around the grand mean, in which the "intercept β_{0j}, is the expected outcome for a subject whose value on X_{ij} is equal to the grand mean" (Raudenbush & Bryk, 2002, p. 33). A second alternative is to center level-1 predictors around the group mean, so that the intercept is the unadjusted mean for a level-2 unit (Raudenbush & Bryk, 2002). There are

other options (e.g., uncentered), but grand- and group-mean centering are the most common and useful.

Although the guidance on centering provided by Raudenbush and Bryk (2002) is not always clear, there were certain rules of thumb followed in the HLM analysis. For example, group-mean centering is necessary to provide an unbiased estimate of slope heterogeneity when the relationship between a level-1 variable and the outcome measure varies across schools (Raudenbush & Bryk, 2002). Thus, SPECED and PRIOR were group-mean centered in the intercepts- and slopes-as-outcomes model. For SES, ETHNIC, and CEC, the lack of slope heterogeneity called for grand-mean centering, which is slightly more precise under that condition (Raudenbush & Bryk, 2002).

Raudenbush and Bryk (2002) are clear that it is convenient to center all level-2 predictors around the grand mean when estimating school effects while adjusting for student-level covariates. Thus, school SES, teacher education, and teacher experience all were centered around their respective grand means in the final HLM model.

Intercepts- and-Slopes-as-Outcomes

For the final model specification, the level-1 equation from the random coefficient model was retained while school-level variables were added as predictors of the level-1 intercept and slopes. The equations for the fitted model are

Level 1

$$Y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{ETHNIC}) + \beta_{2j}\,(\text{CEC}) + \beta_{3j}\,(\text{SES}) + \beta_{4j}\,(\text{SPECED}) + \beta_{5j}\,(\text{PRIOR}) + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{SCHSES}) + \gamma_{02}\,(\text{TEACHEXP}) + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30} + \gamma_{31}\,(\text{SCHSES}) + \gamma_{32}\,(\text{TEACHEDU}) + \gamma_{33}\,(\text{TEACHEXP})$$
$$\beta_{4j} = \gamma_{40} + \gamma_{41}\,(\text{SCHSES}) + u_{4j}$$
$$\beta_{5j} = \gamma_{50} + \gamma_{51}\,(\text{SCHSES}) + u_{5j}$$

According to Raudenbush and Bryk (2002), the outcome measure is a function of the grand mean (γ_{00}), the main effect of SCHSES (γ_{01}), TEACHEXP (γ_{02}), ETHNIC (γ_{10}), CEC (γ_{20}), SES (γ_{30}), SPECED (γ_{40}), and PRIOR (γ_{50}), five cross-level interactions involving SCHSES with SES (γ_{31}), TEACHEDU with SES (γ_{32}), TEACHEXP with SES (γ_{33}), SCHSES with SPECED (γ_{41}), and SCHSES with PRIOR (γ_{51}), and random error (u_{0j}, u_{4j}, u_{5j}, r_{ij}).

Fixed effects. Results from the final intercepts- and-slopes-as-outcomes model are presented in Table 13. School SES was negatively related to school mean achievement ($\hat{\gamma}_{01}$ = -74.85, t = 19.08), and teacher experience was positively related to school mean achievement ($\hat{\gamma}_{02}$ = 1.10, t = 3.40). To interpret these coefficients, an increase in one unit of school SES (i.e., less affluent) results in a decline of 74.85 points on the CSAP fourth-grade reading test for schools with similar levels of teaching experience. An increase in one unit of teacher experience results in a gain of 1.09 points for schools with similar school SES.

School SES has a positive relationship with the SES-achievement slope ($\hat{\gamma}_{31}$ = 8.13, t = 2.29), which indicates that schools with higher percentages of students eligible for free/reduced lunch have steeper SES-achievement slopes (i.e., less equitable). Teacher education has a negative relationship with the SES-achievement slope ($\hat{\gamma}_{32}$ = -20.33, t = -2.38), which indicates that schools with higher levels of teacher education have flatter SES-achievement slopes (i.e., more equitable). Teacher experience has a positive relationship with the SES-achievement slope ($\hat{\gamma}_{33}$ = .761, t = 2.13), which indicates that schools with more experienced teachers have steeper SES-achievement slopes (i.e., less equitable).

Table 13

Results from the Intercepts- and Slopes-as-Outcomes Model

Fixed Effects	Coefficient	*SE*	*t* ratio
Model for school means, β_0			
Intercept, γ_{00}	578.30	1.08	536.30
School SES, γ_{01}	-74.85	3.92	-19.08
Teacher Experience, γ_{02}	1.10	.32	3.40
Model for ETHNIC-Achievement Slopes, β_1			
Intercept, γ_{10}	-5.93	.81	-7.31
Model for CEC-Achievement Slopes, β_2			
Intercept, γ_{20}	-2.66	.91	-2.94
Model for SES-Achievement Slopes, β_3			
Intercept, γ_{30}	-7.44	.92	-8.09
School SES, γ_{31}	8.13	3.55	2.29
Teacher Education γ_{32}	-20.33	8.56	-2.38
Teacher Experience, γ_{33}	.76	.36	2.13
Model for SPECED-Achievement Slopes, β_4			
Intercept, γ_{40}	-21.48	1.37	-15.64
School SES, γ_{41}	9.75	4.31	2.26
Model for PRIOR-Achievement Slopes, β_5			
Intercept, γ_{50}	.596	.008	75.15
School SES, γ_{51}	.144	.028	5.25

Random Effects	Variance	*df*	χ^2	*p* value
School Mean, u_{0j}	250.75	246	2746.33	.000
SPECED-achievement slope, u_{4j}	170.09	247	395.20	.000
PRIOR-achievement slope, u_{5j}	.008	247	525.13	.000
Level-1 effect, r_{ij}	1125.12			

Note. The chi-square statistics are based on only 249 of 254 units that had sufficient data for computation. Fixed effects and variance components are based on all the data.

For the SPECED-achievement slope, school SES has a positive effect ($\hat{\gamma}_{41}$ = 9.75, t = 2.26), which indicates that less affluent schools have steeper SPECED-achievement slopes (i.e., less equitable). For the PRIOR-achievement slope, school SES again has a positive relationship ($\hat{\gamma}_{51}$ = .144, t = 5.25), which indicates that less affluent schools also have steeper PRIOR-achievement slopes (i.e., less equitable).

Random effects. For the fitted model, the residual variance at level 2 after accounting for the school-level predictors was $\hat{\tau}_{00}$ = 250.75. The level-1 effect, r_{ij} = 1125.12, was almost identical to the variance estimate from the random coefficient model. Thus, student SES, ethnicity, prior achievement, mobility, and special education status accounted for 64% of the within-school variance in fourth-grade reading achievement even after controlling for school-level predictors.

The residual variance of the SPECED-achievement slope was reduced from 181.42 in the random coefficient model to 170.09 in the intercepts- and-slopes-as-outcomes model. The residual variance of the PRIOR-achievement slope was reduced from .0095 in the random coefficient model to .0077 in the fitted model. Reductions of 6% and 19% in residual variance, respectively, suggests that most of the variability in the special education and prior achievement slopes was not associated with SCHSES (Raudenbush & Bryk, 2002). The chi-square tests indicate that the SPECED-achievement slope (χ^2 = 395.20, p < .001) and PRIOR-achievement slope (χ^2 = 525.13, p < .001) were heterogeneous across schools after controlling for level-2 predictors (Pituch, 1999). This validates the specification of special education and prior achievement as random variables in the fitted model.

Supplementary statistics. The proportion of variance explained at level 2 can be calculated by entering school variance estimates from the intercepts- and slopes-as-outcomes and random coefficient models into the equation, $\hat{\tau}_{00}$ (random coefficient) - $\hat{\tau}_{00}$ (fitted model)/ $\hat{\tau}_{00}$ (random coefficient). Thus, 1077.41 – 250.75 / 1077.41 = .767, which indicates that school SES and teacher experience accounted for approximately 77% of the between-school parameter variance in school mean reading achievement (Raudenbush & Bryk, 2002). An analysis of the one-way ANOVA model found that 25% of the variance in reading achievement was between schools. If 77% of that variance was due to SCHSES and TEACHEXP, then the remaining 23% (i.e., 5.75% of overall variance) can be attributed to school effects for this sample. Furthermore, the significant chi-square test ($\chi^2 = 2746.33, p < .001$) "indicates that other school-level factors, such as school practice, may account for the remaining part of the variance" (Yu & White, 2002, p. 25).

Model fit. Raudenbush and Bryk (2002) caution against the overspecification of random coefficients because the explained variance is difficult to interpret when it is parsed into minute parts. According to their guidelines, the specification of two random coefficients and one random intercept for a sample with 254 schools and 51 students per school was well justified.

Although the SES-achievement slope varied across schools in the random coefficient model, it was not statistically significant when specified as a random level-1 predictor (with group mean centering) in the original intercepts- and slopes-as-outcomes model. Thus, SES was specified as a fixed variable in the final model, which was further supported by a comparison between the deviances associated with the two models.

Specifically, the more complex model in which SES was specified as random had a deviance of 128,501.04 with 11 parameters. The simplified model with SES as a fixed variable had a deviance of 128,502.21 with 7 parameters. The reduction in deviance of 1.17 with four degrees of freedom was not statistically significant. Therefore, "the simpler model seems justified. We infer that the explanatory power is not significantly enhanced by specifying the residual SES-achievement slopes as random" (Raudenbush & Bryk, 2002, p. 85). Finally, the fitted model analysis was expedited in only 11 iterations, which is well within the standard set forth by Raudenbush and Bryk (2002).

School Effects Estimation

According to the research, there are two types of school effects commonly estimated in SER studies. Type A school effects address the question, "how well would we expect a student with average background characteristics to perform in school *j* relative to the grand mean" (Willms & Raudenbush, 1989, p. 213). Type A effects measure the impact of school context and school practice, as only student characteristics are controlled (Willms & Raudenbush, 1989). Type B school effects ask, "how well a particular school performed relative to other schools with similar student intakes, contextual effects, and wider social influences" (Willms & Raudenbush, 1989, p. 213). Type B effects are the focus of most school accountability systems because they assess individual schools on policies and practices that they can control (Raudenbush & Willms, 1995).

Performance Differences

To provide a framework for identifying under and overperforming schools, actual scale scores were compared with fitted or predicted values on the CSAP fourth-grade

reading test (see Appendix E). Performance differences ranged from 46.27 points below the expected score to 76.61 points above the predicted value. However, the mean difference between actual and predicted scale scores for this sample was .04. There also was a strong positive correlation ($r = .85$) between the predicted values and actual scores for these schools. Based on this analysis, a total of 123 schools were identified as overperforming (i.e., exceeded predicted mean score) while 131 schools were identified as underperforming (i.e., trailed predicted mean score).

There was a statistically significant difference ($p = .019$) in the fitted scores for these two groups, with the overperforming group predicted to score 582.03 and the underperforming group expected to score 574.80. For the outcome measure, there was a statistically significant difference ($p < .001$) between these two groups, with the overperforming group averaging 596.17 and the underperforming group averaging 561.61. Thus, schools in the overperforming group scored an average of 14.14 points above their fitted value while schools from the underperforming group averaged 13.19 points below their predicted score.

The two groups varied only on school SES and prior achievement, as there were no significant differences for teacher education, teacher experience, student-teacher ratio, and school special education status. Specifically, the overperforming group had an average of 40.9% of students eligible for free/reduced lunch benefits while the underperforming group had an average of 49.4% of students eligible. The overperforming group averaged 571.77 on the CSAP third-grade reading test while the underperforming group averaged only 541.48.

School Rankings

Schools were ranked on value-added school effects estimates generated from the residuals for the prior achievement and special education slopes. However, the contribution of special education status was almost negligible, as prior achievement accounted for almost all of the variance in school effects. Furthermore, it is more common to interpret school effects using prior achievement or student aptitude (Pituch, 1999; Yu & White, 2002). Thus, special education status was dropped from the school effects analysis. The appropriate equation to estimate school effects in this case was u_{0j} + u_{5j}(PRIOR) where u_{0j} is the effect of school practice on the adjusted level-1 mean, and u_{5j} is the effect of school practice on the within-school slope (Pituch, 1999). The EB residuals for the intercept and PRIOR-achievement slope are displayed in Appendix F.

To estimate school effects, three prior achievement values were selected: the grand mean (PRIOR = 556.14), one standard deviation below the grand mean (PRIOR = 520.36), and one standard deviation above the grand mean (PRIOR = 591.92). Estimates of school effects and the respective school rankings are presented in Appendix G. At the average prior achievement level, the school effects ranged from –93.82 to 101.00 with a mean of zero. To interpret these value-added residuals, a school effects estimate of 14.40 indicates that school practice helped students with average prior achievement to score 14.4 points above the grand mean (Pituch, 1999). However, schools with negative school effects estimates are simply adding less value to student achievement than would be expected based on their school context and student characteristics (Bingham et al., 1991).

As displayed in Table 14, schools were compared on performance differences and school effects rankings. There were overperforming schools that ranked very high on

school effects (e.g., School 158) and underperforming schools that ranked very low (e.g., School 41). However, the school that posted the largest performance gain (i.e., School 59) had a school effects estimate of –28.99, which ranked 213 out of 254 schools. The school with the second largest performance deficit (i.e., School 182) had a school effects estimate of 42.92 and a ranking of 24. In addition, there was no apparent relationship between school proficiency levels and performance differences and rankings. Specifically, over and underperforming schools and schools with positive and negative school effects estimates were distributed across all proficiency levels (i.e., unsatisfactory, partially proficient, proficient, and advanced) on the CSAP test.

Table 14

School Effects Estimates for Over and Underperforming Schools

School	Performance Difference	School Effects Estimate	Rank
59	+76.61	-28.99	213
158	+25.57	63.54	6
49	+22.94	-77.76	253
79	+10.07	101.00	1
132	-46.27	-10.88	160
182	-41.51	42.92	24
41	-31.04	-93.82	254
252	-8.45	83.22	2

There were no discernable patterns in school context for schools of different performance levels and effectiveness ranks. For example, the school that posted the third largest performance difference (i.e., School 49) and the school that had largest performance deficit (i.e., School 132) each had 75% of students eligible for free/reduced lunch benefits. The school ranked first on value-added school effects (i.e., School 79) had

experienced and well educated teachers while the school ranked second (i.e., School 252) had inexperienced and less educated teachers.

Differential Effectiveness

According to Pituch (1999), "the use of a uniform effects model, which implies that the school practice effect is constant across the range of the student attribute, is appropriate only when the school residual term for each slope is zero for all schools" (p. 198). Thus, the crux of differential effectiveness is the notion that "school effects differ on the different values of the slope instead of being uniform across all the students in a school" (Yu & White, 2002, p. 26). It is clear from the school effects equation that if slopes are ignored and schools are ranked only on the intercepts, then schools with differential effectiveness might be incorrectly ranked (Pituch, 1999).

As expected, most schools had different school effects estimates and rankings across prior achievement levels (Yu & White, 2002). For example, School 1 had a residual estimate of –18.27 for u_{0j} and a residual estimate of .12 for u_{5j}. Substituting the three prior achievement values into the school effects equation for School 1, –18.27 + .12 (PRIOR), yielded school effects of 42.19 for students with below average prior achievement, 46.35 for students with average prior achievement, and 50.51 for students with above average prior achievement. Even with the large negative intercept, the large positive value of u_{5j} resulted in a very high overall ranking and demonstrated that School 1 was more effective in adding value to the test scores of higher achieving students. This result was reversed for School 3, in that the negative intercept (u_{0j} = -4.40) and slope (u_{5j} = -.02) resulted in lower achieving students gaining more value from school practice.

CHAPTER 5: DISCUSSION

Although somewhat exploratory in design, the present study confirmed the most powerful and consistent predictors of student achievement while validating a multilevel modeling approach for conducting SER. The following conclusions, interpretations, implications, limitations, and recommendations are derived from these key findings.

Conclusions

Before conclusions about student achievement predictors, explained variance in HST results, and school accountability systems are presented, the HLM assumptions detailed in Chapter 4 are recalled to facilitate systematic inquiry into their validity. Assumptions one, three, and five were supported by the analyses of within-school errors and residual school effects. Specifically, the findings indicate that level-1 errors had equal variance, level-2 random effects were normally distributed, and level-2 predictors had linear relationships with the corresponding intercept and slope outcomes. However, assumptions two and four were harder to test and served more as conceptual guidelines in the specification of the level-1 and level-2 models (Raudenbush & Bryk, 2002).

Significant Predictors

The results from the HLM analysis showed that SES, special education status, ethnicity, mobility, and prior achievement were significant level-1 predictors of student achievement on the CSAP fourth-grade reading test. As anticipated, school SES accounted for most of the level-2 variance. Teacher experience and teacher education

were significant, albeit subtle, level-2 predictors of CSAP scores, which validated the inclusion of teacher quality as a school-level variable.

In light of recent research confirming that teacher certification is a strong predictor of elementary student achievement (e.g., Goldhaber & Anthony, 2004), a surprising finding was the absence of a significant effect for this school-level input. With most schools in the sample reporting over 90% teacher certification, it is possible that the variable was overly broad by including several licensure designations. If teacher certification was more precisely defined (e.g., national certification), its usefulness in SER might be enhanced. The same is true for student-teacher ratio, in that a more sensitive measure of class size might reveal a significant effect for this level-2 variable.

Explained Variance

As shown in Table 15, 75% of the variance on CSAP fourth-grade reading test scores was within schools and 25% was between schools. At level-1, SES, special education status, prior achievement, ethnicity, and continuous enrollment accounted for 64% of the within-school variance in student achievement. At level-2, school SES and teacher experience accounted for 77% of the between-school variance in test scores.

For the present study, approximately 6% of the overall variance in fourth-grade reading performance was attributed to school practice. This finding is comparable to other HLM analyses conducted with a SER approach. One explanation for why this percentage is not greater is that high-stakes tests traditionally do a poor job of measuring school practice.

Table 15

Percentage of Variance Explained by HLM Analysis

Predictor	Within-School	Between-School	Overall
Unconditional	75%	25%	100%
Student-level	64%		48%
Random Error	36%		27%
School-level		77%	19%
School Practice		23%	6%

School Accountability

The findings from this dissertation can be used to bolster both sides of the debate over test-based school accountability. The strong influence of SES and prior achievement underscores the assertion that high-stakes tests do not accurately measure student learning. Furthermore, the lack of a definitive interpretation of school effects bolsters the argument against using a single test to determine the fate of schools and students. However, the results could be interpreted to provide support for using test scores as evaluative and diagnostic measures for school improvement.

The present study also provides encouragement to adherents and critics of school effectiveness research. The findings indicate that schools with similar contexts and demographics produce different achievement levels and school effects. Thus, the SER hypothesis that school practice is responsible for some of the variance in student achievement was supported. However, this investigation was unable to answer many of the questions regarding the psychometric, methodological, and interpretative limitations of SER.

Interpretations

The following interpretations about the conceptual framework, achievement gaps, and school effects should be considered in context of the myriad permutations for hierarchical linear modeling. That being said, the methods and results were fully consistent with similar studies from the SER evidence base.

Theoretical

The conceptual framework was empirically supported, as the outcome variable was shown to be a function of school context, student characteristics, within-school random error, and unmeasured school practice. School context was directly related to student achievement and interacted with certain student characteristics to have an indirect effect on CSAP test scores. Student characteristics also had both a direct and indirect association with the dependent variable. After accounting for these relationships and controlling for random error, the remaining variance was attributed to school practice. However, the theoretical model relies on "the implausible assumption that schools of varying context are assigned at random to elements of school practice" (Raudenbush & Willms, 1995, p. 313).

Achievement Gaps

For this sample, affluent schools were more effective in mitigating the achievement gaps related to student SES, special education status, and prior achievement than were less affluent schools. This interpretation is inconsistent with Raudenbush and Bryk's (2002) finding that high-poverty schools were more equitable in regard to the SES-achievement relationship.

Although schools with better-educated teachers were more equitable on the SES-achievement slope, schools with experienced teachers were less equitable in this regard. One possible interpretation is that schools with veteran teachers are less likely to adopt innovative instructional and organizational strategies, where schools with less experienced educators are more apt to implement reforms designed to help disadvantaged students keep pace with their peers.

School Effects

The most interesting aspect of this study was the analysis of school effects estimates. Although expected to score 7.23 points higher than the underperforming group on the CSAP fourth-grade reading test, schools from the overperforming group actually averaged 34.56 points higher. The groups were quite similar on all school characteristics except for SES and prior achievement, which favored the overperforming schools. Although these variables likely accounted for some of the performance difference, school practice also contributed to the variation in student achievement.

The most provocative interpretation is that schools identified as overperforming based on the analysis of predicted values were ranked significantly lower than were underperforming schools on the analysis of value-added school effects. Specifically, the mean rank of overperforming schools for students with below average prior achievement was 142 while the mean rank for underperforming schools was 114. This same trend held, and in fact became more pronounced, for students with average and above average prior achievement. This may be a statistical artifact of the slope relationship, however, as the overperforming group had significantly higher mean prior achievement than did the underperforming group.

A differential effectiveness analysis also was conducted to explore the interaction between prior achievement and achievement on the CSAP fourth-grade reading test. As in previous studies (e.g., Yu & White, 2002), there was a differential effect of school practice on students within a school based on prior achievement. Specifically, positive school effects were more marked for students with above average prior achievement. Thus, school accountability systems that rank schools on a single residual term are inadequate in regard to estimating school effects (Pituch, 1999).

Each of these school effects measures has strengths and weaknesses. For example, school rankings based on value-added residuals are harder to explain to stakeholders but are more sensitive to achievement gap issues. Although more intuitive, the performance difference analysis has its flaws, in that schools identified as overperforming had more advantageous input variables than did underperforming schools. Perhaps the fairest interpretation is that all schools have something to offer in regard to school effectiveness.

Implications

This dissertation has theoretical and practical ramifications for researchers in the areas of high-stakes testing, multilevel modeling, and school effectiveness research. The study also has potential implications for politicians and educators in regard to school accountability policy and classroom practice.

Research

The present study was successful in building upon the best evidence from the HLM and SER disciplines. Specifically, this study extended Ma and Klinger's (2000) research by specifying a more replicable model based on accessible school-level predictors in a U.S. educational setting. This research added value to Yu and White's

(2002) work by incorporating essential school context and student characteristic variables to account for more of the variance in student achievement on high-stakes tests. Finally, this study improved upon the external and internal validity of Wenglinsky's (2002) research, by using statewide assessments and fourth-grade test scores to measure school effects.

Beyond corroborating prior school effectiveness research, this dissertation opened up new avenues for exploration. The collaboration between educational researchers from local and state entities facilitated a wider consideration of student achievement predictors, which allowed for a more accurate estimation of school effects. A potential implication is that additional foresight and funding will be given to the development and synthesis of secondary databases. A possible unintended but positive consequence is that multilevel modeling could become a more prevalent course of study in schools of education, so future researchers are better prepared to meet the challenges posed by high-stakes testing.

Policy

Perhaps the strongest implication is that school accountability systems that do not account for student characteristics and school context in the analysis of high-stakes tests are inherently unfair (Yu & White, 2002). Specifically, comparisons of uncorrected test scores have led to schools being sanctioned based on inaccurate data interpretations (AERA, 2000). As a result, state policymakers face mounting pressure to control for educational inputs in the interpretation and reporting of HST data (Reichardt, 2001). In addition, the continued implementation of NCLB will have policy ramifications for years to come, as new financial and administrative burdens stretch educational budgets to the breaking point.

Practice

The implications of high-stakes testing often are articulated in the stated beliefs and perceptions of teachers and administrators (Hoffman et al., 2001). For example, most teachers believe that poor test scores will negatively impact their job security and compensation (Barksdale-Ladd & Thomas, 2000). Many teachers perceive that HST will decrease their autonomy while raising questions about their professional judgment (Langenfeld et al., 1996). Educators also perceive that HST is responsible for the looming teacher shortage, the declining number of administrator candidates, and the migrating of students to private schools (Waite, et al., 2001).

The present study has fewer implications for practitioners, in that teachers have little or no control over the demographics of the student population in their school. However, the results demonstrate that teacher quality has a significant effect on student achievement in the primary grades. On a more practical note, the slope interpretations are somewhat instructive for educators by identifying where schools are being differentially effective.

Limitations

Many authors have concluded that school effectiveness research is not an exact science (e.g., Pituch, 1999). Willms and Raudenbush (1989) readily acknowledge the methodological, technical, and interpretative limitations of using HLM to measure school effects on high-stakes tests.

Methodological

The foremost methodological limitation is that SER is correlational and does not provide definitive evidence of causality regarding the impact of predictors on student

achievement (Battistich et al., 1995). Bryk and Raudenbush (1988) also concede that HLM cannot adequately represent every type of school effect impacting test scores. Similar to traditional regression analyses, multilevel modeling is limited by inappropriate sampling procedures (Willms & Raudenbush, 1989), unreliable instrumentation (Stringfield & Herman, 1996), and nonrandom assignment (Bernstein, 1990). Additionally, cross-sectional research is based on the assumption that "end-of-year test results capture that year's contribution. If the contributions are lagged and show up in later years, estimates will be inaccurate" (Bingham et al., 1991, p. 201).

Statistical

The most serious statistical limitation is that the reliability of Type B school effects estimates is undermined by unmeasured school practice (Raudenbush & Willms, 1995). As data on school practice were not available in this study, it is possible that some of its effect was confused with school context (Willms & Raudenbush, 1989). Furthermore, school effects are underestimated when schools with more advantaged students also have more resources for instruction, curriculum, and organization, as too much credit goes to school context and too little is given to school practice (Raudenbush & Willms, 1995). In addition, there is still uncertainty about the stability of school effects because of the inconsistent performance of schools from year-to-year (Mandeville & Heidari, 1988).

As for other technical limitations, Wang (1999) argues that HLM creates statistical artifacts that cloud the analysis of HST data when too many fixed effects are specified. According to de Leeuw and Kreft (1995), HLM may not be appropriate for very large samples that have small intraclass correlations. Hierarchical linear modeling

also is constrained by the requirement that only one output variable be modeled during a specific analysis (Schagen, 1991). Finally, the lack of validity for high-stakes tests limits the statistical claims of SER (Domenech, 2000).

Interpretative

The precise interpretation of school effects estimates is most threatened by the inadequate specification of educational inputs. According to Willms and Raudenbush (1989), "unless the control variables capture all of the effects associated with selection into schools . . . high SES schools appear to perform better than they are and low SES schools appear to perform worse" (p. 228). Furthermore, SER does not always account for the notion that schools can have different effects within the same building for students of diverse abilities and backgrounds (Pituch, 1999).

School effectiveness research suffers from unwarranted interpretations because of the unreliability of value-added estimates used to rank schools (Goldstein, 1997). Specifically, the instability of school effects results from organizational flux, instructional change, teacher turnover, and wider social, economic, and political influences (Willms & Raudenbush, 1989). Furthermore, it is unclear how much of the variance between schools should be attributed to constant and organized school properties (Yu & White, 2002).

As school rankings provide a relative benchmark for school quality rather than an absolute criterion of excellence, interpretations fail to "identify either what makes particular teachers successful or even any of the in-class characteristics of success beyond high levels of test performance" (Bingham et al., 1991, p. 201). Lastly, predictive models may not be generalizable to other school settings based on differing definitions and interactions of educational input and output variables (Bingham et al., 1991).

Recommendations

Although the present study has perhaps raised more questions than it has answered, the research approach has hopefully paved the way for more advanced empirical work in this area. To facilitate this process, recommendations for accessing educational inputs, reporting HST results, and conducting future research are discussed.

School Practice

According to Raudenbush and Willms (1995), the biggest challenge facing SER is that "it is far more difficult to adequately measure school practice than to obtain good measures of family background and prior student aptitude or achievement" (p. 313). Thus, a major recommendation is that quantitative measures for school instruction, curriculum, and organization should be conceptualized, operationalized, and collected. For example, powerful level-2 predictors such as academic press should be made more accessible. Finally, a more accurate proxy for SES than free/reduced lunch status would improve the fit of hierarchical linear models designed to measure school effects.

Reporting

The Colorado school accountability report is a good model for presenting HST results in a concise, understandable, and meaningful way. Specifically, the inclusion of student-, school-, and district-level data along with performance comparisons for similar schools provides educational consumers with important information (Schafer et al., 2000). This type of reporting mechanism would be improved by the incorporation of subgroup analyses and the jargon-free explanation of data analysis procedures. A more open and transparent reporting of HST results would assist policymakers in garnering public and professional support for school accountability systems.

Future Research

The most obvious recommendation is that future studies should employ HLM to account for the nested nature of HST data. Pituch (1999) argues that multiple outcome measures from different content areas are needed to triangulate the school effectiveness evidence generated by a multilevel modeling approach. Additionally, researchers might add a third level to the HLM analysis to explore district-level inputs such as PPE.

Although SER typically focuses on the secondary grades, the cumulative nature of education calls for more elementary school research to be conducted (Ma & Klinger, 2000). Of course, longitudinal studies would provide more stable results while allowing for a value-added tracking of student achievement over time (Sanders, 1998). The use of random sampling in future SER would allow predictive models to be applicable in other educational settings. Future research also is needed to explore the statistical and theoretical assumptions underlying the various school effectiveness measures.

According to Yu and White (2002), the objective of SER "is to identify effective school characteristics so as to disseminate them within similar schools and enhance the quality of public education" (p. 32). However, quantitative research seeks only to answer the "what works" question and not the "why does it work" query. Therefore, grounded theory, ethnography, and other qualitative research designs are needed to gain a better understanding of how school effects operate (Coe & Fitz-Gibbon, 1998). For example, in-depth case studies could identify the instructional, organizational, and curricular practices responsible for high and low performance on high-stakes tests (Yu & White, 2002). Once trends in effective school practice are illuminated, outreach and training with practitioners can begin so that students truly reap the benefits of school accountability.

REFERENCES

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in pre K-12 education. *Educational Researcher, 29*(8), 24-25.

Ascher, C., & Fruchter, N. (2001). Teacher quality and student performance in New York City's low-performing schools. *Journal of Education for Students Placed At Risk, 6*, 199-214.

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education, 51*, 384-397.

Battistich, V., Solomon, D., Kim, D, Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance: A multilevel analysis. *American Educational Research Journal, 32*, 627-658.

Bernstein, L. (1990). *Developing an adequately specified model of state level student achievement with multilevel data.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

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

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*, 65-108.

Burns, M. (1998). *Interpreting the reliability and validity of the Michigan Educational Assessment Program.* Saginaw, MI: Michigan Association of School Psychologists. (ERIC Document Reproduction Service No. ED 418138)

Caldas, S. J. (1993). Reexamination of input and process factor effects on public school achievement. *Journal of Educational Research, 86*, 206-214.

Centra, J. A., & Potter, D. A. (1980). School and teacher effects: An interrelational model. *Review of Educational Research, 50*, 273-291.

Coe, R., & Fitz-Gibbon, C. T. (1998). School effectiveness research: Criticism and recommendations. *Oxford Review of Education, 24,* 421-438.

Coleman, A. L. (2000). Fair testing. *American School Board Journal, 187*(6), 32-35.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: Government Printing Office.

Crone, L. J., Lang, M. H., Franklin, B. J., & Halbrook, A. M. (1994). Composite versus component scores: Consistency of school effectiveness classification. *Applied Measurement in Education, 7*, 303-321.

Darling-Hammond, L. (1991). The implications of testing policy for quality and equality. *Phi Delta Kappan, 73*, 220-225.

Darling-Hammond, L. (2000, January 1). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved January 10, 2001, from http://epaa.asu.edu/epaa/v8n1/

Darling-Hammond, L. (2002, September 6). Research and rhetoric on teacher certification. *Education Policy Analysis Archives, 10*(36). Retrieved October 9, 2002, from http://epaa.asu.edu/epaa/v10n36.html

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher, 31*(9), 13-25.

de Leeuw, J., & Kreft, I. G. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics, 20*, 171-189.

Domenech, D. A. (2000). My stakes well done. *School Administrator, 57*(11), 16-19.

Dounay, J. (2000). High-stakes assessments bring out the critics. *State Education Leader, 18*(1), 4-6.

Ediger, M. (2001). *Assessing state report cards on student achievement*. Columbia, MO. (ERIC Document Reproduction Service No. ED452260)

Glasman, N. S., & Biniaminov, I. (1981). Input-output analyses of schools. *Review of Educational Research, 51*, 509-539.

Gliner, J. A., & Morgan, G. A. (2000). *Research methods in applied settings: An integrated approach to design and analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Goddard, R. D., Sweetland, S. R., & Hoy, W. K. (2000). Academic emphasis of urban elementary schools and student achievement in reading and mathematics: A multilevel analysis. *Educational Administration Quarterly, 36*, 683-702.

Goldhaber, D. (2002, Spring). The mystery of good teaching. *Education Next, 2*(1). Retrieved December 19, 2002, from http://www.educationnext.org/20021/50.html

Goldhaber, D., & Anthony, E. (2004, March 8). *Can teacher quality be effectively assessed?* Retrieved March 24, 2004, from Urban Institute Web site: http://www.urban.org/url.cfm?ID=410958

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

Gordon, S. P., & Reese, M. (1997). High-stakes testing: Worth the price? *Journal of School Leadership, 7*, 345-368.

Gottlieb, A. (2002). Economically segregated schools hurt poor kids, study shows. *The Term Paper, 1*(2), 1-2, 5-6.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*, 361-396.

Griffith, J. (1996). Relation of parental involvement, empowerment, and school traits to student academic performance. *Journal of Educational Research, 90*, 33-41.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Haladyna, T. M., Haas, N. S., & Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education, 74*, 262-273.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*, 1141-1177.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45-52.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Education Evaluation and Policy Analysis, 19*, 141-164.

Harter, E. A. (1999). How educational expenditures relate to student achievement: Insights from Texas elementary schools. *Journal of Education Finance, 24*, 281-303.

Heinlein, L. M., & Shinn, M. (2000). School mobility and student achievement in an urban setting. *Psychology in the Schools, 37*, 349-357.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12*(4), 20-25, 41-42.

Hess, F. M., & Brigham, F. (2000). None of the above. *American School Board Journal, 187*(1), 26-29.

Hoffman, J. V., Assaf, L. C., & Paris, S. G. (2001). High-stakes testing in reading: Today in Texas, tomorrow? *Reading Teacher, 54*, 482-492.

Holdaway, E. A., & Johnson, N. A. (1993). School effectiveness and effectiveness indicators. *School Effectiveness and School Improvement, 4*, 165-188.

Hubler, E. (2002, May 3). Latest scores bring joy, concern to Denver's principals, staffs. *The Denver Post*, p. 22A.

Jennings, T. A., Kovalski, T. M., & Behrens, J. T. (2000). *Predicting academic achievement using archival mobility data.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Johnson, R. A., & Lindblad, A. H. (1991). Effect of mobility on academic performance of sixth grade students. *Perceptual and Motor Skills, 72*, 547-552.

Jones, M. G., Jones, B. D., & Hardin, B. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81*, 199-203.

Kiplinger, V. L. (1998). *Assessing the CSAP: How can we assess the quality of a large-scale, standards-based assessment?* Paper presented at the annual Standards and Assessments Conference, Breckenridge, CO.

Kohn, A. (2000). Burnt at the high stakes. *Journal of Teacher Education, 51*, 315-327.

Kreft, I. G., de Leeuw, J., & van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *American Statistician, 48*, 324-335.

Laczko-Kerr, I., & Berliner, D. C. (2002, September 6). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case for harmful public policy. *Education Policy Analysis Archives, 10*(37). Retrieved October 9, 2002, from http://epaa.asu.edu/epaa/v10n37/

Langenfeld, K. L., Thurlow, M. L., & Scott, D. L. (1996). *High-stakes testing for students: Unanswered questions and implications for students with disabilities* (Synthesis Report No. 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 24, 2001, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis26.htm

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*, 37-62.

Lattimore, R. (2001). The wrath of high-stakes tests. *Urban Review, 33*(1), 57-67.

Lindjord, D. (2001). Overhauling public schools: President Bush's education proposal and the effects on children and families. *Journal of Early Education and Family Review, 8*(5), 5-6.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29-36.

Ma, X., & Klinger, D. A. (2000). Hierarchical linear modeling of student and school effects on academic achievement. *Canadian Journal of Education, 25*, 41-55.

Madaus, G. F. (1991). The effects of important tests on students: Implications for a national examination system. *Phi Delta Kappan, 73*, 226-231.

Mandeville, G. K., & Heidari, K. (1988). *Measuring school effectiveness using hierarchical linear models.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

McGee, G. W. (1997). *What state tests test.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Nick, S. (2002, December 1). Schools focus on minorities, poor: CSAP results show lower test scores. *Fort Collins Coloradoan*, pp. B1, B7.

Noble, A. J., & Smith, M. L. (1994). Old and new beliefs about measurement-driven reform: "Build it and they will come." *Educational Policy, 8*, 111-136.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice, 11*(2), 9-15.

Norton, M. S. (1998). Teacher absenteeism: A growing dilemma in education. *Contemporary Education, 69*, 95-99.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2002). Do low-achieving students benefit more from small classes? Evidence from the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 24*, 201-217.

Oakes, J. (1989). What educational indicators? The case for assessing the school context. *Educational Evaluation and Policy Analysis, 11*, 181-199.

Paul, P. (2002). Raising our standards. *American Demographics, 24*(6), 12.

Pituch, K. A. (1999). Describing school effects with residual terms: Modeling the interaction between school practice and student background. *Evaluation Review, 23*, 190-211.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Popham, W. J. (2000). The score-boosting game. *American School Board Journal, 187*(6), 36-39.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2001). *HLM 5: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

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

Reichardt, R. (2001). *Toward a comprehensive approach to teacher quality.* Aurora, CO: Mid-Continent Research for Education and Learning.

Riddle Buly, M., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis, 24*, 219-239.

Rose, L. C., & Gallup, A. M. (2000). The 32nd annual Phi Delta Kappa/Gallup poll of the public attitudes toward the public schools. *Phi Delta Kappan, 82*, 41-57.

Ryan, K. (2002). Shaping educational accountability systems. *American Journal of Evaluation, 23*, 453-468.

Sanders, W. L. (1998). Value-added assessment. *School Administrator, 55*(11), 24-27.

Schafer, W. D., Yen, S. J., & Rahman, T. (2000). School effects indices: Stability of one- and two-level formulations. *Journal of Experimental Education, 68*, 239-250.

Schagen, I. (1991). Beyond league tables. How modern statistical methods can give a truer picture of the effects of schools. *Educational Research, 33*, 216-222.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*, 232-238.

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

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*, 334-344.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11.

Strand, S. (1997). Pupil progress during key stage 1: A value added analysis of school effects. *British Educational Research Journal, 23*, 471-487.

Stringfield, S., & Herman, R. (1996). Assessment of the state of school effectiveness research in the United States of America. *School Effectiveness and School Improvement, 7*, 159-180.

Thompson, S. (2001). The authentic standards movement and its evil twin. *Phi Delta Kappan, 82*, 358-362.

Waite, D., Boone, M., & McGhee, M. (2001). A critical sociocultural view of accountability. *Journal of School Leadership, 11*, 182-203.

Walker, S. F. (2000). High-stakes testing: Too much? Too Soon? *State Education Leader, 18*(1), 1-3.

Wang, J. (1999). Reasons for hierarchical linear modeling: A reminder. *Journal of Experimental Education, 68*, 89-93.

Wenglinsky, H. (1997). How money matters: The effect of school district spending on academic achievement. *Sociology of Education, 70*, 221-237.

Wenglinsky, H. (2002, February 13). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10*(12). Retrieved October 9, 2002, from http:// epaa.asu.edu/epaa/v10n12/

Whaley, M. (2002, February 7). Delay sought in school report cards. *The Denver Post*, p. 14A.

Whaley, M., & Hubler, E. (2002, May 3). 3rd-graders show gains on CSAP. *The Denver Post*, pp. 1A, 26A.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*, 209-232.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Yen, W. M., & Ferrara, S. (1997). The Maryland School Performance Assessment Program: Performance assessment with psychometric quality suitable for high stakes usage. *Educational and Psychological Measurement, 57*, 60-84.

Yettick, H. (2002, September 3). CSAP 'no scores' may hurt schools. *Rocky Mountain News*, pp. 4A, 13A.

Yu, L., & White, D. B. (2002). *Measuring value added school effects on Ohio sixth-grade proficiency test results using two-level hierarchical linear modeling.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

APPENDIX A – SCHOOL DISTRICT CONSENT FORM

RE: Approval for Release of Just For the Kids Data

This letter demonstrates support for the study proposed by Marc Winokur, a doctoral candidate in Educational Leadership at Colorado State University. I have read the attached cover letter and dissertation abstract and I am comfortable with the scope and process of the research approach as outlined. I also understand the objective of the study, which is to determine the most powerful and consistent predictors of student achievement on the Colorado Student Assessment Program. Furthermore, the information from this study will be helpful to the students, teachers, and administrators in my school district.

I am aware that no new data will be collected during this study and that my only role is to provide permission for the release of the requested data from the 2001-2002 school year. I am satisfied that the confidentiality of students, schools, and districts will be strictly protected in this study. In addition, I acknowledge that my district's participation in this study is voluntary and that we may choose to end our involvement at any time.

My signature below indicates approval for Just for the Kids to release student CSAP and demographic data from my school district to Marc Winokur and Dr. Brian Cobb at the Research & Development Center for the Advancement of Student Learning.

_____ _____

Superintendent's Signature Date

APPENDIX B – LETTERS OF COOPERATION

National Center for Educational Accountability

Sponsoring organizations:
The Education Commission of the States
The University of Texas at Austin
Just for the Kids

RE: Approval for Release of Just for the Kids (JFTK) Data

This letter of cooperation confirms that Just for the Kids has received letters with original signatures from 14 Colorado superintendents or their designees granting permission for JFTK to release demographic and achievement data from their school districts.

I am comfortable with the scope and approach of the research study as described and understand that the only role for JFTK is to provide access to the requested data. I am also satisfied that the confidentiality of students, schools, and districts will be strictly protected in this study. In addition, I acknowledge that JFTK's participation in this study is voluntary and that we may choose to end our involvement at any time.

My signature below indicates approval for JFTK to release, as per permission from 14 Colorado school districts, demographic and achievement data to Marc Winokur, a doctoral candidate in Educational Leadership at Colorado State University (CSU) and Dr. Brian Cobb, a Professor of Education at CSU and co-director of the Research & Development Center for the Advancement of Student Learning.

Chrys Dougherty — Print Name
Director of Research — Title

Chrys Dougherty — Signature
6/20/03 — Date

8030-2 W. Braker Lane, 2nd Floor, Austin, Texas 78759 ■ Phone: 512.232.0770 ■ Fax: 512.232.0777 ■ www.nc4ea.org ■ www.just4kids.org

COLORADO DEPARTMENT OF EDUCATION

201 East Colfax Avenue [Central Office 303.866.6600]
Denver, Colorado 80203-1704 • www.cde.state.co.us

William J. Moloney
Commissioner of Education

Roscoe Davidson
Deputy Commissioner

RE: Approval for Release of Colorado Department of Education (CDE) Data

This letter demonstrates support for the study proposed by Marc Winokur, a doctoral candidate in Educational Leadership at Colorado State University. I have spoken with Marc and I am comfortable with the scope and process of the research approach as described. I also understand the objective of the study, which is to determine the most powerful and consistent predictors of student achievement on the Colorado Student Assessment Program. Furthermore, the information from this study will be helpful to the students, teachers, and school administrators in the state of Colorado.

I am aware that the only role for CDE is to provide permission for Mid-Continent Research for Education and Learning (McREL) to release the requested data originally collected by CDE as part of regular administrative procedures. I am satisfied that the confidentiality of students, teachers, schools, and districts will be strictly protected in this study. In addition, I acknowledge that the participation of CDE in this study is voluntary and that we may choose to end our involvement at any time.

My signature below indicates approval for McREL to release aggregated teacher quality data from the 2001-2002 school year to Marc Winokur and Dr. Brian Cobb at the Research & Development Center for the Advancement of Student Learning.

Susan P. Schafer
Print Name

Regional Coordinator
Title

Susan P. Schafer
Signature

6-4-03
Date

cde *Improving Academic Achievement*

McREL
Making a Difference
2550 S. Parker Road, Suite 500 • Aurora, CO 80014-1678
303.337.0990 • Fax: 303.337.3005 • www.mcrel.org

RE: Approval for Release of Mid-Continent Research for Education and Learning (McREL) Data

This letter demonstrates support for the study proposed by Marc Winokur, a doctoral candidate in Educational Leadership at Colorado State University. I have spoken with Marc. I understand the objective of the study, which is to determine the most powerful and consistent predictors of student achievement on the Colorado Student Assessment Program. Furthermore, the information from this study will be helpful to the students, teachers, and school administrators of Colorado.

I am aware that my only role is to release the requested data from the *Teacher Supply and Demand in the State of Colorado* study conducted by McREL in 2003. I am satisfied that the confidentiality of students, schools, and districts will be strictly protected in this study. In addition, I acknowledge that McREL's participation in this study is voluntary and that we may choose to end our involvement at any time.

My signature below indicates approval for McREL to release aggregated teacher quality data from the 2001-2002 school year to Marc Winokur and Dr. Brian Cobb at the Research & Development Center for the Advancement of Student Learning.

Robert E Reichardt
Print Name

Senior Researcher
Title

Robert E Reichardt
Signature

6/12/03
Date

APPENDIX C – HUMAN SUBJECTS RESEARCH APPLICATION

Protocol Information

1. Objectives of proposed research – the main objective of the proposed research is to identify the most powerful, consistent, and efficient predictors of elementary student achievement. The primary goal of the study is to provide a fair and accurate interpretation of the Colorado Student Assessment Program (CSAP).
2. Source of participant population – participants are fourth grade students (during the 2001-2002 school year) from Colorado elementary schools that are involved with the Just for the Kids project (JFTK). Students are the participants in this study because their CSAP scores will serve as the dependent variable in the data analysis.
3. Number of participants – the accessible sample includes approximately 34,000 fourth-grade students from 477 elementary schools representing 17 school districts in Colorado.
4. Characteristics of participants – the accessible sample of Front Range elementary schools contains students with diverse demographic backgrounds.
5. Recruiting procedures – in March 2003, superintendents from the 17 selected districts were asked to provide permission for the student-level data collected by JFTK to be released to Dr. Brian Cobb at the Research & Development Center. Since these individuals represent districts already participating in JFTK, they should be willing to cooperate with this research.
6. Recruiting materials – each superintendent was sent a cover letter, dissertation abstract, and district permission form as recruiting materials.

7. Criteria for including or excluding participants – participants will be included in the study if they have complete student-level data. Students with disabilities and English language learners also will be included.
8. Rationale for using "at-risk" populations – elementary school students were chosen because demographic data and high-stakes testing results often are more complete at this level.
9. HRC agreement/approval letters – letters of approval from selected school districts are currently being collected and will be submitted upon receipt to JFTK.
10. Other pertinent matters – N/A
11. Location of study – the study will be conducted at Colorado State University.
12. Variables to be studied – At the student-level, prior achievement, ethnicity, special education status, continuous enrollment, and free/reduced lunch status will be studied. At the school-level, teacher quality indicators (e.g., certification, education, experience) and school demographics will be analyzed.
13. Data collection methods – two secondary databases and one publicly available dataset will be accessed to collect the appropriate student- and school-level variables. Data from the JFTK and Mid-Continent Research for Education and Learning (McREL) databases will be collected via email attachments of customized Microsoft Excel files. Data collected from the school accountability reports (SAR) available on the Colorado Department of Education (CDE) website will be transcribed and entered into an Excel worksheet.
14. Research activities – there are no proposed research activities with participants.
15. Equipment – N/A

16. Procedures causing stress – there are no anticipated causes for physical or emotional stress for participating students.
17. Biological samples description – N/A
18. De-briefing method and materials – district superintendents and school administrators will have the opportunity to review the findings from this study.
19. Other procedural aspects – N/A
20. Describe potential risks – there are no known risks for participating districts, schools, students, and administrators in this study.
21. Methods for minimizing risks – N/A
22. Other methods – N/A
23. Other matters relative to participant risk – N/A
24. Describe direct benefits – there are no direct or tangible benefits for participating districts, schools, students, and administrators in this study.
25. Describe benefits accruing to the class of participants – participating students, schools, and districts would benefit from a more equitable analysis and interpretation of the CSAP.
26. Describe benefits accruing to society-at-large – the Colorado school system would benefit from rigorous research on educational accountability measures.
27. Other aspects of benefits to participants – N/A
28. Describe how potential participants will be informed – there are no proposed activities for participants, so students will not be informed about the study.

29. Informed consent form – as no new data will be collected from students, and existing student-level data will not have identifiers attached, the proposal does not require documentation of informed consent.
30. Other aspects of the consent process – letters of cooperation from CDE, McREL, and JFTK giving consent for the release of data from the secondary databases will be collected and submitted for human research review upon receipt.
31. Methods used to protect confidentiality – the secondary databases from McREL and JFTK will be transmitted with school and district identifiers attached to the data, so that they can be linked together and with data from the SAR. Once all three databases are combined, dummy codes will be given to each school and district case so that the data are totally unidentifiable. The linked list containing the original identifiers then will be destroyed. The data collected from the secondary databases do not contain student identifiers, so there are no confidentiality concerns for participants. As the data are all in electronic form, the resultant files will be stored and maintained on the co-PIs home computer.
32. Plans for maintaining data – a summary of all relevant data will be prepared and submitted to the Principal Investigator in an anonymous format. The PI will keep the data summary in a locked location at the Research & Development Center where it will be available for audit. The PI will store the letters of cooperation in locked storage for three years after the conclusion of the study.
33. Retention of consent forms – N/A
34. Tape storage use – N/A
35. Other aspects of confidentiality – N/A

APPENDIX D – SPSS CODEBOOK

Student-level File

Variable Name		Position
SCH_ID	School identification number Measurement Level: Scale	1
ETHNIC	Ethnicity Measurement Level: Nominal Value Label 0 White (not Hispanic) 1 Minority	2
CEC	Continuous enrollment on campus Measurement Level: Nominal Value Label 0 3 years or more 1 2 years	3
SES	Free/reduced lunch status Measurement Level: Nominal Value Label 0 Not eligible for free/reduced lunch 1 Eligible for free/reduced lunch	4
ESL	English as second language status Measurement Level: Nominal Value Label 0 Not eligible for ESL services 1 Eligible for ESL services	5
SPECED	Special education status Measurement Level: Nominal Value Label 0 Not eligible for special education services 1 Eligible for special education services	6

PRIOR	2001 Grade 3 reading total scale score Measurement Level: Scale	7
READSS	2002 Grade 4 reading total scale score Measurement Level: Scale	8

School-level File

Variable Name		Position
SCH_ID	School identification number Measurement Level: Scale	1
SCHSES	Proportion of students in school that are eligible for free/reduced lunch benefits Measurement Level: Scale	2
TEACHEDU	Proportion of teachers in school that have a graduate degree Measurement Level: Scale	3
TEACHCRT	Proportion of teachers in school that are fully certified or have a Master certification Measurement Level: Scale	4
TEACHEXP	Mean years of experience for teachers in school Measurement Level: Scale	5
RATIO	Students per teacher ratio in grade 4 Measurement Level: Scale	6

APPENDIX E – ACTUAL AND PREDICTED TEST SCORES BY SCHOOL

Descriptive Statistics for Actual and Predicted Test Scores by School

School	Actual Score *M*	Predicted Score *M*	Difference *M*
1	541.36	564.23	-22.87
2	550.52	543.98	6.54
3	606.37	606.55	-.18
4	597.08	597.72	-.64
5	584.23	605.67	-21.44
6	599.00	601.66	-2.66
7	579.67	585.02	-5.35
8	537.04	537.69	-.65
9	544.18	541.21	2.97
10	535.46	541.73	-6.27
11	558.33	540.48	17.85
12	598.14	598.67	-.53
13	538.98	547.07	-8.09
14	627.44	613.59	13.85
15	637.62	612.09	25.53
16	619.45	595.35	24.10
17	632.79	608.77	24.02
18	586.56	585.54	1.02
19	619.16	614.15	5.01
20	595.88	595.14	.74
21	573.68	573.12	.56
22	564.80	569.12	-4.32
23	596.42	608.74	-12.32
24	628.63	608.37	20.26
25	539.44	550.46	-11.02
26	564.11	540.86	23.25
27	605.70	587.26	18.44
28	543.87	563.07	-19.20
29	615.61	602.98	12.63
30	547.88	587.21	-39.33
31	540.05	560.59	-20.54
32	568.95	559.53	9.42
33	622.00	610.04	11.96
34	575.30	599.21	-23.91
35	560.50	564.98	-4.48
36	555.69	554.90	.79
37	529.70	564.98	-35.28
38	581.73	587.07	-5.34
39	629.61	604.46	25.15
40	569.82	572.32	-2.50
41	504.10	535.14	-31.04

School	Actual Score *M*	Predicted Score *M*	Difference *M*
42	645.33	612.04	33.29
43	587.36	597.87	-10.51
44	564.84	556.44	8.40
45	555.09	541.55	13.54
46	525.32	545.25	-19.93
47	537.71	548.14	-10.43
48	542.72	537.39	5.33
49	578.10	555.16	22.94
50	578.18	590.13	-11.95
51	582.53	570.54	11.99
52	623.53	601.56	21.97
53	580.25	605.71	-25.46
54	641.23	610.77	30.46
55	606.32	597.79	8.53
56	520.91	538.19	-17.28
57	612.54	614.54	-2.00
58	623.21	610.53	12.68
59	669.79	593.18	76.61
60	609.92	603.21	6.71
61	592.46	584.07	8.39
62	532.44	569.93	-37.49
63	552.61	563.82	-11.21
64	569.65	564.01	5.64
65	552.98	563.06	-10.08
66	640.15	609.44	30.71
67	639.51	588.07	51.44
68	563.96	554.90	9.06
69	606.59	607.66	-1.07
70	535.98	541.71	-5.73
71	594.26	615.62	-21.36
72	524.58	557.39	-32.81
73	595.19	582.60	12.59
74	513.35	534.87	-21.52
75	582.60	567.71	14.89
76	573.85	606.43	-32.58
77	588.82	605.53	-16.71
78	607.30	608.58	-1.28
79	598.60	588.53	10.07
80	577.00	607.66	-30.66
81	568.58	567.10	1.48
82	525.84	536.87	-11.03
83	630.61	592.81	37.80
84	540.23	537.49	2.74

School	Actual Score *M*	Predicted Score *M*	Difference *M*
85	515.30	531.11	-15.81
86	555.20	545.02	10.18
87	570.65	574.44	-3.79
88	579.86	556.02	23.84
89	516.53	540.83	-24.30
90	586.34	583.46	2.88
91	534.93	546.44	-11.51
92	573.60	580.92	-7.32
93	552.12	569.58	-17.46
94	552.03	583.59	-31.56
95	525.57	546.02	-20.45
96	546.38	547.14	-.76
97	543.66	545.37	-1.71
98	553.54	538.89	14.65
99	563.95	587.22	-23.27
100	628.70	615.26	13.44
101	543.31	563.08	-19.77
102	539.08	545.27	-6.19
103	542.53	534.20	8.33
104	613.25	565.86	47.39
105	614.64	596.88	17.76
106	562.54	598.25	-35.71
107	610.12	607.65	2.47
108	577.78	573.27	4.51
109	587.50	586.57	.93
110	599.23	581.80	17.43
111	579.47	582.05	-2.58
112	617.37	612.49	4.88
113	626.06	611.69	14.37
114	589.55	594.98	-5.43
115	604.93	612.09	-7.16
116	574.39	551.51	22.88
117	615.83	605.92	9.91
118	552.24	573.68	-21.44
119	535.38	541.53	-6.15
120	622.89	602.63	20.26
121	596.56	583.55	13.01
122	555.68	561.64	-5.96
123	536.48	539.34	-2.86
124	595.71	573.53	22.18
125	631.57	609.39	22.18
126	562.55	590.78	-28.23
127	600.45	575.85	24.60

School	Actual Score *M*	Predicted Score *M*	Difference *M*
128	591.05	595.54	-4.49
129	590.00	590.51	-.51
130	587.16	593.95	-6.79
131	607.49	597.93	9.56
132	503.03	549.30	-46.27
133	582.21	587.40	-5.19
134	580.00	563.00	17.00
135	603.77	584.90	18.87
136	605.57	605.05	.52
137	556.26	555.44	.82
138	576.48	586.69	-10.21
139	580.11	584.12	-4.01
140	553.17	564.66	-11.49
141	604.64	608.01	-3.37
142	538.58	544.35	-5.77
143	625.64	602.93	22.71
144	576.75	588.11	-11.36
145	561.77	552.14	9.63
146	581.91	570.57	11.34
147	602.19	595.32	6.87
148	595.55	600.07	-4.52
149	586.37	582.62	3.75
150	612.30	602.01	10.29
151	518.33	532.90	-14.57
152	567.59	562.28	5.31
153	596.03	597.84	-1.81
154	593.98	596.39	-2.41
155	537.78	557.65	-19.87
156	546.62	547.94	-1.32
157	556.12	569.18	-13.06
158	599.29	573.72	25.57
159	581.97	565.86	16.11
160	600.09	606.91	-6.82
161	529.11	534.22	-5.11
162	613.74	610.89	2.85
163	537.59	543.33	-5.74
164	566.45	564.41	2.04
165	541.64	567.87	-26.23
166	560.13	580.96	-20.83
167	562.71	557.49	5.22
168	565.30	588.71	-23.41
169	571.00	587.39	-16.39
170	602.02	592.78	9.24

School	Actual Score *M*	Predicted Score *M*	Difference *M*
171	596.68	594.75	1.93
172	624.92	606.57	18.35
173	529.25	552.31	-23.06
174	603.04	583.35	19.69
175	572.64	587.14	-14.50
176	580.97	568.73	12.24
177	593.62	583.95	9.67
178	538.95	540.61	-1.66
179	603.66	608.80	-5.14
180	507.46	547.90	-40.44
181	548.31	566.58	-18.27
182	559.28	600.79	-41.51
183	503.43	538.29	-34.86
184	574.04	580.00	-5.96
185	603.22	603.85	-.63
186	592.21	593.31	-1.10
187	598.67	590.43	8.24
188	573.78	564.27	9.51
189	530.08	535.77	-5.69
190	599.11	595.25	3.86
191	556.18	555.80	.38
192	621.41	610.64	10.77
193	540.44	560.19	-19.75
194	524.64	553.56	-28.92
195	568.73	586.38	-17.65
196	612.45	605.70	6.75
197	587.63	581.75	5.88
198	578.39	582.75	-4.36
199	597.23	596.05	1.18
200	588.04	599.94	-11.90
201	550.19	540.31	9.88
202	551.51	554.20	-2.69
203	587.07	579.76	7.31
204	587.79	605.54	-17.75
205	574.82	586.99	-12.17
206	626.72	607.62	19.10
207	627.39	606.18	21.21
208	527.81	535.97	-8.16
209	508.49	536.54	-28.05
210	580.45	577.36	3.09
211	609.37	600.24	9.13
212	667.67	607.47	60.20
213	610.21	591.46	18.75

School	Actual Score *M*	Predicted Score *M*	Difference *M*
214	553.71	544.67	9.04
215	594.00	589.51	4.49
216	575.18	567.28	7.90
217	569.32	568.03	1.29
218	569.18	587.61	-18.43
219	597.51	603.91	-6.40
220	595.29	597.92	-2.63
221	517.35	532.73	-15.38
222	593.88	592.52	1.36
223	606.32	584.77	21.55
224	584.96	565.70	19.26
225	559.40	577.58	-18.18
226	608.06	592.44	15.62
227	610.70	607.25	3.45
228	589.49	605.80	-16.31
229	592.23	594.07	-1.84
230	599.23	607.30	-8.07
231	600.18	591.75	8.43
232	578.45	561.96	16.49
233	560.93	568.88	-7.95
234	610.22	598.96	11.26
235	539.81	541.18	-1.37
236	604.69	607.29	-2.60
237	589.89	572.57	17.32
238	627.54	608.29	19.25
239	649.49	605.19	44.30
240	576.49	593.27	-16.78
241	631.66	607.48	24.18
242	598.70	598.71	-.01
243	584.34	589.26	-4.92
244	532.14	537.17	-5.03
245	559.45	552.55	6.90
246	544.54	535.17	9.37
247	557.46	584.50	-27.04
248	643.70	614.68	29.02
249	600.34	602.03	-1.69
250	584.19	572.84	11.35
251	582.34	602.98	-20.64
252	536.78	545.23	-8.45
253	576.19	596.96	-20.77
254	562.23	538.84	23.39

APPENDIX F – EMPIRICAL BAYES RESIDUALS

EB Residual Estimates of School Intercepts and PRIOR-achievement Slopes

School	Intercept (u_{0j})	PRIOR-achievement Slope (u_{5j})
1	-18.27	.1162
2	9.67	.0344
3	-4.40	-.0163
4	-3.41	-.0716
5	-24.60	.0976
6	-6.06	-.0150
7	-6.41	.0286
8	1.71	.0228
9	5.15	.0683
10	-1.51	.0627
11	13.78	.1072
12	-3.39	.0360
13	-3.01	.0029
14	9.88	.0026
15	18.69	-.0420
16	19.03	-.1185
17	17.98	-.0602
18	-.70	.0432
19	2.95	-.0246
20	-1.13	-.0421
21	3.04	-.0635
22	-2.06	-.0513
23	-17.04	.0201
24	12.23	-.0098
25	-5.87	-.0405
26	23.35	-.0705
27	15.07	-.0496
28	-14.02	-.0013
29	8.10	.0306
30	-32.16	.1024
31	-15.12	.1471
32	12.19	-.0324
33	8.34	-.0754
34	-25.31	.0742
35	-.59	.0022
36	5.25	-.0900
37	-29.50	.1368
38	-1.50	-.0165
39	17.41	-.1071
40	-3.23	.0299
41	-22.50	-.1282

School	Intercept (u_{0j})	PRIOR-achievement Slope (u_{5j})
42	25.31	-.0550
43	-12.28	.0353
44	7.97	-.0458
45	15.65	-.0437
46	-12.86	-.0169
47	-4.28	.0253
48	9.74	-.0557
49	22.74	-.1807
50	-16.02	.0770
51	11.73	-.0179
52	16.03	-.0014
53	-27.81	.0842
54	20.93	.0397
55	5.36	-.1027
56	-11.42	.0253
57	-3.89	.0745
58	7.04	-.0926
59	59.15	-.1585
60	2.75	-.0827
61	8.57	-.0710
62	-23.89	.0543
63	-7.29	-.0345
64	4.65	-.0243
65	-6.57	.0528
66	22.52	-.0210
67	45.00	-.2033
68	11.75	-.1077
69	-6.12	.0077
70	.00	-.0118
71	-23.49	.0927
72	-24.94	.1211
73	11.81	-.1194
74	-11.53	-.0556
75	14.34	.0580
76	-34.52	.1517
77	-17.77	-.0072
78	-5.68	.0621
79	4.46	.1736
80	-31.36	.0579
81	1.81	.0348
82	-5.02	.0315
83	28.20	-.1341
84	8.18	-.0376

School	Intercept (u_{0j})	PRIOR-achievement Slope (u_{5j})
85	-6.04	-.0442
86	10.61	-.0050
87	-3.27	-.0495
88	24.53	-.1064
89	-16.94	.0618
90	1.05	-.0090
91	-8.50	.1098
92	-6.98	.0431
93	-13.49	.0328
94	-26.74	.1138
95	-12.76	.0441
96	2.74	.1093
97	4.07	-.0692
98	19.48	-.1561
99	-19.26	.0738
100	5.66	.0376
101	-17.44	.1272
102	.38	-.0244
103	13.45	-.0289
104	36.96	-.1166
105	10.39	-.0586
106	-35.16	.1046
107	-2.63	-.0397
108	4.19	.0304
109	-.06	.0215
110	13.83	.0604
111	-2.30	-.0385
112	-2.10	-.0350
113	12.25	-.0497
114	-9.62	.1001
115	-11.02	-.0205
116	23.34	-.0430
117	7.40	-.0746
118	-19.70	.1659
119	1.38	-.0583
120	15.19	-.0545
121	11.66	-.0495
122	-1.42	-.0411
123	1.96	-.0457
124	24.35	-.1576
125	16.76	-.0434
126	-23.59	-.0283
127	22.01	-.0565

School	Intercept (u_{0j})	PRIOR-achievement Slope (u_{5j})
128	-5.60	-.0165
129	-1.96	.0019
130	-8.40	.0480
131	5.67	-.0047
132	-35.28	.0439
133	-6.69	.0377
134	15.08	-.0529
135	15.13	-.0120
136	-4.08	.0559
137	4.34	.0043
138	-10.81	.0290
139	-2.49	-.0065
140	-6.45	.0430
141	-7.45	.0293
142	-1.47	.0543
143	14.96	-.0043
144	-7.99	-.0484
145	8.93	.0210
146	11.60	.0653
147	2.92	.0296
148	-6.91	.0084
149	.67	.0971
150	7.24	-.1208
151	-6.95	.0043
152	8.80	-.0846
153	-5.44	.0551
154	-3.61	-.0166
155	-14.86	.0500
156	2.63	.0219
157	-8.17	-.0155
158	18.63	.0808
159	16.56	-.0300
160	-11.10	.0090
161	1.54	.0150
162	-1.11	.0256
163	-1.26	.0021
164	3.87	-.0797
165	-19.38	.0479
166	-18.92	.0471
167	6.74	-.0313
168	-21.14	-.0242
169	-13.24	.0980
170	6.84	-.0536

School	Intercept (u_{0j})	PRIOR-achievement Slope (u_{5j})
171	2.28	-.0311
172	11.20	.0543
173	-17.91	.1041
174	16.93	-.0530
175	-11.05	.0019
176	12.40	-.0168
177	8.31	.0478
178	1.84	.0464
179	-11.31	.0534
180	-28.64	.0554
181	-12.53	-.0148
182	-38.97	.1472
183	-21.34	.0669
184	-6.19	.0140
185	-4.10	-.0261
186	-3.14	.0342
187	5.57	.0391
188	9.72	-.0615
189	-.63	.0154
190	1.82	-.0081
191	1.54	.0073
192	5.07	-.0106
193	-13.00	-.0212
194	-22.50	.1337
195	-14.73	-.0246
196	1.41	.0220
197	6.06	-.0448
198	-.44	-.0425
199	-1.71	.0772
200	-12.82	.0006
201	12.31	-.0137
202	2.01	-.0575
203	9.63	-.0028
204	-19.21	.0466
205	-13.02	.1045
206	12.16	.0065
207	14.78	-.1473
208	-1.40	.0259
209	-18.89	-.0429
210	2.02	-.0054
211	5.48	.0492
212	47.11	-.1451
213	14.20	.0780

School	Intercept (u_{0j})	PRIOR-achievement Slope (u_{5j})
214	12.62	-.0348
215	.08	.0257
216	10.12	-.0372
217	4.86	.0381
218	-16.26	.0547
219	-11.16	.0773
220	-5.24	-.0127
221	-9.70	.1115
222	.46	-.0734
223	18.16	-.0736
224	19.92	-.1357
225	-16.15	.0891
226	11.68	-.0458
227	-2.19	-.0231
228	-17.99	.0149
229	-4.63	.0680
230	-12.47	-.0226
231	5.95	.0129
232	12.90	-.0546
233	-3.88	.0580
234	8.78	-.0343
235	3.91	-.0319
236	-5.48	.0096
237	15.71	-.0653
238	12.88	-.0816
239	36.60	-.0841
240	-18.00	.0262
241	18.00	-.0300
242	-3.19	.0121
243	-6.24	.0139
244	.95	-.0263
245	9.91	-.0600
246	13.74	-.0323
247	-23.72	.0763
248	21.34	-.0929
249	-5.15	.0087
250	10.18	-.0823
251	-22.04	.0212
252	-3.10	.1552
253	-20.77	.1025
254	21.32	-.0765

APPENDIX G – SCHOOL EFFECTS ESTIMATES AND RANKINGS

School Effects Estimates and Differential Effectiveness Rankings by School

School	Below Average PRIOR	Rank	Average PRIOR	Rank	Above Average PRIOR	Rank
1	42.19	19	46.35	19	50.51	16
2	27.56	39	28.79	41	30.02	44
3	-12.87	172	-13.45	168	-14.04	164
4	-40.66	238	-43.23	239	-45.79	237
5	26.21	43	29.71	39	33.20	39
6	-13.89	177	-14.43	173	-14.96	171
7	8.47	95	9.49	94	10.51	92
8	13.59	74	14.40	76	15.22	78
9	40.71	23	43.16	22	45.60	24
10	31.10	34	33.35	35	35.59	36
11	69.54	3	73.37	3	77.20	4
12	15.33	69	16.62	69	17.90	71
13	-1.50	130	-1.40	130	-1.30	129
14	11.22	88	11.32	90	11.41	90
15	-3.19	136	-4.69	139	-6.20	141
16	-42.62	243	-46.86	243	-51.10	242
17	-13.36	175	-15.52	180	-17.67	182
18	21.79	53	23.34	54	24.88	55
19	-9.86	155	-10.74	159	-11.62	157
20	-23.06	205	-24.56	205	-26.07	206
21	-30.02	223	-32.29	221	-34.56	220
22	-28.76	218	-30.59	216	-32.43	215
23	-6.59	147	-5.87	142	-5.15	138
24	7.15	100	6.80	103	6.45	106
25	-26.92	212	-28.37	211	-29.82	211
26	-13.32	174	-15.84	181	-18.36	183
27	-10.74	160	-12.51	163	-14.29	167
28	-14.71	182	-14.75	174	-14.80	170
29	24.00	51	25.09	52	26.19	54
30	21.10	54	24.77	53	28.43	49
31	61.41	5	66.67	5	71.93	5
32	-4.67	140	-5.83	141	-6.99	144
33	-30.89	225	-33.58	223	-36.28	224
34	13.28	77	15.93	70	18.59	68
35	.56	121	.64	121	.72	120
36	-41.56	241	-44.78	241	-48.00	241
37	41.68	20	46.57	18	51.47	15
38	-10.08	157	-10.67	158	-11.26	154
39	-38.35	235	-42.18	235	-46.01	238
40	12.34	82	13.41	81	14.48	81
41	-89.23	254	-93.82	254	-98.40	254

School	Below Average PRIOR	Rank	Average PRIOR	Rank	Above Average PRIOR	Rank
42	-3.29	137	-5.26	140	-7.23	145
43	6.10	102	7.37	100	8.63	98
44	-15.86	184	-17.49	185	-19.13	184
45	-7.10	148	-8.66	148	-10.23	153
46	-21.65	198	-22.25	195	-22.86	191
47	8.89	93	9.79	93	10.70	91
48	-19.26	190	-21.25	191	-23.25	194
49	-71.29	253	-77.76	253	-84.23	253
50	24.07	50	26.83	48	29.59	45
51	2.41	114	1.77	116	1.13	119
52	15.29	70	15.24	74	15.19	79
53	15.99	65	19.01	62	22.02	60
54	41.56	21	42.98	23	44.40	26
55	-48.07	245	-51.74	245	-55.42	245
56	1.75	116	2.65	114	3.56	113
57	34.88	30	37.54	30	40.21	31
58	-41.16	239	-44.47	240	-47.79	240
59	-23.32	207	-28.99	213	-34.66	221
60	-40.26	236	-43.22	238	-46.18	239
61	-28.40	216	-30.94	219	-33.48	218
62	4.34	108	6.28	106	8.22	101
63	-25.25	211	-26.48	209	-27.72	209
64	-8.01	151	-8.88	149	-9.75	151
65	20.91	55	22.80	56	24.69	56
66	11.58	86	10.83	91	10.08	94
67	-60.81	250	-68.09	252	-75.36	252
68	-44.27	244	-48.12	244	-51.97	244
69	-2.11	132	-1.83	131	-1.55	130
70	-6.14	145	-6.56	145	-6.98	143
71	24.72	48	28.04	44	31.35	40
72	38.07	26	42.40	25	46.73	23
73	-50.31	246	-54.58	246	-58.85	246
74	-40.45	237	-42.44	236	-44.43	236
75	44.52	16	46.60	17	48.67	21
76	44.40	17	49.82	14	55.25	14
77	-21.52	197	-21.78	193	-22.04	189
78	26.62	41	28.84	40	31.06	41
79	94.79	1	101.00	1	107.21	1
80	-1.24	129	.83	120	2.90	114
81	19.92	57	21.17	58	22.41	58
82	11.37	87	12.50	87	13.62	85
83	-41.57	242	-46.37	242	-51.17	243
84	-11.39	162	-12.74	165	-14.08	166

School	Below Average PRIOR	Rank	Average PRIOR	Rank	Above Average PRIOR	Rank
85	-29.02	221	-30.60	217	-32.18	214
86	7.98	97	7.80	99	7.62	104
87	-29.01	220	-30.78	218	-32.55	216
88	-30.85	224	-34.66	228	-38.47	228
89	15.24	71	17.45	67	19.66	66
90	-3.62	138	-3.94	136	-4.26	136
91	48.61	11	52.54	11	56.47	12
92	15.43	68	16.97	68	18.51	69
93	3.57	111	4.74	109	5.92	108
94	32.48	33	36.55	31	40.63	30
95	10.16	91	11.74	89	13.31	87
96	59.60	7	63.51	7	67.42	6
97	-31.95	227	-34.43	226	-36.90	227
98	-61.75	251	-67.34	251	-72.93	251
99	19.13	60	21.77	57	24.41	57
100	25.24	46	26.58	49	27.93	50
101	48.72	10	53.27	10	57.82	10
102	-12.33	166	-13.20	167	-14.07	165
103	-1.56	131	-2.60	133	-3.63	135
104	-23.72	208	-27.89	210	-32.07	213
105	-20.09	191	-22.18	194	-24.28	196
106	19.26	59	23.00	55	26.74	53
107	-23.28	206	-24.70	206	-26.12	207
108	19.98	56	21.07	59	22.16	59
109	11.12	90	11.89	88	12.66	88
110	45.25	15	47.41	16	49.57	19
111	-22.36	202	-23.74	201	-25.12	199
112	-20.33	194	-21.59	192	-22.84	190
113	-13.63	176	-15.40	179	-17.18	179
114	42.46	18	46.05	20	49.63	18
115	-21.67	199	-22.41	196	-23.14	193
116	.99	119	-.55	126	-2.09	132
117	-31.43	226	-34.10	225	-36.77	226
118	66.65	4	72.59	4	78.53	3
119	-28.97	219	-31.06	220	-33.14	217
120	-13.19	173	-15.14	178	-17.09	178
121	-14.11	179	-15.88	182	-17.65	181
122	-22.79	204	-24.26	203	-25.73	204
123	-21.81	200	-23.45	200	-25.08	198
124	-57.66	249	-63.30	249	-68.93	249
125	-5.81	143	-7.36	147	-8.91	147
126	-38.31	234	-39.33	232	-40.34	231
127	-7.36	150	-9.38	150	-11.40	155

School	Below Average PRIOR	Rank	Average PRIOR	Rank	Above Average PRIOR	Rank
128	-14.17	180	-14.76	175	-15.35	173
129	-.95	128	-.88	128	-.81	125
130	16.59	62	18.31	65	20.02	65
131	3.21	112	3.04	113	2.87	115
132	-12.45	168	-10.88	160	-9.31	149
133	12.93	78	14.28	78	15.63	75
134	-12.45	167	-14.34	172	-16.24	177
135	8.87	94	8.44	96	8.01	102
136	25.01	47	27.01	47	29.01	47
137	6.57	101	6.73	104	6.88	105
138	4.27	109	5.31	108	6.35	107
139	-5.88	144	-6.11	144	-6.34	142
140	15.93	66	17.47	66	19.01	67
141	7.79	98	8.84	95	9.88	95
142	26.76	40	28.70	42	30.64	42
143	12.73	80	12.58	86	12.42	89
144	-33.18	229	-34.91	229	-36.64	225
145	19.87	58	20.63	60	21.38	62
146	45.60	14	47.94	15	50.28	17
147	18.30	61	19.36	61	20.42	63
148	-2.56	134	-2.26	132	-1.96	131
149	51.20	9	54.68	9	58.15	9
150	-55.62	248	-59.94	248	-64.26	248
151	-4.69	141	-4.54	138	-4.38	137
152	-35.23	231	-38.26	231	-41.29	232
153	23.21	52	25.19	51	27.16	52
154	-12.24	165	-12.84	166	-13.43	163
155	11.17	89	12.95	85	14.74	80
156	14.03	73	14.81	75	15.60	76
157	-16.25	185	-16.81	183	-17.36	180
158	60.65	6	63.54	6	66.43	7
159	.94	120	-.13	124	-1.21	128
160	-6.40	146	-6.08	143	-5.75	140
161	9.33	92	9.86	92	10.40	93
162	12.20	84	13.11	83	14.03	83
163	-.16	123	-.09	122	-.01	124
164	-37.60	232	-40.45	234	-43.30	234
165	5.53	104	7.24	102	8.95	97
166	5.59	103	7.27	101	8.96	96
167	-9.52	154	-10.64	157	-11.76	158
168	-33.71	230	-34.58	227	-35.44	223
169	37.74	27	41.25	27	44.76	25
170	-21.03	195	-22.94	198	-24.86	197

School	Below Average PRIOR	Rank	Average PRIOR	Rank	Above Average PRIOR	Rank
171	-13.90	178	-15.02	176	-16.13	176
172	39.46	24	41.41	26	43.35	29
173	36.24	29	39.96	29	43.69	28
174	-10.63	159	-12.53	164	-14.42	168
175	-10.05	156	-9.98	152	-9.91	152
176	3.68	110	3.08	112	2.49	116
177	33.21	31	34.92	33	36.63	33
178	25.97	44	27.63	45	29.28	46
179	16.48	63	18.39	64	20.30	64
180	.17	122	2.15	115	4.13	111
181	-20.22	193	-20.75	189	-21.28	188
182	37.65	28	42.92	24	48.19	22
183	13.48	75	15.88	72	18.27	70
184	1.10	117	1.60	117	2.10	117
185	-17.70	187	-18.64	186	-19.57	186
186	14.66	72	15.88	71	17.11	72
187	25.92	45	27.32	46	28.72	48
188	-22.28	201	-24.48	204	-26.68	208
189	7.36	99	7.91	98	8.46	99
190	-2.40	133	-2.69	134	-2.98	134
191	5.31	105	5.57	107	5.83	109
192	-.44	124	-.82	127	-1.20	127
193	-24.04	209	-24.80	207	-25.56	201
194	47.06	13	51.84	13	56.62	11
195	-27.53	214	-28.40	212	-29.28	210
196	12.87	79	13.66	80	14.45	82
197	-17.27	186	-18.88	187	-20.48	187
198	-22.57	203	-24.09	202	-25.62	203
199	38.47	25	41.24	28	44.00	27
200	-12.53	169	-12.51	162	-12.49	160
201	5.20	106	4.72	110	4.23	110
202	-27.93	215	-29.99	214	-32.05	212
203	8.17	96	8.07	97	7.97	103
204	5.03	107	6.69	105	8.36	100
205	41.38	22	45.12	21	48.86	20
206	15.52	67	15.75	73	15.99	74
207	-61.86	252	-67.13	250	-72.40	250
208	12.06	85	12.98	84	13.91	84
209	-41.20	240	-42.73	237	-44.27	235
210	-.76	127	-.95	129	-1.14	126
211	31.07	35	32.83	37	34.59	38
212	-28.42	217	-33.61	224	-38.80	230
213	54.76	8	57.55	8	60.34	8

School	Below Average PRIOR	Rank	Average PRIOR	Rank	Above Average PRIOR	Rank
214	-5.49	142	-6.74	146	-7.98	146
215	13.43	76	14.35	77	15.27	77
216	-9.24	153	-10.57	156	-11.90	159
217	24.67	49	26.04	50	27.40	51
218	12.21	83	14.16	79	16.12	73
219	29.08	38	31.85	38	34.61	37
220	-11.86	163	-12.32	161	-12.77	161
221	48.33	12	52.32	12	56.31	13
222	-37.73	233	-40.36	233	-42.98	233
223	-20.14	192	-22.77	197	-25.41	200
224	-50.68	247	-55.54	247	-60.39	247
225	30.19	37	33.38	34	36.57	34
226	-12.18	164	-13.82	170	-15.46	174
227	-14.21	181	-15.04	177	-15.86	175
228	-10.23	158	-9.69	151	-9.16	148
229	30.77	36	33.21	36	35.64	35
230	-24.25	210	-25.06	208	-25.87	205
231	12.67	81	13.14	82	13.60	86
232	-15.51	183	-17.47	184	-19.42	185
233	26.31	42	28.39	43	30.46	43
234	-9.09	152	-10.32	155	-11.55	156
235	-12.70	170	-13.85	171	-14.99	172
236	-.47	125	-.13	123	.22	122
237	-18.24	188	-20.58	188	-22.91	192
238	-29.58	222	-32.50	222	-35.42	222
239	-7.17	149	-10.18	153	-13.19	162
240	-4.39	139	-3.45	135	-2.52	133
241	2.41	115	1.34	119	.26	121
242	3.13	113	3.56	111	4.00	112
243	1.01	118	1.51	118	2.01	118
244	-12.74	171	-13.68	169	-14.62	169
245	-21.29	196	-23.44	199	-25.58	202
246	-3.07	135	-4.23	137	-5.38	139
247	16.00	64	18.73	63	21.46	61
248	-27.01	213	-30.33	215	-33.65	219
249	-.63	126	-.32	125	-.01	123
250	-32.64	228	-35.58	230	-38.53	229
251	-11.01	161	-10.26	154	-9.50	150
252	77.67	2	83.22	2	88.78	2
253	32.55	32	36.21	32	39.88	32
254	-18.50	189	-21.24	190	-23.98	195