

Development of an Analytical Process to Measure Teacher Effectiveness Based on Student Growth to Augment an Educator Evaluation System

by

Lamar D. Adams

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 4, 2012

Keywords: Student Growth, Teacher Evaluation System, Linear Mixed Model, Quantile Regression, Principal Component Analysis, Cluster Analysis

Copyright 2012 by Lamar D. Adams

Approved by

Dr. Jeffrey S. Smith, Chair, Joe W. Forehand Jr. Professor, Industrial and Systems Engineering
Dr. Saeed Maghsoodloo, Professor Emeritus of Industrial and Systems Engineering
Dr. David Shannon, Humana-Germany-Sherman Distinguished Professor, Educational Foundations, Leadership and Technology
Dr. Joni Lakin, Assistant Professor of Educational Foundations, Leadership and Technology

Abstract

Teacher quality is one of the most important school related variables associated with student achievement. Therefore, raising the quality of the U.S. public education teaching force is essential to ensure that every child has the opportunity to achieve academic success. In order to accomplish this, significant analytical inspection of teachers is needed to assist with the determination of whether teachers contribute appropriately to students attaining adequate yearly growth. The primary objective of this research was to fill the need of augmenting Alabama's formative educator evaluation system, EDUCATEAlabama, with a precise and stable teacher effectiveness index based on student growth. The methodology of computing such an index consisted of three phases. Phase I entailed calculating four teacher effectiveness metrics. Subject-specific and overall teacher index values were calculated in Phase II utilizing the Phase I metrics and principal component analysis. The principal components served as the inputs to Phase III, Cluster Analysis, with Ward's clustering method employed as a general prescription to illuminate teachers with similar characteristics (principal components) in the data. A medium-sized, suburban district in Alabama and a dataset from the National Center for Education Statistics consisting of 17 urban districts from across the United States provided the requisite student and teacher data to fully implement the process, which concluded with successfully placing teachers into effectiveness categories by grade, subject(s), and year.

Acknowledgments

Marlin Lavon Adams passed away on October 13, 2009, at the age of 76. I had completed a mere semester of doctoral studies at Auburn University on that day. My father wanted me to earn a Ph.D. before I even considered pursuing one. He had vision. I want to thank him and tell him that I miss him.

Children of military service members seldom have a choice in where they live. My children are no exception. David has lived through eight assignments, seven moves, and eight schools spanning six states. Leslie has lived through seven assignments, six moves, and six schools spanning five states. They are “military” children in the true sense of the word by being wonderfully supportive, adaptive, and selfless. I want to thank them for being who they are and tell them that I am proud of them.

My bride has proudly stood next to me through every promotion, graduation, and award ceremony for the past 17 years. More importantly, Jennifer has stood next to me through every disappointment, failure, and setback over those same 17 years. I want to thank her for the love and support she has given me while I pursued this degree.

Dr. Jeffrey Smith took a chance when he agreed to support this educational research. I thank him for serving as my committee chairman and allowing me to tackle an area of research outside of the department. I would also like to thank Dr. David Shannon for serving as a committee member. His efforts to acquire data were significant, and I would not have obtained the requisite data from an Alabama district without his help. Lastly, I would like to thank my

remaining committee members, Dr. Saeed Maghsoodloo and Dr. Joni Lakin. Their insight and assistance have been immeasurable.

Bonnie “G.B.” Adams moved to Auburn in 2009 to support our family while I completed my doctoral studies and Jennifer completed her degree at Auburn. G.B. packed up her house in Lancaster, California, and moved east to establish a home nearby. Never one to shy away from work, she completed any and every “duty” to help our family to be successful. I want to thank my mother for her love and support during our time in Auburn.

Table of Contents

Abstract	ii
Acknowledgments.....	iii
List of Tables	ix
List of Figures	xi
List of Abbreviations	xiii
Chapter 1 : Introduction	1
1.1 Teacher Quality.....	1
1.2 No Child Left Behind.....	1
1.3 Alabama’s Race to the Top.....	3
1.4 Educator Evaluation System	5
1.5 Research Objectives.....	7
1.6 Organization of Research.....	8
Chapter 2 : Literature Review	11
2.1 Introduction.....	11
2.2 Linear Mixed Models (LMMs).....	12
2.2.1 Linear Mixed Model General Specification	14
2.2.2 Linear Mixed Model Hierarchical Specification	17
2.2.3 Linear Mixed Model Implementation in Texas	19
2.3 Quantile Regression (QR).....	20
2.3.1 Computing QR Coefficients	22
2.3.2 Quantile Regression Extended to Cubic Splines	23

2.3.3 Quantile Regression and B-Splines	25
2.3.4 Quantile Regression Implementation in Colorado.....	26
2.4 Discussion.....	29
2.4.1 Measures of Effective Teaching (MET) Project.....	31
2.4.2 A Risk-Mitigated Approach.....	33
2.5 Principal Component Analysis	35
2.6 Cluster Analysis	37
2.7 Summary.....	40
Chapter 3 : Risk-Mitigated Teacher Effectiveness Index	42
3.1 Introduction.....	42
3.2 Phase I: Teacher Effectiveness Metrics	43
3.2.1 Linear Mixed Model Teacher Effect.....	45
3.2.2 Linear Mixed Model Value-Added Measure.....	46
3.2.3 Median Student Growth Percentile.....	48
3.2.4 Quantile Regression Value-Added Measure.....	49
3.3 Phase II: Principal Component Analysis (PCA).....	50
3.4 Phase III: Cluster Analysis.....	57
3.4.1 Comparison of Clustering Results with Phase I Metrics	62
3.5 Summary.....	63
Chapter 4 : Data Analysis	66
4.1 Introduction.....	66
4.2 Alabama Data Analysis.....	67
4.2.1 Alabama Reading and Mathematics Test	69
4.2.2 Testing Histories	70
4.2.3 Student, Teacher, and School Level Predictors for Alabama Data.....	72

4.3 National Center for Education Statistics Data Analysis	74
4.3.1 NCES Student Achievement Data Recorded as Z-Scores	76
4.3.2 Student, Teacher, and School Level Predictors for NCES Data.....	77
4.4 Checking Model Assumptions for the Final Linear Mixed Models	80
4.5 Diagnostics for the Final Quantile Regression Models	82
4.6 Principal Component Analysis (PCA)	84
4.7 Cluster Analysis	87
4.7.1 Comparison of Clustering Results with Phase I Metrics	90
4.8 Subject-Specific and Overall RMTEI Values.....	92
4.9 Precision of the Risk-Mitigated Teacher Effectiveness Index.....	96
4.10 Stability of the Risk-Mitigated Teacher Effectiveness Index.....	99
4.10.1 Comparison of Stability Results with Phase I Metrics	105
4.11 Summary.....	108
Chapter 5 : Assessment of Teacher Evaluation in Alabama.....	110
5.1 Introduction.....	110
5.2 Professional Education Personnel Evaluation (PEPE)	111
5.3 EDUCATEAlabama	111
5.4 NCES Teacher Effectiveness Scoring Analysis	113
5.5 Summary	116
Chapter 6 : Research Summary.....	118
6.1 Conclusion	118
6.2 Limitations of the Risk Mitigated Teacher Effectiveness Index	121
6.2.1 Limitations in Reporting	121
6.2.2 RMTEI Values and Small Populations of Teachers	123
6.3 Future Study.....	124

References.....	127
Appendix 1: Establishing Longitudinal Student Achievement Data Linked with Teacher Information	134
Appendix 2: SAS Code.....	142
Appendix 3: Alabama District Results	165
Appendix 4: NCES District Results.....	168

List of Tables

Table 2.1: Linear and Quantile Regression Comparison	21
Table 3.1: Excerpt of Example Dataset for Phase I Metrics	44
Table 3.2: Partial Output of Solution for Random Effects	46
Table 3.3: Partial Output of LMM Value-Added Measure.....	47
Table 3.4: Partial Output of Median SGP Metric	48
Table 3.5: Partial Output of Overall QR Teacher Value-Added Measure.....	49
Table 3.6: Output of Phase I Metrics	51
Table 3.7: Correlation Matrix of Phase I Metrics	52
Table 3.8: Eigenvectors of the Correlation Matrix	53
Table 3.9: Output of Principal Component Scores	54
Table 3.10: Eigenvalues of the Correlation Matrix	55
Table 3.11: Correlation of Phase I Metrics with Principal Component 1.....	56
Table 3.12: Sample of Teachers with Extraordinary Gains in Student Growth	61
Table 3.13: Sample of Teachers with Poor Gains in Student Growth.....	61
Table 4.1: Excerpt from 6 th grade, 2009 Dataset	69
Table 4.2: Components of Alabama Reading and Mathematics Test.....	69
Table 4.3: Summary of Model Predictors for Alabama Data	74
Table 4.4: NCES Testing Histories.....	75
Table 4.5: Summary of Model Predictors for NCES Data	80
Table 4.6: Proportion of the Total Variance for First Principal Component in the Alabama District.....	85

Table 4.7: Proportion of the Total Variance for First Principal Component in a NCES District	86
Table 4.8: Results from PCA for 4 th Grade, 2011, in the Alabama District	87
Table 4.9: Excerpt of Final Results Sorted by the 2009 Mathematics RMTEI value	93
Table 4.10: Excerpt of Treatment Status Analysis	96
Table 4.11: Correlation of Yearly RMTEI Math Values for the Alabama District.....	101
Table 4.12: Correlation of Yearly Reading, Overall RMTEI Values for the Alabama District	102
Table 4.13: Correlation of Yearly Mathematics RMTEI Values for NCES Data	103
Table 4.14: Correlation of Yearly Reading and Overall RMTEI Values for NCES Data.....	104
Table 5.1: Correlation of RMTEI Reading Values and Observational Evaluations.....	115
Table 6.1: Alabama Student Assessment Program Overview	122
Table 6.2: Excerpt of Schedule Dataset.....	136
Table 6.3: Excerpt of Course Counts Dataset.....	137
Table 6.4: Excerpt of Teacher Information File	138
Table 6.5: Excerpt from 6 th grade, 2009 Dataset	140

List of Figures

Figure 2.1: Linear Parameterization of Quantile Regression.....	21
Figure 2.2: Point/Line Duality (Edgeworth, 1888).....	22
Figure 2.3: Cubic B-Spline Parameterization of Student Growth Percentiles.....	27
Figure 3.1: Scree Plot.....	55
Figure 3.2: Subject-specific RMTEI Value along Dominant Principal Component Axis.....	57
Figure 3.3: Ward’s Clustering Method Statistics for Determining Number of Clusters	59
Figure 3.4: Ward’s Clustering of Teachers by Effectiveness	60
Figure 3.5: Comparison of Clustering Results with Phase I Metrics.....	62
Figure 4.1: Distribution Analysis of Mathematics LMM Teacher Effects	80
Figure 4.2: Distribution Analysis of Studentized Residuals from Mathematics LMM.....	81
Figure 4.3: Paired Comparison of MathGain and Conditional Predicted Value	82
Figure 4.4: Variance Analysis of Standardized Residuals from 0.5 QR Mathematics Model	83
Figure 4.5: Agreement Plot of MathGain versus 0.5 QR Prediction of MathGain	84
Figure 4.6: Distribution Analysis of Standardized Residuals from 0.5 QR Mathematics Model	84
Figure 4.7: Linear Relationship of the Number of Clusters and Teachers for the Alabama District.....	88
Figure 4.8: 4 th Grade, 2011, Effectiveness Categories for the Alabama District	89
Figure 4.9: Linear Relationship of Number of Clusters and Teachers for a NCES District	90
Figure 4.10: Mathematics Clustering Results for 4 th Grade, 2011, in the Alabama District.....	91
Figure 4.11: Mathematics Clustering Results for 3 rd Grade, 2006, in a NCES District	91
Figure 4.12: Final 4 th Grade Mathematics RMTEI values.....	94

Figure 4.13: Control Chart Analysis for Alabama 4 th Grade Mathematics Teachers	98
Figure 4.14: Control Chart Analysis for 3 rd Grade Mathematics Teachers in a NCES District ..	98
Figure 4.15: Excerpt of Final Stability Analysis for the Alabama District	100
Figure 4.16: Excerpt of Final Stability Analysis for NCES Data	103
Figure 4.17: Mathematics Correlation Results in the Alabama District for 2009	105
Figure 4.18: Mathematics Correlation Results in the AL District between 2010 and 2011	106
Figure 4.19: Mathematics Correlation Results for 2006 in a NCES District.....	107
Figure 4.20: Mathematics Correlation Results between 2007 and 2008 for a NCES District...	107
Figure 5.1: Scatter Plots of RMTEI Reading Values versus Observational Evaluations	115
Figure 6.1: Establishing Longitudinal Student Achievement Data Linked with Teacher Information	141

List of Abbreviations

AL	Alabama
ALSDE	Alabama State Department of Education
ARMT	Alabama Reading and Mathematics Test
ANOVA	Analysis of Variance
AQTS	Alabama Quality Teaching Standards
AYP	Adequate Yearly Progress
CCC	Cubic Clustering Criterion
COMPETES	Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science
CSAP	Colorado Student Assessment Program
CSV	Comma Separated Values
EBLUP	Empirical Best Linear Unbiased Predictor
ERIC	Education Reform and Innovation Council
ESEA	Elementary and Secondary Education Act
ETS	Educational Testing Service
EVAAS	Education Value Added Assessment System
HLM	Hierarchical Linear Model
IRT	Item Response Theory
LEA	Local Education Agency
LMM	Linear Mixed Model

MET	Measures of Effective Teaching
NCES	National Center for Education Statistics
NCLB	No Child Left Behind Act
NTC	New Teacher Center
PCA	Principal Component Analysis
PEPE	Professional Education Personnel Evaluation
RMTEI	Risk Mitigated Teacher Effectiveness Index
SGP	Student Growth Percentile
SQL	Structured Query Language
TPM	Texas Projection Measure
TVAAS	Tennessee Value-Added Assessment System
QR	Quantile Regression
VAM	Value Added Model

Chapter 1 : Introduction

1.1 Teacher Quality

Teacher quality is one of the most important school related variables associated with student achievement. Many education accountability systems built around student test results ignore the initial conditions of the students and merely measure a status at a specific point in time. Teacher evaluations based solely on a snapshot of the number of students that have attained a particular testing proficiency level without consideration of the initial conditions of those students are flawed. Certainly, status model (snapshot) evaluations provide useful information to administrators, but they do not account for student growth and are “confounded with other non-school factors” (Lissitz & Doran, 2009, p. 39). A snapshot of student scores could portray a teacher to be ineffective despite tremendous student growth occurring within the classroom, or effective despite little student growth occurring within the classroom. A favorable alternative to status evaluations is to determine which teachers contribute appropriately to students attaining adequate yearly growth. Therefore, a statistically supportable measure of teacher effectiveness based on student growth is desired.

1.2 No Child Left Behind

The No Child Left Behind Act (NCLB) of 2001, a reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965, is a performance-based accountability system built around student test results. It requires Adequate Yearly Progress (AYP) determinations for schools to be based on a snapshot of the number of students that have attained proficiency.

Therefore, several states gained approval from the U.S. Department of Education to use “growth models” to augment AYP calculations. These states could use longitudinal testing data to extrapolate the expected growth for individual students over time thereby accounting for “differences in the initial conditions of the students” (Pearson & Stecher, 2004, p. 99). This expected growth would then become benchmarks for students to meet in order to be counted for making AYP. “Fifteen states now have approved growth models: North Carolina, Tennessee, Delaware, Arkansas, Florida, Iowa, Ohio, Alaska, Arizona, Michigan, Missouri, Colorado, Minnesota, Pennsylvania and Texas” (“Secretary Spellings,” 2009, para. 5).

The subsequent reauthorization of the ESEA titled “A Blueprint for Reform”, released by the U.S. Department of Education in March 2010, now requires all state accountability systems to “recognize progress and growth” (“A Blueprint for Reform,” 2010, p. 9). It goes further by requiring states to “identify effective and highly effective teachers and principals on the basis of student growth” (“A Blueprint for Reform,” 2010, p. 4). Mathematical techniques presently employed in growth models that augment AYP calculations can be applied to this new state requirement of measuring teacher effectiveness.

A specific application of “growth models” is a Value Added Model (VAM). VAMs can measure the influence of educational entities on student growth using student longitudinal testing data. This measure of influence is the additional value that a teacher brings to the classroom above that of his/her peers. Several states currently calculate a quantitative indicator of value that identifies teachers who employ pedagogical strategies or exhibit certain behaviors that positively impact student learning. Alabama (AL) is absent from this group of states that have such a measure, yet it desires to improve its existing educator evaluation system and meet the requirement of A Blueprint for Reform (Bice, 2010).

1.3 Alabama's Race to the Top

The U.S. Department of Education created the Race to the Top grant program to allow states to compete for federal funding that supports needed education reform. The structure of the program allows \$4 billion to be funded for state reforms in the following four areas:

1. "Adopting standards and assessments that prepare students to succeed in college and the workplace;
2. Building data systems that measure student growth and success, and inform teachers and principals how to improve instruction;
3. Recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most; and
4. Turning around their lowest-performing schools" ("Delaware and Tennessee," 2011, para. 6).

Alabama submitted a Phase I application to the U.S. Department of Education in January 2010 along with 39 other states and the District of Columbia. Tennessee and Delaware won grants for Phase I and were awarded \$500 million and \$100 million respectively ("Delaware and Tennessee," 2011, para. 3). "Delaware and Tennessee [had] aggressive plans to improve teacher and principal evaluation, use data to inform instructional decisions, and turn around their lowest-performing schools" ("Delaware and Tennessee," 2011, para. 8). Phase II of the program commenced after the announcement of the Phase I winners in March 2010 and had \$3.4 billion still available for reform grants.

Based on recommendations from the reviewers of Alabama's Phase I application, the Alabama State Department of Education (ALSDE) attempted to bolster its Phase II application by including reforms in teacher and principal evaluation in which teacher and principal effectiveness ratings are tied to student growth. In order to include such reforms in the Phase II application, stakeholder support had to come from the ALSDE Board. In a contentious 5-4 vote

on May 27, 2010, with Governor Bob Riley providing the swing vote, the ALSDE Board passed the Educator Effectiveness Resolution tying teacher and principal effectiveness to student performance (Morton, 2010, p. 3). This special session of the ALSDE board on the eve of the Phase II application deadline of June 1, 2010, was critical to being able to submit the comprehensive document of reforms that the ALSDE believed it needed.

Alabama submitted its Phase II Race to the Top grant application, “Advancing Education as the 21st Century Civil Right”, to the U.S. Department of Education on June 1, 2010. The Phase II application contained language throughout that supports research in measuring teacher and principal effectiveness based on student growth. A sample of the reforms follows:

1. “Create data systems...that are readily available to colleges and universities for research” (“Alabama’s Race,” 2010, p. 4)
2. “Alabama will redesign the current accountability system ...as measured by student growth” (“Alabama’s Race,” 2010, p. 9).
3. “Alabama will apply a growth model to existing data...to develop predictive trajectories for its students” (“Alabama’s Race,” 2010, p. 9).
4. “The Educator Effectiveness Resolution allow[s] the use of multiple and objective measures of student growth outcomes as the predominant factor for determining teacher and principal effectiveness” (“Alabama’s Race,” 2010, p. 86).

Alabama came in last place out of the 36 states that submitted Phase II applications. Dr. Joseph Morton, State Superintendent of Education, sent an open letter to Arne Duncan, U.S. Secretary of Education, dated September 1, 2010, following the announcement of the Phase II winners. In the letter Dr. Morton critically addressed the grading of the applications with his assertion that reviewers placed unnecessarily high importance on the ability of states to have charter schools, teacher union support of measuring teacher effectiveness based on student achievement, and adoption of the Common Core Standards. At the time of the application, Alabama had none of those elements. Dr. Morton stated that despite knowing Alabama’s

application would not be competitive in Phase II, it would serve as a foundation for needed reforms. The letter demonstrates the conviction of Dr. Morton to provide a document of reforms for the State in spite an anticipated poor outcome and the difficulty to even obtain the authority to submit it (Morton, 2010).

1.4 Educator Evaluation System

Alabama desires an objective effectiveness index in its educator evaluation system to augment its presently used formative assessment, EDUCATEAlabama. The combination of the two components would produce a yearly effectiveness score for each teacher in order to place teachers into “at least” four categories:

1. Extraordinary gains in student growth.
2. Meets student growth.
3. Did not meet student growth but produce evidence that progress is being made.
4. Consistently failed to produce student growth for multiple or consecutive years (“Alabama’s Race,” 2010, p. 90).

The most important aspect of obtaining such categories is being able to implement policies and practices aimed at preparing every student “to graduate from high school ready for college and a career” (“A Blueprint for Reform,” 2010, p. 3). For example, effectiveness categories could be used to ensure the equitable distribution of effective teachers, target incentives at effective teachers to teach in high-need schools, and target professional development for those teachers not able to demonstrate effectiveness. Following the development of the new evaluation system, “alignment” of the two components is desired through refinement of the formative component (“Alabama’s Race,” 2010, p. 90).

Alabama's Race to the Top application contains an aggressive but systematic plan to implement an educator evaluation system that has teacher effectiveness ratings tied to student growth. The application as well as the Educator Effectiveness Resolution passed by the ALSDE Board on May 27, 2011, stipulate that the Education Reform and Innovation Council (ERIC) will be formed to identify an approach to measure teacher and principal effectiveness based on student growth. According to the application, the council was to convene in the summer of 2010 and develop approaches to measuring student growth and teacher and principal effectiveness by December 2010 ("Alabama's Race," 2010, p. 86,89).

The Educator Effectiveness Resolution stated that the ALSDE Board would receive the recommendations in early 2011 to review, discuss, and take possible action in order for "full implementation to occur with the first day of school in 2011" ("Educator Effectiveness Resolution," 2010). The Race to the Top application was less optimistic. It required the development and implementation of an evaluation system that includes an objective measure based on student growth by the 2012-2013 school year ("Alabama's Race," 2010, p. 86).

Regardless of the implementation date for the new evaluation system, the ERIC never convened to provide the required recommendations. The ALSDE acknowledged that it "did not know where to start" (Bice, 2010). More study was necessary before the council could come together to determine how to include teacher effectiveness, as measured by student growth, in an educator evaluation system (Bice, 2010).

In addition to determining how to best measure teacher effectiveness with student growth, a practical matter exists for Alabama to link student achievement data with teachers. According to the Alabama's Race to the Top application, Alabama currently meets 10 of the 12 data elements of the America Creating Opportunities to Meaningfully Promote Excellence in

Technology, Education, and Science (COMPETES) Act (“Alabama’s Race,” 2010, pp. 59–61). On January 4, 2011 President Obama signed into law the America COMPETES reauthorization Act which further authorizes the investment in “research and development, education, innovation, and competitiveness” (Holdren, 2011, para. 2).

Alabama stated clearly that it met data element No. 8 of America COMPETES which requires a “teacher identifier with the ability to link teachers to students” (“Alabama’s Race,” 2010, p. 60). Certainly a difference exists between having the “ability” to link student data to teachers and analyzing student data incorporating teacher identifiers. The State does not provide the linked testing data to the districts nor does it provide analysis as a result of the linkage (Crouse, 2011; DiChiara, 2011). Presently, the ALSDE considers it difficult to link student data with teachers due to the requirement to merge a teacher scheduling database with a student achievement database (Larson, 2010). Districts only receive yearly student achievement data and are not provided longitudinal student data to allow longitudinal analysis to occur (Crouse, 2011). Therefore, the lack of teacher-linked, longitudinal data endures as an administrative issue, not a technical one.

1.5 Research Objectives

The primary objective of this research is to fill the need of augmenting Alabama’s formative educator evaluation system, EDUCATEAlabama, with a precise and stable teacher effectiveness index based on student growth. This index will be determined with an analytical process involving principal component and cluster analysis that uses multiple and objective measures of teacher effectiveness to minimize the risk inherent in isolating the value that a teacher brings to the classroom above that of his/her peers. In order to accomplish the primary objective, this research intends to accomplish the following sub-objectives:

- 1) Develop techniques with Alabama's database infrastructure to streamline the process of establishing longitudinal data from existing yearly data in order to make this information readily accessible to school districts.
- 2) Develop techniques with Alabama's database infrastructure to streamline the process of linking student achievement data with teachers in order to make this information readily accessible to school districts.
- 3) Confirm or modify Alabama's effectiveness rating categories based on being able to detect statistically different groups of teachers by effectiveness.
- 4) Report all teachers in accordance with the rating categories for an Alabama school district using the newly developed objective effectiveness index.
- 5) Provide an assessment of Alabama's educator evaluation system, compare and contrast the results of an objective effectiveness index with observational teacher evaluation data, and propose an observational, summative assessment for Alabama that is a predictor of student achievement gains and correlated with an objective effectiveness index.

1.6 Organization of Research

The structure needed to accomplish the research objectives of Chapter 1 consists of five additional chapters. Chapter 2 contains a literature review to provide the technical context for the development of a teacher effectiveness index. Through the use of an example, Chapter 3 describes the methodology of computing the Risk-Mitigated Teacher Effectiveness Index (RMTEI) within a three phase process. Phase I entails calculating four teacher effectiveness metrics:

- 1) Linear Mixed Model Teacher Effect - statistical prediction of the relative value of a particular teacher by measuring the teacher deviation from the district mean; greater is better.
- 2) Overall Linear Mixed Model Value-Added Measure - a teacher's average of the difference between students' actual achievement and predicted achievement had they been taught by the average teacher in the district; greater is better.
- 3) Median Student Growth Percentile - serves as an indicator of student growth associated with each teacher by calculating the median of a teacher's student growth

percentiles. A student's growth percentile is obtained by determining what percentage of other students had less growth in testing achievement; greater is better.

4) Overall Quantile Regression Value-Added Measure - a teacher's average of the difference between students' actual achievement and predicted achievement of a typical student within the district; greater is better.

Subject-specific and overall teacher index values are calculated in Phase II with an analytical process involving principal component analysis. Since teachers' Phase I metrics for a given subject, grade, and year have varying units of measurement, the principal components are obtained from the standardized version of those metrics by calculating the eigenvectors of the metrics' correlation matrix. The variance of the principal component scores from the dominant principal component is the largest eigenvalue associated with the corresponding eigenvector. This eigenvalue absorbs the preponderance of the variation of the system, and the principal component scores from the dominant principal component become teachers' subject-specific RMTEI values. If teachers instruct mathematics and reading, then an overall teacher effectiveness index is obtained by taking the mean of their two subject indexes. Otherwise, the single, subject index is the teachers' overall values.

The principal components serve as the inputs to Phase III, Cluster Analysis. Since the RMTEI process produces a dominant principal component that continually leads to elliptically shaped clusters in two dimensions, Ward's clustering method, which expects elliptically shaped clusters, is employed as a general prescription of the RMTEI process. Ward's clustering method illuminates teachers with similar characteristics (principal components) in the data, which provides better assignments to teacher effectiveness categories compared to clustering with a single Phase I metric.

The methodology is placed into practice in Chapter 4 by examining five years of student and teacher data from an Alabama school district and four years of student and teacher data from

418 elementary schools in 17 urban districts from across the United States. Further analysis is undertaken in Chapter 4 to confirm the desired outcome of a precise and stable teacher effectiveness index along with a discussion of the results. Chapter 5 provides an assessment of Alabama's educator evaluation system, and then propose an observational, summative assessment for Alabama that is a predictor of student achievement gains and correlated with the Risk-Mitigated Teacher Effectiveness Index. Lastly, Chapter 6 concludes the dissertation and provides recommendations for future study.

Chapter 2 : Literature Review

2.1 Introduction

Most prominent methods of analytically measuring teacher effectiveness are rooted in the application of Linear Mixed Models. Another analytical method that is currently applied to determining school effectiveness in the state of Colorado, which can be extended to teacher effectiveness, employs Quantile Regression (Betebenner, 2007). Methods rooted in the application of Linear Mixed Models based on longitudinal student testing data either derive individual teacher effects in order to make teacher comparisons or make student test score projections for the successive year. For the latter case students' projections can be compared against their actual performance at the end of the year. If students' actual scores are above the projections, then their teachers have provided instruction that contributed to obtaining appropriate growth. Conversely, if students' actual scores are below projections, then their teachers have not provided instruction that contributed to obtaining appropriate growth.

Quantile Regression has been applied to student testing scores within a school district in order to calculate a growth percentile for each student which can be viewed similarly to a child's growth chart assessment following a visit to the doctor. The median of a school's aggregated student growth percentiles can be calculated and compared to other schools' medians in the district. These comparisons contribute to schools obtaining recognition for performance ("Colorado," 2008, p. 10). The calculations needed to make these comparisons, however, deserve investigation as a technique to be applied to measuring teacher effectiveness.

2.2 Linear Mixed Models (LMMs)

The most prominent LMM approach to measure teacher effectiveness belongs to Dr. William Sanders of the SAS institute who implemented the Tennessee Value-Added Assessment System (TVAAS) in 1992 (Braun, Chudowsky, & Koenig, 2010, p. 3; McCaffrey, Lockwood, Koretz, & Hamilton, 2004, p. 2). This methodology is now packaged as the SAS Education Value Added Assessment System (EVAAS) for K-12 and commercially available for implementation by states and school districts. Tennessee, Pennsylvania, Ohio, and North Carolina presently use SAS EVAAS for K-12 as a means to measure the effects of teachers on the academic growth of their students (“SAS EVAAS,” 2010).

LMMs are statistical models that “quantify the relationship between a continuous dependent variable and various predictor variables” (West, Welch, & Galecki, 2007, p. 9). Datasets that can be analyzed with LMMs include: nested data (i.e. students in classrooms) and longitudinal or repeated measures studies (subjects are measured regularly over time or under different conditions) (West et al., 2007, p. 1). LMMs get their name from the fact that the model is a linear function of the predictor coefficients (parameters), and the predictors, themselves, can be a mix of fixed and random effects (West et al., 2007, p. 1). Fixed effects can be either continuous or categorical and describe the relationship between the predictors and the response for the entire population (West et al., 2007, p. 9). If an effect has factor levels that can expand during the course of study, then the effect can be considered a sample of the population and thus random (i.e., teachers will change during a study as new teachers emerge each year) (Lissitz & Doran, 2009, p. 24). Random effects model the random variation in the response variable for different levels in the data (West et al., 2007, p. 9).

Whether to estimate teacher effectiveness as a fixed or random effect has consequences for LMMs. If one considers teachers a random effect, then the variance of the estimates is reduced at the expense of introducing more bias (Braun et al., 2010, p. 52). Conversely, modeling teacher effectiveness as a fixed effect reduces bias but tends to produce “quite volatile” estimates particularly for teachers with small numbers of students (Braun et al., 2010, p. 52). By believing something is known about the distribution of teacher effects, “a large positive or negative estimate of the teacher effect is unlikely and is probably the result of random errors” (Braun et al., 2010, p. 52). Most well-known models estimating teacher effects specify the effects to be random. The random effects are calculated as Empirical Best Linear Unbiased Predictors (EBLUPs) and shrunk toward the mean with reduced variance but, perhaps, with the introduction of some bias (Braun et al., 2010, p. 52). One can then obtain variability estimates of the random teacher effects in order to make inferences about the random effects within the population (West et al., 2007, p. 2).

Students are often the subjects of analysis nested within teachers nested within schools. Longitudinal data exists when multiple evaluations are made on the same subject (student) over time. Evaluations of the same subject over time are most likely correlated, and LMMs capture this correlation by estimating the covariance parameters. With improvements in software, LMMs can now fit different covariance structures to the data while capturing this correlation. Depending on the covariance structure specified in formulation, efficiency can be obtained by not having to estimate the full covariance structure of the multivariate normal model. For example, one could specify that the covariance between random effects is zero, thus the structure of the D matrix (see section 2.2.1) for two random effects can be reduced:

$$D = \text{Var}(u_i) = \begin{pmatrix} \text{Var}(u_{1i}) & 0 \\ 0 & \text{Var}(u_{2i}) \end{pmatrix} \rightarrow \theta_D = \begin{pmatrix} \text{Var}(u_{1i}) \\ \text{Var}(u_{2i}) \end{pmatrix}$$

In addition to the benefit of being able to specify the covariance structure in the model formulation, one can also fit a LMM to a dataset with missing observations (West et al., 2007, pp. 2–3).

2.2.1 Linear Mixed Model General Specification

LMMs can be specified in a general or hierarchical manner. Statistical software packages such as SAS, SPSS, and R follow the general formulation, whereas Hierarchical Linear Model (HLM) software follows the hierarchical formulation (West et al., 2007, p. 1). This section follows the general formulation by West et al. in which a Linear Mixed Model is initially presented for just a single student i with scores from $1, \dots, n_i$ followed by a description of the matrices for the collection of students (2007, pp. 16–22)

$$Y_i = \underset{\text{fixed}}{X_i} \beta + \underset{\text{random}}{Z_i} u_i + \varepsilon_i \quad \begin{array}{l} u_i \sim N(0, D) \\ \varepsilon_i \sim N(0, R_i) \end{array}$$

Y_i is an $n_i \times 1$ observation vector of test scores for the i th student.

$$Y_i = \begin{pmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{n_i i} \end{pmatrix}$$

X_i is a known $n_i \times p$ matrix, which represents the values of the p predictors, such as previous test scores in a particular subject. In a model including an intercept term, the first column would be equal to 1 for all observations.

$$X_i = \begin{pmatrix} X^{(1)}_{1i} & X^{(2)}_{1i} & \cdots & X^{(p)}_{1i} \\ X^{(1)}_{2i} & X^{(2)}_{2i} & \cdots & X^{(p)}_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ X^{(1)}_{n_i i} & X^{(2)}_{n_i i} & \cdots & X^{(p)}_{n_i i} \end{pmatrix}$$

β is an unknown $p \times 1$ vector of fixed effects to be estimated from the data.

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Z_i is a known $n_i \times q$ matrix of observed values for the q predictor variables for the i th student that vary randomly across students. A model in which only the intercept is assumed to be random from student to student, the Z matrix would be a column of 1's.

$$Z_i = \begin{pmatrix} Z^{(1)}_{1i} & Z^{(2)}_{1i} & \cdots & Z^{(q)}_{1i} \\ Z^{(1)}_{2i} & Z^{(2)}_{2i} & \cdots & Z^{(q)}_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ Z^{(1)}_{n_i} & Z^{(2)}_{n_i} & \cdots & Z^{(q)}_{n_i} \end{pmatrix}$$

u_i is an unknown $q \times 1$ vector of random effects to be estimated from the data:

$$u_i = \begin{pmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{qi} \end{pmatrix} \sim N(0, D)$$

D is a $q \times q$ variance-covariance matrix that reflects the correlation among the random effects.

Elements along the main diagonal represent the variances of each random effect in u_i and off diagonal elements represent the covariance between two corresponding random effects. D is symmetric and positive definite (an $n \times n$ real symmetric matrix M is positive definite if

$z^T M z > 0$ for all non-zero vectors z with real entries).

$$D = \text{Var}(u_i) = \begin{pmatrix} \text{Var}(u_{1i}) & \text{cov}(u_{1i}, u_{2i}) & \cdots & \text{cov}(u_{1i}, u_{qi}) \\ \text{cov}(u_{1i}, u_{2i}) & \text{Var}(u_{2i}) & \cdots & \text{cov}(u_{2i}, u_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(u_{1i}, u_{qi}) & \text{cov}(u_{2i}, u_{qi}) & \cdots & \text{Var}(u_{qi}) \end{pmatrix}$$

ε is a non-observable $n_i \times 1$ random vector variable representing unaccountable random variation.

$$\varepsilon_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{n_i} \end{pmatrix} \sim N(0, R_i)$$

R_i is a positive definite symmetric covariance matrix of the form:

$$R_i = \text{Var}(\varepsilon_i) = \begin{pmatrix} \text{Var}(\varepsilon_{1i}) & \text{cov}(\varepsilon_{1i}, \varepsilon_{2i}) & \cdots & \text{cov}(\varepsilon_{1i}, \varepsilon_{n_i}) \\ \text{cov}(\varepsilon_{1i}, \varepsilon_{2i}) & \text{Var}(\varepsilon_{2i}) & \cdots & \text{cov}(\varepsilon_{2i}, \varepsilon_{n_i}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\varepsilon_{1i}, \varepsilon_{n_i}) & \text{cov}(\varepsilon_{2i}, \varepsilon_{n_i}) & \cdots & \text{Var}(\varepsilon_{n_i}) \end{pmatrix}$$

One assumes the residuals of different subjects are independent of each other and the vector of residuals, $\varepsilon_1, \dots, \varepsilon_m$, and random effects, u_1, \dots, u_m , are independent of each other.

One can also specify the LMM for all students as:

$$Y = \underset{\text{fixed}}{X} \beta + \underset{\text{random}}{Z} u + \varepsilon. \quad \begin{array}{l} u \sim N(0, G) \\ \varepsilon \sim N(0, R) \end{array}$$

Y is a $n \times 1$ vector where $n = \sum n_i$. This is a result of placing all Y_i as defined above on top of each other. The X matrix is $n \times p$ obtained by placing all X_i on top of each other. Z is a block-diagonal matrix, with the Z_i 's stated above on the diagonal. The u vector places all u_i on top of each other. The ε vector places all ε_i on top of each other. The G matrix is a block-diagonal matrix representing the variance-covariance matrix for all random effects with blocks of D as stated above for each subject along the diagonal. The R matrix is an $n \times n$ block-diagonal matrix representing the variance-covariance matrix for all residuals with blocks of R_i along the diagonal.

2.2.2 Linear Mixed Model Hierarchical Specification

The Hierarchical Model Specification belongs to the work of Raudenbush and Bryk (2002) and will involve three levels of data and models. It begins with the student-level predictor variables and one student-level outcome variable: test score gain. This constitutes the Level 1 Model (Student):

$$TestGain_{ijk} = b_{0jk} + \beta_1(X_{ijk}^{(1)}) + \beta_2(X_{ijk}^{(2)}) + \varepsilon_{ijk}$$

where $\varepsilon_{ijk} \sim N(0, \sigma^2)$. The outcome variable for student i with teacher j nested in school k depends on an unobserved intercept specific to teacher j nested in school k , the fixed effects β_1 and β_2 for the student predictors of $X^{(1)}$ and $X^{(2)}$, and student residuals.

Level 2 Model (Teacher):

$$b_{0jk} = b_{0k} + \beta_3(T_{jk}^{(1)}) + u_{jk}$$

where $u_{jk} \sim N(0, \sigma^2_{teacher})$. The level 1 intercept, b_{0jk} , teacher j nested in school k , depends on an unobserved intercept specific to school k , b_{0k} , a teacher specific effect β_3 for teacher predictor $T^{(1)}$, and a random effect, u_{jk} , associated with teacher j within school k .

Level 3 Model (School):

$$b_{0k} = \beta_0 + \beta_4(S_k^{(1)}) + u_k$$

where $u_k \sim N(0, \sigma^2_{school})$. The level 2 school specific intercept, b_{0k} , depends on the overall fixed intercept β_0 , a school specific effect β_4 for school predictor $S^{(1)}$, and the random effect u_k associated with the intercept for school k .

The nesting of students within teachers within schools is problematic with longitudinal data consisting of more than one test score for a student due to students not receiving instruction

from the same teacher every year. Even in a two-level model consisting of students and schools, the HLM structure is difficult to apply with longitudinal data due to the transient nature of some students as they move to different schools after a school year. Removing transient students in a two-level model as a remedy to the problem would likely not be appropriate for a researcher trying to develop the best model incorporating every type of student. With multiple yearly test scores for students as predictors, the HLM structure cannot be sustained. With a single test score as a predictor, the elegant structure of the HLM is easily satisfied (Lissitz & Doran, 2009, p. 25).

In order to implement an HLM and longitudinal data, the process would have to be a yearly one. A researcher would need to separate the longitudinal data into single year data and apply an HLM every year based on a single test score as a predictor. Although practical and efficient, this approach would yield less accurate predictions of student achievement since all prior known information would not be used.

In either the general or hierarchical specification of the LMM, a final model can be developed to predict MathGain scores for all students using data from the previous cohort of students. This is done to offset the inherent lack of randomization of students being assigned to teachers and the subsequent confounding nature of student test scores with other characteristics of students (Lissitz & Doran, 2009, p. 20). “This predicted score can be thought of as each student’s counterfactual level of achievement – that is, their predicted achievement had they been taught by a different teacher (say, the average teacher in the district)” (Corcoran, 2010, p. 10). The difference between the prediction and the student’s actual performance becomes the teacher’s value-added measure for that student. Averaging the value-added measures for a teacher’s students becomes his/her final value-added measure for the year (Corcoran, 2010, p. 10).

2.2.3 Linear Mixed Model Implementation in Texas

In order to meet several of its own legislative acts as well as Federal requirements to include growth in its AYP calculations, Texas developed the Texas Projection Measure (TPM) (Texas Education Agency, 2009, p. 6). The TPM uses Linear Mixed Models to augment their AYP calculations as discussed in Section 1.2. The Texas Education Agency's Growth Model Pilot Application to the U.S. Department of Education in January 2009 describes its Linear Mixed Model implementation methods whereby Texas is credited for students projected to meet proficiency and met proficiency with regard to AYP (Texas Education Agency, 2009).

The TPM is used to make projections of student achievement test scores in reading and mathematics at selected evaluation grades in the future (Texas Education Agency, 2009, p. 1). The models use current year scale scores in reading and mathematics and school-level mean scores in the projection subject as predictors. Development of the models, however, is accomplished the year prior with data from the previous cohort of students. For example, 3rd and 4th grade data from cohort 2019 can be used to develop a model to make a 4th grade prediction using Linear Mixed Models. The model is applied to the successive cohort (class that just finished 3rd grade, cohort 2020). The development of the models using the data of the previous year allows school administrators to know projected scores of their current students prior to the beginning of the school year.

In terms of AYP, projections contribute to the calculations by adding the number of students who are projected to meet the proficiency target at a specific evaluation grade in the future to the number of students who already meet the proficiency target at the current grade (Texas Education Agency, 2009, p. 12). This sum divided by the number of students in the particular grade determines the AYP percentage. This percentage is compared with the state

objectives for that grade. Whether aggregation occurs at the state, district, or school level, AYP above stated objectives defines success for the year.

2.3 Quantile Regression (QR)

A linear regression model identifies the conditional-mean function. It assumes constant variance and normality with its residuals. Difficulties arise when trying to overcome model inadequacy should assumption violations be present. Outliers also negatively affect a linear regression model by having an undue influence on the results. As a result, some linear regression models may not account for the full distributional properties of the response and can be invalid. An alternative modeling approach is desired to remedy these inadequacies (Hao & Naiman, 2007, pp. 22,24–25).

To address these issues of conditional-mean estimates, Quantile Regression was developed to estimate the effect of predictors on various quantiles that make up the distribution of the data. Similar to linear regression, Quantile regression models can deal with continuous response variables. Unlike linear regression, quantile regression can account for the full distributional properties of the response (Hao & Naiman, 2007, p. 29). In terms of educational data, estimation of a quantile is accomplished with students' scores at time t using the students' prior scores at times $1, 2, \dots, t-1$ as the conditioning variables (Betebenner, 2007, p. 3).

As a means to showcase Quantile Regression, a comparison with Linear Regression is developed in Table 2.1 using the QR notation of Hao and Naiman (2007).

Table 2.1: Linear and Quantile Regression Comparison

Model	Formulation	Equation
Linear Regression	$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$ $\varepsilon_i \sim N(0, \sigma^2)$	$E[y_i x_i] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Quantile Regression	$y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \varepsilon_i^{(p)}$ $0 < p < 1$	$Q^{(p)}[y_i x_i] = \beta_0^{(p)} + \beta_1^{(p)} x_i + Q^{(p)}(\varepsilon_i) = \beta_0^{(p)} + \beta_1^{(p)} x_i$ thus $Q^{(p)}(\varepsilon_i) = 0$

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of least-squares estimation minimize the sum of squared distances between data points (x_i, y_i) and the fitted line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Extending this concept to QR, one seeks to find estimators that minimize the sum of weighted vertical distances between data points and a fitted line where points below the fitted line are weighted $1 - p$ and points above the fitted line are weighted p . Each p (.05, .25, .50 for example) leads to a different fitted line, called a conditional-quantile function, that has p number of points below the fitted line and $1 - p$ points above the fitted line (Hao & Naiman, 2007, pp. 33–34). An example of the linear parameterization of QR is shown in Figure 2.1.

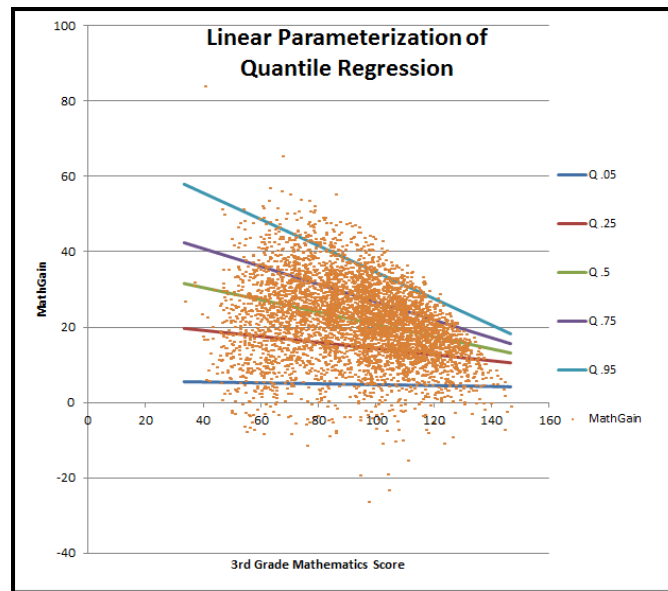


Figure 2.1: Linear Parameterization of Quantile Regression

2.3.1 Computing QR Coefficients

Each line in the (x, y) plane of the form $y = \beta_0 + \beta_1 x$ has a corresponding point (β_0, β_1) in the (β_0, β_1) plane. Conversely, lines in the (β_0, β_1) plane of the form $\beta_1 = (y/x) - (1/x)\beta_0$ correspond to points in the (x, y) plane. The goal is to find a line in the (x, y) plane that minimizes the sum of weighted vertical distances between the data points and the line. This line corresponds to a point in the (β_0, β_1) plane. For example, a sample of lines in the (x, y) plane has a corresponding set of points in the (β_0, β_1) plane that form polygonal regions. Figure 2.2 depicts lines in the (x, y) plane corresponding to points in the (β_0, β_1) plane of the same color with lines in the (β_0, β_1) plane corresponding to points in the (x, y) plane of the same color.

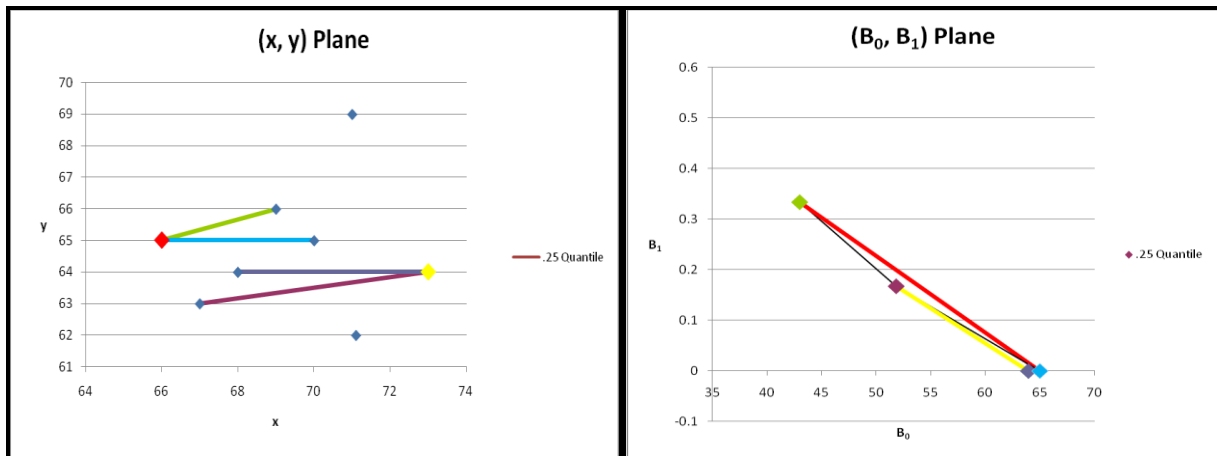


Figure 2.2: Point/Line Duality (Edgeworth, 1888)

The vertices of these polygonal regions in the (β_0, β_1) plane are extreme points. Each polygonal region in the (β_0, β_1) plane corresponds to a family of lines in the (x, y) plane that maintain the same number of points above or below the line. Based on exterior point algorithms for solving linear-programming problems, one can start at one of the vertices in the (β_0, β_1) plane and

iteratively move from vertex to vertex along the edge of the polygonal region, choosing at each vertex the correspondingly smallest value of:

$$p \sum_{y_i \geq \hat{y}_i} |y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)| + (1-p) \sum_{y_i < \hat{y}_i} |y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)| \quad (2-1)$$

In the (x, y) plane, one is iteratively moving from line to line defined by pairs of data points, at each step “deciding which new data point to swap with one of the two current ones by picking the one that leads to the smallest value in Equation (2-1) (Hao & Naiman, 2007, pp. 34–38).

Practically, the Quantile Regression Procedure in SAS “offers simplex, interior point, and smoothing algorithms for estimation” (“SAS/STAT 9.2,” 2009, p. 5354). Figure 2.2 depicts the point in the (β_0, β_1) plane that corresponds to the line in the (x, y) plane that minimizes Equation (2-1) for $p = .25$ to arrive at the .25 quantile.

2.3.2 Quantile Regression Extended to Cubic Splines

“A spline of degree 3 is a piecewise cubic curve whose values, slopes, and curvature coincide at the knots. Visually, a cubic spline is a smooth curve, and it is the most commonly used spline when a smooth fit is desired” (“SAS/STAT 9.2,” 2009, p. 387).

Given $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, one can approximate the data by fitting a cubic spline through two consecutive data points. By extending the formulation of quadratic splines by Kaw and Keteltas (2009), the cubic splines are the following (2009, p. 6):

$$f(x) = a_1 x^3 + b_1 x^2 + c_1 x + d_1 \quad x_0 \leq x \leq x_1 \quad (2-2)$$

$$= a_2 x^3 + b_2 x^2 + c_2 x + d_2 \quad x_1 \leq x \leq x_2 \quad (2-3)$$

⋮

$$= a_n x^3 + b_n x^2 + c_n x + d_n \quad x_{n-1} \leq x \leq x_n$$

These cubic splines generate $4n$ coefficients that can be determined by simultaneously solving $4n$ equations. $2n$ equations are created as a result of each spline going through two consecutive data points:

$$\begin{aligned} a_1x_0^3 + b_1x_0^2 + c_1x_0 + d_1 &= f(x_0) \\ a_1x_1^3 + b_1x_1^2 + c_1x_1 + d_1 &= f(x_1) \\ &\vdots \\ a_nx_{n-1}^3 + b_nx_{n-1}^2 + c_nx_{n-1} + d_n &= f(x_{n-1}) \\ a_nx_n^3 + b_nx_n^2 + c_nx_n + d_n &= f(x_n) \end{aligned}$$

The cubic splines must be continuous at interior points. At a particular interior point (where two splines meet), the first derivatives must also be equal (have the same slope). For example the first derivative of Equations (2-2) and (2-3) are equal at x_1 :

$$\begin{aligned} 3a_1x_1^2 + 2b_1x_1 + c_1 &= 3a_2x_1^2 + 2b_2x_1 + c_2 \\ \Rightarrow 3a_1x_1^2 + 2b_1x_1 + c_1 - 3a_2x_1^2 - 2b_2x_1 - c_2 &= 0 \\ &\vdots \\ 3a_{n-1}x_{n-1}^2 + 2b_{n-1}x_{n-1} + c_{n-1} - 3a_nx_{n-1}^2 - 2b_nx_{n-1} - c_n &= 0 \end{aligned}$$

These equations produce $n-1$ equations. Lastly, the second derivative of each spline at each interior point must be equal (curvature must be the same) (House, 2010, p. 6):

$$\begin{aligned} 6a_1x_1 + 2b_1 &= 6a_2x_1 + 2b_2 \\ \Rightarrow 6a_1x_1 + 2b_1 - 6a_2x_1 - 2b_2 &= 0 \\ &\vdots \\ 6a_{n-1}x_{n-1} + 2b_{n-1} - 6a_nx_{n-1} - 2b_n &= 0 \end{aligned}$$

This also produces $n-1$ equations. The sum of the equations generated is $2n + n - 1 + n - 1 = 4n - 2$. In order to produce two additional equations, assume the second derivatives are zero at the endpoints to produce a natural spline (Mathews, 2004, para. 2). A natural spline has endpoints that are inflection points.

$$f''(x_0) = 6a_1x_0 + 2b_1 = 0$$

$$f''(x_n) = 6a_nx_n + 2b_n = 0$$

With the $4n$ equations with $4n$ unknowns, one can solve for the coefficients to arrive at each cubic spline. An extension of the process above is to express each individual cubic spline $f(x)$ as a linear combination of the cubic spline basis functions to create a single interpolating function, which is more stable during numerical calculations (Kincaid & Cheney, 2002, p. 366). The interpolating function evaluates the spline curves in basis form.

2.3.3 Quantile Regression and B-Splines

The coefficients or weights of the cubic splines obtained in the previous section are determined explicitly given a select number of points that connect the cubic splines. Given a large dataset, however, one must determine the best coefficients for an interpolating function that describes the data knowing the boundaries and interior locations (knots) that help give the spline its shape. This is achieved through quantile regression in which the coefficients are chosen that minimize the sum of weighted residuals for a specific quantile. Following the text of Hao and Naiman, the p^{th} quantile-regression coefficients are the values that minimize the weighted sum of distances between \hat{y}_i and y_i , where a weight of $1-p$ is used if the fitted value \hat{y}_i overpredicts the observed value y_i and a weight of p is used if it underpredicts the observed value. Specifically, minimization of a weighted sum of residuals, $y_i - \hat{y}_i$, is desired where positive residuals receive a weight of p and negative residuals receive a weight of $1-p$:

$$p \sum_{y_i \geq \hat{y}_i} |y_i - \hat{y}_i| + (1-p) \sum_{y_i < \hat{y}_i} |y_i - \hat{y}_i| \quad (2007, p. 37).$$

The Colorado Department of Education currently implements a model that “parameterize[s] the conditional quantile functions as linear combinations of B-spline basis functions” (Betebenner, 2007, p. 5). As Betebenner (2007) points out, models of B-spline basis functions do a good job of interpolating data that is skewed or does not have constant variance:

“Using B-splines is attractive both theoretically and computationally in that they provide excellent data fit, seldom lead to estimation problems, and are simple to implement in available software. As will be seen when examining goodness-of-fit, use of B-splines instead of linear percentile curves leads to appreciable improvement in goodness-of-fit over the more common linear parameterization of the conditional percentile functions” (2007, p. 5).

The implication with education data is that B-splines can account for “slightly greater variability for higher... scale scores than for lower scores” (Betebenner, 2007, p. 5).

2.3.4 Quantile Regression Implementation in Colorado

Quantile Regression Analysis is used by the state of Colorado to calculate Student Growth Percentiles (SGP) and determine whether a student has made a year’s worth of growth over a period of a year. The resulting model calculates a SGP for each student based on a normative comparison with all other students with the same testing history. The minimum testing history is two successive Colorado Student Assessment Program (CSAP) tests in at least one academic subject (Betebenner, 2007, p. 2). SGPs can also be compared to the “50th percentile representing typical growth or one year’s growth in one year’s time” and evaluated in order to determine whether growth is sufficient “to reach proficient and advanced levels of achievement within one, two, and three years” (“Colorado,” 2008, p. 10).

For example, models are developed for five different percentiles with Quantile Regression using a dataset of scaled 3rd Grade Mathematics test scores and the MathGain obtained following the administration of the 4th Grade Mathematics Test. Three equally spaced

knots positioned between the minimum and maximum values of 3rd Grade Mathematics Score are chosen. Individual Student Growth Percentiles (SGP) are determined by plotting the individual's 3rd Grade Mathematics Score versus his/her MathGain to provide a reference for individuals with respect to the population. For example, a student with a score of 50 on the 3rd Grade Mathematics test obtains a 77 on the 4th Grade Mathematics Test. The MathGain of 27 is used to complete the plot in Figure 2.3 and obtain a SGP of 75%. Therefore, the student's growth was greater or equal to 75% of students with the same testing history ("Colorado," 2008, p. 9).

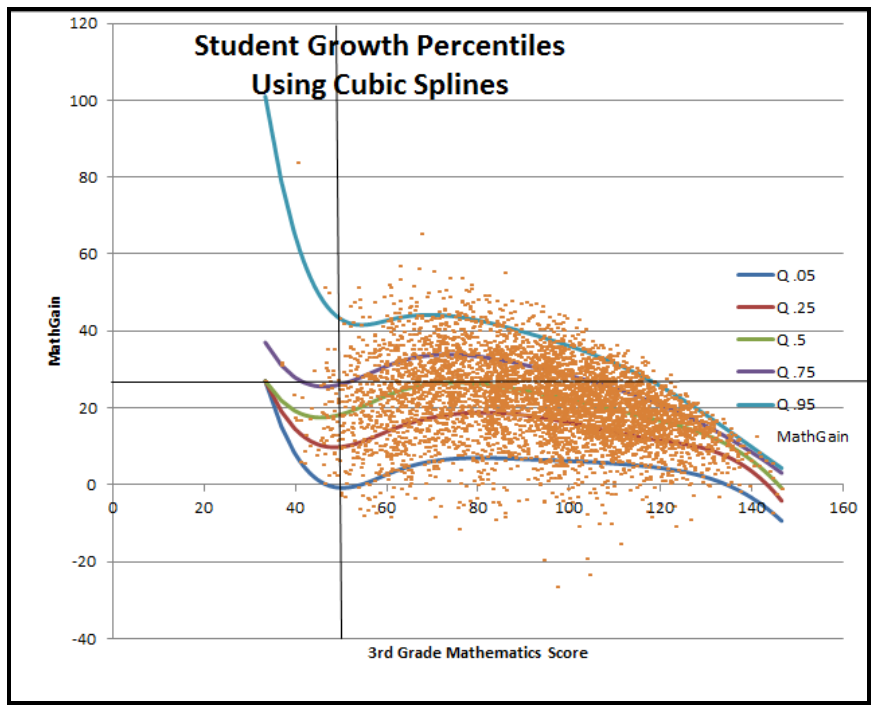


Figure 2.3: Cubic B-Spline Parameterization of Student Growth Percentiles

Colorado currently calculates the medians of individual SGPs to “quantify the level of student growth attained at specific schools and districts relative to other schools and districts within the state” (“Colorado,” 2008, p. 6).

“The median SGP computed for each school serves as an indicator of student growth associated with each school, ...describes a characteristic of a school’s students as a group and can be used to evaluate school outcomes, ...[and] measures the relative growth that has occurred for the group attending a specific school” (“Colorado,” 2008, p. 10).

An extension of this process to be discussed in Chapter 3 can be to calculate the median of individual SGPs for a teacher and compare it to the median of other teachers within the same population.

As stated previously, Quantile Regression chooses the best coefficients for each quantile interpolating function given boundaries of the dataset and interior locations, knots, which help give the spline its shape. One can choose any location for the desired knots when determining an interpolating function for each quantile. If knots that are not equidistant from each other are selected, then one has a Non-Uniform Rational B-Spline (NURB). Significant literature exists for different knot placement techniques that may lead to better fitting models. The Colorado Department of Education author, Betebenner, followed the work of Wei and He (2006) in which they preselected a set of knots for growth charts “using [their] general understanding of growth patterns” (Wei & He, 2006, p. 2073). Wei and He (2006) placed more knots during rapid changes in the data, such as during infancy and puberty, compared to other times. They stated, “In this paper, we do not go into the issue of automated knot selection” (Wei & He, 2006, p. 2073). Betebenner (2007), however, never states how he chose knots other than to say that the number and placement of knots would change the fit of the percentile curves (2007, p. 10).

Colorado had initially used Linear Mixed Models in 2004, but after a few years switched to student growth percentiles to address the main objective of a Colorado law (HB 07-1048) to determine whether students attain adequate yearly growth. The Colorado Department of Education determined that it could not adequately address this with the Linear Mixed Model approach, thus it changed due to the following shortcomings of the Model:

1. “The model fit a linear growth trajectory to each student and used that trajectory to predict future achievement. Longitudinal achievement of students across vertical CSAP scale is not linear but displays a negative concavity. The use of a linear trend resulted in higher predicted achievement than was likely for low achieving students” (“Colorado,” 2008, p. 7).
2. “The percentage of students projected to be proficient was strongly correlated with current status measures and likely confounded growth of students at a given school with their initial status” (“Colorado,” 2008, p. 7).

2.4 Discussion

The ability to definitively state that a student’s yearly growth in testing achievement is attributable to a particular teacher serves as one of the main goals of value-added modeling. The method to ideally determine teachers’ effectiveness would be to randomly assign students to teachers, measure the achievement of those students, and then make inferences regarding the differences in effectiveness between the teachers based on the scores. Although the randomization of students assigned to teachers is ideal for measuring teacher effectiveness, it does not exist practically. As a result, efforts to offset this lack of randomization are accomplished by constructing a model to predict a student’s yearly gain in testing with data from a previous cohort of students. The model provides predictions for successive students with the theory that those students were taught by the average teacher in the population being studied.

Model construction takes on added importance when determining which variables to include in addition to prior achievement that predict the yearly gain in testing achievement. Lissitz and Doran (2009) propose that variable selection should be based on the intended purpose of the study. For example, if the desire is to determine what student characteristics are associated with successful students, then the study should include background variables such as socioeconomic status, gender, or race. If the desire is to look for academic attributes associated

with the learning environment that can be changed for the betterment of students, then student background characteristics should not be included in a model (Lissitz & Doran, 2009, p. 27).

The No Child Left Behind Act of 2001 and the next reauthorization of the Elementary and Secondary Education Act does not/will not permit the use of student background characteristics in predictive growth models for determining whether a student can be counted for making Adequate Yearly Progress. Their exclusion prevents the implication of “different expectations for students of different sociodemographic classes” (Braun et al., 2010, p. 43). Sanders, Saxton, and Horn (1997) assert that the use of student longitudinal test data precludes including student background characteristics in their LMM (1997, p. 138). They state, “Each child can be thought of as a blocking factor that enables the estimation of school system, school, and teacher effects free of the socio-economic confounding that historically have rendered unfair any attempt to compare districts and schools based on the inappropriate comparison of group means”(Sanders et al., 1997, p. 138). Including socio-economic factors into analysis that is not longitudinal in nature does allow the comparison of districts, schools, and teachers without bias; however, the effort to gather accurate data that is often incomplete creates other problems that cannot be overcome (Sanders et al., 1997, p. 138). In addition to the difficulty of obtaining accurate and complete information regarding demographic attributes of students, Ballou, Sanders, and Horn (2004) determined that including such variables in the analysis had a negligible impact on the estimates of teacher effects.

Small sample sizes also present challenges for measuring teacher effectiveness. Precision of the results becomes greater for middle school teachers who may teach a greater number of students compared to an elementary school teacher who may teach a single class multiple subjects. Research performed by McCaffrey et al. (2004) consistently found large standard

errors such that about two-thirds of teacher effects from teacher effectiveness models are not statistically different from the mean (Braun et al., 2010, p. 45). The precision of the results may also lead to a lack of stability from year to year. McCaffrey, Sass, and Lockwood (2009) compared the teacher effectiveness results from two successive cohorts of students in four counties of Florida elementary and middle schools and found low correlations of teacher effectiveness between the two years (Braun et al., 2010, pp. 45–46).

2.4.1 Measures of Effective Teaching (MET) Project

The Bill and Melinda Gates Foundation initiated the MET project in the Fall of 2009 to “develop and test multiple measures of teacher Effectiveness” (“Working with Teachers,” 2010, p. 1). Similar to “A Blueprint for Reform”, the MET project cites research declaring that teachers have the greatest impact on student learning compared to other factors controlled by school systems. Therefore, the project aims at increasing the quality of teacher effectiveness information that is presently provided to education leaders to improve teacher feedback, direct professional development, and make informed decisions regarding teacher placement and retention (“Working with Teachers,” 2010, p. 3).

The MET project is led by several prominent academic institutions, nonprofit organizations, and for-profit education consultants. The project clearly states that teacher evaluation has two components: one based on student growth in standardized testing and another based on classroom-observed “aspects of teaching” that are valid predictors of student learning (“Working with Teachers,” 2010, pp. 4–5). The two components of teacher evaluation are constructed with five measures related to teacher effectiveness:

- 1) Student achievement gains on assessments
- 2) Classroom observations and teacher reflections

- 3) Teachers' pedagogical content knowledge
- 4) Student perceptions of the classroom instructional environment
- 5) Teachers' perceptions of working conditions and instructional support at their schools ("Working with Teachers," 2010, pp. 6–8)

Stage 1 of the project consists of measuring the unique influence of individual teachers on student growth for 2009-10 with a single VAM to establish baseline values. The VAM will use three years of student testing data and control for "student demographics and teacher characteristics (such as degrees, certification, licensing scores, tenure, district performance review ratings, years of experience, and [National Board for Professional Teaching Standards] (NBPTS) status)" ("Working with Teachers," 2010, p. 8). Stage 2 consists of combining measures two through five "to form a composite indicator of effective teaching" by assigning a weight to each measure based on how much each measure contributes to predicting student achievement gains ("Working with Teachers," 2010, p. 8). Lastly, stage 3 will attempt to show that the composite score of effective teaching is a stable predictor of teachers' student achievement gains. Included in stage 3 is a process to show if baseline value added measures accurately predict student achievement gains for 2010-11 in which students were randomly assigned to teachers unlike in 2009-10 ("Working with Teachers," 2010, p. 9).

The MET project is ambitious and comprehensive in scope. A preliminary report of the project was published in December 2010 with four general findings:

- 1) Teachers' value added estimates are one of the strongest predictors of teachers' future student achievement gains ("Learning about Teaching," 2010, p. 4).
- 2) "Teachers with the highest value-added scores on state tests also tend to help students understand math concepts or demonstrate reading comprehension through writing" ("Learning about Teaching," 2010, p. 4).
- 3) "The average student knows effective teaching when he or she experiences it" ("Learning about Teaching," 2010, p. 5).

4) “Valid feedback need not be limited to test scores alone” (“Learning about Teaching,” 2010, p. 5).

The ultimate goal of the MET project as with this research is to improve the quality of teachers by providing quality information to education leaders to make decisions regarding teachers’ professional development, placement, and retention.

2.4.2 A Risk-Mitigated Approach

Despite some of the concerns addressed above, quantifying teacher effectiveness using student growth has significant merit. Each approach comes with some risk and should not be the sole basis of measuring teacher effectiveness. Risk takes the form of identifying teachers that are effective when they are ineffective, and identifying teachers as ineffective when they are effective.

The amount of risk allowed in either scenario depends upon one’s perspective of the situation. If one views the situation from a statistical and student perspective where the null hypothesis is framed negatively, then the risk of identifying teachers as effective when they are ineffective should be extremely small. Therefore, the risk of identifying teachers as ineffective when they are effective would be correspondingly higher. For example, consider rejecting the

following null hypothesis: H_0 : Teacher is ineffective
 H_1 : Teacher is effective . If the teacher is, in fact, ineffective, then

this would be classified as Type I error. Alternatively, consider failing to reject the null

hypothesis: H_0 : Teacher is ineffective
 H_1 : Teacher is effective . If the teacher is, in fact, effective, then this would be

classified as Type II error.

Conversely, viewing the situation from the perspective of the teacher would require assuming smaller risk for identifying teachers as ineffective when they are effective. For example, consider rejecting the following null hypothesis: H_0 : Teacher is effective
 H_1 : Teacher is ineffective . If the teacher is, in fact, effective, then this would be classified as Type I error. Alternatively, consider failing to reject the null hypothesis: H_0 : Teacher is effective
 H_1 : Teacher is ineffective . If the teacher is, in fact, ineffective, then this would be classified as Type II error.

With either a student or teacher perspective regarding risk, a process is needed to combine different indicators of teacher effectiveness to arrive at an index of teacher effectiveness that contains less overall risk (sum of Type I and Type II error) than that of any single measurement. Ultimately, the goal is to produce a teacher effectiveness index with the smallest error as possible. Although the desire is to have an index that will provide a more nuanced approach to effectiveness (i.e., “at least” four effectiveness categories compared to the binary approach discussed above), placement of teachers across the boundary between “did not meet growth” and “meets student growth” must be met with redundancy to ensure any errors are minimized (“Alabama’s Race,” 2010, p. 89). In the end classroom observations by administrators serve to augment any objective result.

Several techniques to harness the desired redundancy exist in the literature. Principal Component Analysis can reduce dimensionality of highly correlated variables to allow Cluster Analysis to assign a set of items into groups of comparable quality. As a result, Principal Component and Cluster Analysis are reviewed in the following sections to show their applicability in producing a risk-mitigated teacher effectiveness index.

2.5 Principal Component Analysis

“Principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are 1) data reduction, 2) interpretation. Although p components [variables] are required to reproduce the total system variability, often much of this variability can be accounted for by a small number k of the principal components. The k principal components can then replace the initial p variables, and the original dataset, consisting of n measurements on p variables, is reduced to a dataset consisting of n measurements on k principal components. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result” (Johnson & Wichern, 2007, p. 430).

Following the development of Johnson and Wichern (2007), let random vector $X' = [X_1, X_2, \dots, X_p]$ have a covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$, where $\lambda = 0$ if and only if the column vectors are not linearly independent (2007, pp. 431–437). Consider the linear combinations:

$$\begin{aligned} Y_1 &= a_1' X = a_{11} X_1 + a_{12} X_2 + \dots + a_{1p} X_p \\ Y_2 &= a_2' X = a_{21} X_1 + a_{22} X_2 + \dots + a_{2p} X_p \\ &\vdots \\ Y_p &= a_p' X = a_{p1} X_1 + a_{p2} X_2 + \dots + a_{pp} X_p \end{aligned}$$

Obtain

$$\begin{aligned} \text{Var}(Y_i) &= a_i' \Sigma a_i \quad i = 1, 2, \dots, p \\ \text{Cov}(Y_i, Y_k) &= a_i' \Sigma a_k \quad i, k = 1, 2, \dots, p \end{aligned}$$

The principal components are those linear combinations Y_1, Y_2, \dots, Y_p whose variances are as large as possible, coefficient vectors are of unit length, and covariance between them is zero.

Let random vector $X' = [X_1, X_2, \dots, X_p]$ have a covariance matrix Σ with eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. The principal component is given by

$$Y_i = e_i' X = e_{i1} X_1 + e_{i2} X_2 + \dots + e_{ip} X_p \quad i = 1, 2, \dots, p$$

$$\text{with } \text{Var}(Y_i) = e_i' \Sigma e_i = \lambda_i \quad i = 1, 2, \dots, p$$

$$\text{Cov}(Y_i, Y_k) = e_i' \Sigma e_k = 0 \quad i \neq k$$

$$\rho_{Y_i, X_k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \dots, p$$

The proportion of total population variance due to the k th principal component =

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p}, \quad k = 1, 2, \dots, p.$$

“If most (for instance, 80 to 90%) of the total population variance, for large p , can be attributed to the first one, two, or three components, then these components can “replace” the original p variables without much loss of information” (Johnson & Wichern, 2007, p. 433).

Principal components can be calculated similarly if the variables are standardized.

Standardized variables need to be calculated if ranges of the variables are significantly different or units of measurement are dissimilar (Johnson & Wichern, 2007, p. 439). The i th principal component of the standardized variables $Z' = [Z_1, Z_2, \dots, Z_p]$ with $\text{Cov}(Z) = \rho$ is:

$$Y_i = e_i' Z \quad i = 1, 2, \dots, p \quad \text{with} \quad \sum_{i=1}^p \text{Var}(Y_i) = \sum_{i=1}^p \text{Var}(Z_i) = p \quad \text{and}$$

$$\rho_{Y_i, Z_k} = e_{ik} \sqrt{\lambda_i} \quad i, k = 1, 2, \dots, p$$

$(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ are the eigenvalue-eigenvector pairs for ρ , where

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. The proportion of standardized population variance due to k th principal

$$\text{component} = \frac{\lambda_k}{p}, \quad k = 1, 2, \dots, p.$$

Principal components with variables X_i that have large positive or negative coefficients typically have large correlations between that variable X_i and Y_i . Thus, both measures (coefficients and correlations) provide similar results of what they contribute to the component. Johnson and Wichern (2007) recommend, however, that both coefficients and correlations be examined to determine variable contribution to a component (2007, p. 434).

The development of principal components is normally an intermediate step leading to some final analysis. Principal component analysis can be an input to cluster analysis. To provide context for teacher effectiveness, principal components can be calculated to represent the varying metrics of teacher effectiveness for each teacher. Cluster analysis can then identify teachers with similar characteristics (principal components) in the data.

2.6 Cluster Analysis

Cluster analysis can be considered as an assignment of a set of items into groups so that items of the same group are of comparable quality. It is conducted without any assumptions regarding the number or structure of groups present in the data (Johnson & Wichern, 2007, p. 671). The groups are formed based on “distances” where “closeness” equates to being similar. Many distance measures exist in the literature to determine similarity (Euclidean, Minkowski, Canberra, etc.) (Johnson & Wichern, 2007, pp. 673–674). Due to the computationally expensive nature of examining all grouping possibilities, clustering algorithms have been developed to find good clusters without having to check all possible clustering configurations.

Hierarchical Clustering Methods form groups by either “agglomerative” or “divisive” techniques. With agglomerative techniques the most similar objects are first grouped together followed by combining those groups that are most similar. Divisive techniques place all objects

in a single group with subsequent subgroups partitioned away that are farther from the other objects in another subgroup (Johnson & Wichern, 2007, pp. 680–681).

An agglomerative hierarchical clustering method of J.H. Ward is based on minimizing the increase in an error sum of squares criterion (sum of squared deviations of every item in the cluster to the centroid). Each cluster begins as a single object with that object being the centroid. An iteration of the method considers every combination of merging two clusters. The merger of two clusters that produces the smallest increase in error sum of squares is completed. Iterations are performed until all objects are contained in a single cluster (Johnson & Wichern, 2007, pp. 692–693).

Several statistics exist to aid in determining the number of clusters that naturally exist in the data. For compact or slightly elongated clusters with a preference for roughly multivariate normal clusters, the three best statistics for hierarchical clustering methods are the pseudo F statistic, pseudo t^2 statistic, and the Cubic Clustering Criterion (CCC) (“SAS/STAT 9.2,” 2009, p. 245). The clustering methods produce the statistics at each step of the algorithm to evaluate the cluster solution. The pseudo F statistic measures “the separation among all the clusters at the current level” with local maximum values indicating a good number of clusters (“SAS/STAT 9.2,” 2009, p. 1267). It equals the ratio, $\frac{(T - P_G)/(K - 1)}{P_G / (n - K)}$, with T equal to the

total sum of squares, P_G equal to the within group sum of squares, and K equal to the number of clusters (“SAS/STAT 9.2,” 2009, p. 1258). Essentially it is “a ratio of the mean sum of squares between groups to the mean sum of squares within group” (Lim, Acito, & Rusetski, 2006, p. 508). The pseudo t^2 statistic measures “the separation between the two clusters most recently joined” (“SAS/STAT 9.2,” 2009, p. 1267). Good candidates for the number of clusters are “the

number of clusters one greater than the level at which [a] large pseudo t^2 value is displayed” (“SAS/STAT 9.2,” 2009, p. 1283). The pseudo t^2 statistic is a variant of Hoetelling’s T^2 in which large values provide evidence that the “two clusters being considered should not be combined since the mean vectors of these two clusters can be regarded as different” (Schmidhammer, 2010, p. 20) . Therefore, clusters can be combined for small values of the pseudo t^2 . “The CCC is based on the assumption that a uniform distribution on a hyperrectangle will be divided into clusters shaped roughly like hypercubes” (“SAS/STAT 9.2,” 2009, p. 245). Local maximum values of the CCC suggest a good number of clusters by rejecting the null hypothesis that the data has been sampled from a uniform distribution on a hyperrectangle. The alternative is then accepted that “the data has been sampled from a mixture of spherical multivariate normal distributions with equal variances and sampling probabilities...[and] the obtained R^2 value is greater than would be expected if the sampling was from a uniform distribution” (Schmidhammer, 2010, p. 17).

Nonhierarchical clustering methods are less computationally expensive than hierarchical methods as a matrix of distances between clusters do not have to be calculated. Larger datasets can be examined with nonhierarchical methods as a result. The method begins with user defined initial clusters or a set of seed points that form the centroid of the clusters. A popular nonhierarchical clustering method is the K-Means Method (MacQueen, 1967). The algorithm consists of three steps:

1. Place all items into user defined K clusters and calculate the initial centroid. Alternatively one can just specify K initial centroids.
2. Review each item and determine which centroid is closest. Assign item to the cluster of the nearest centroid. Recalculate the centroid for the gaining and losing cluster after assignment of each item.
3. Repeat Step 2 until no item switches clusters.

Johnson and Wichern contend that the final assignment of items is often dependent upon the initial partition or seed point of the cluster. Also, clustering methods that fix the number of clusters prior to analysis can be ineffective if the data includes outliers, if the data does not support the specified K clusters, and/or if random seed points place centroids near each other during step one above (Johnson & Wichern, 2007, pp. 696,701–702).

2.7 Summary

A review of the literature to evaluate the effectiveness of teachers has highlighted two general techniques. Firstly, LMMs produce estimates of the random teacher effect in the form of EBLUPs, and, secondly, LMMs produce predictions of student achievement scores in order to calculate teachers' value-added measures by comparing the students' actual achievement scores with the predicted scores. Colorado currently calculates SGPs using Quantile Regression to determine school effectiveness by comparing the medians of schools' student growth percentiles.

The two general techniques to measure teacher effectiveness produce risk of not properly identifying the quality of teachers. As a result, they should not be the sole basis of measuring teacher effectiveness. Multiple objective measures of teacher effectiveness must be employed to isolate the value that a teacher brings to the classroom. A teacher's "specific causal impact on learning cannot be discerned only from [a] single descriptive measure" ("Colorado," 2008, p. 10).

To expand the measures of teacher effectiveness, this research proposes to calculate two additional metrics. Specifically, Quantile Regression will be used to calculate a third objective measure of teacher effectiveness by analyzing student test scores within a school district and calculating a growth percentile for each student. The median of a teacher's aggregated student growth percentiles will be calculated and compared to other teachers' medians within the

population. Lastly, a fourth method to measure teacher effectiveness will be developed by combining Quantile Regression with the LMM practice of making achievement score predictions to derive teachers' value-added measures. The 0.5 quantile model generated by Quantile Regression will produce predictions of student achievement scores in order to calculate teachers' value-added measures by comparing the students' actual achievement scores with the predicted scores.

In order to exploit these multiple measures of teacher effectiveness, an analytical process involving principal component and cluster analysis will be developed. The principal components of the four objective measures of teacher effectiveness become composite indicators of teacher effectiveness, unlike Stage 2 of the MET Project that augments a single measure based on student achievement scores with a composite indicator based on measures not tied to student achievement scores. The principal components then serve as the inputs to Cluster Analysis. The process will isolate the value that a teacher brings to the classroom above that of his/her peers while minimizing the risk of not properly identifying the quality of teachers. The desired outcome is a precise and stable teacher effectiveness index to augment Alabama's formative educator evaluation system, EDUCATEAlabama.

Chapter 3 : Risk-Mitigated Teacher Effectiveness Index

3.1 Introduction

Various methods exist to measure teacher effectiveness using student growth. Each method generates some risk of not properly identifying the quality of teachers. As a result, one method should not be the sole basis of measuring teacher effectiveness. An analytical process that uses multiple and objective measures of teacher effectiveness to isolate the value that a teacher brings to the classroom above that of his/her peers warrants development for the state of Alabama.

The development of the Risk-Mitigated Teacher Effectiveness Index (RMTEI) consists of three phases. Phase I begins with the calculation of four teacher effectiveness metrics: Linear Mixed Model Teacher Effect, Overall Linear Mixed Model Value-Added Measure, Median Student Growth Percentile, and Overall Quantile Regression Value-Added Measure. Phase II consists of determining the principal components of the four metrics determined in Phase I and providing each teacher with a quantitative indicator of effectiveness. The principal components from Phase II serve as the inputs to Phase III, Cluster Analysis. Phase III will then illuminate teachers with similar characteristics (principal components) in the data in order to place them into effectiveness categories.

Through the use of an example, Chapter 3 describes the methodology of computing the RMTEI within the three phase process. The data for this example was obtained from the Early Childhood Longitudinal Study [United States] of the Kindergarten class of 1998-1999 ("Fifth Grade Data Codebook," 2006). Researchers of this study recorded a vast amount of data of a

singular class of students from kindergarten to fifth grade. The data involved every aspect of childhood education with investigation of schools, teachers, parents, and students. The nested, hierarchical nature of the data supports study at multiple levels. In order to provide an instrument to develop the RMTEI process, the data received some structuring to align with a typical school district of six elementary schools with five teachers per grade. Although teachers were appropriately linked with students in the Early Childhood Longitudinal Study, the number of students per teacher rarely resembled a classroom. However, the number of students per school presented desirable numbers for classroom analysis. Therefore, schools were regarded as the “teachers”, and the actual schools’ region were then analyzed at the “school” level.

This research finds the data restructuring appropriate for the purpose of demonstrating the RMTEI process. The analysis of Chapter 3 is truly at the school and region level while being called “teacher” and “school”. The literature supports the study of any two-level hierarchical data structure. Therefore, lowering the examination of the data by one level in name only does not alter the results or render them inadequate.

3.2 Phase I: Teacher Effectiveness Metrics

Every year a student earns a score in a particular subject on a standardized test that is vertically scaled to allow for comparisons from year to year. The process to compute scaled scores is called *equating* with one such method being Item Response Theory (IRT). The scores of the Chapter 3 example are IRT scores. “IRT uses the pattern of right, wrong, and omitted responses to the items actually administered in an assessment and the difficulty, discriminating ability, and ‘guess-ability’ of each item to place each child on a continuous ability scale. IRT scoring makes possible longitudinal measurement of gain in achievement over time, even though the assessments that are administered are not identical at each point. The common items

present...allow the scores to be placed on the same scale” (“Fifth Grade Data Codebook,” 2006, pp. 3–5, 3–6).

Every teacher within a district has a unique teacher identification code. The student’s teacher in that particular subject and year has his/her teacher identification code recorded against the student’s standardized test score. The result is a row vector of data with student, teacher, school, yearly score, ..., and yearly score. Given data of this nature for all students in a district, the desire is to extract what teachers contribute to their students obtaining a year’s growth in a year’s time as measured by the difference in the current year score and the previous year’s score. This difference is captured in a new variable, MathGain, to allow further analysis to focus on the current teacher. An excerpt of the example dataset is shown in Table 3.1 that illustrates the structure described above.

Table 3.1: Excerpt of Example Dataset for Phase I Metrics

CHILDDID	Teacher	School	3rd Grade Read	3rd Grade Math	4th Grade Math	MathGain
0011002C	11	1	155.81	98.47	135.40	36.93
0011004C	11	1	131.13	112.68	142.86	30.18
0011006C	11	1	119.44	87.24	105.69	18.45
0011007C	11	1	121.18	101.59	118.06	16.47
0011008C	11	1	119.68	88.27	109.84	21.57
0011009C	11	1	111.56	76.05	107.94	31.89
0011010C	11	1	98.41	76.52	115.36	38.84
0011011C	11	1	103.43	87.73	119.23	31.50
0011012C	11	1	103.93	85.12	114.84	29.72
0011013C	11	1	139.73	128.77	145.61	16.84

This example dataset consists of 718 students nested in 33 teachers nested in four schools and will be used to calculate all of the teacher effectiveness metrics.

The application of LMMs to longitudinal student testing data will provide two objective measures of teacher effectiveness by calculating individual teacher effects (see Section 2.2.1) and teacher value-added measures by comparing students’ actual achievement scores with their

predicted scores (see Section 2.2.2). The application of QR to longitudinal student testing data will provide two additional objective measures of teacher effectiveness. Subsequent to calculating a growth percentile for each student (see Section 2.3.4), the median of a teacher's aggregated student growth percentiles will be calculated and compared to other teachers' medians within the population. Lastly, the 0.5 quantile model generated by QR will produce predictions of student achievement scores in order to calculate teachers' value-added measures by comparing the students' actual achievement scores with the predicted scores. With all four Phase I metrics, larger values represent greater value that teachers bring to the classroom above that of their peers.

3.2.1 Linear Mixed Model Teacher Effect

The general LMM specification for an individual response follows:

$$\text{MathGain}_{ijk} = \beta_0 + \beta_1(X^{(1)}_{ijk}) + \beta_2(X^{(2)}_{ijk}) + \beta_3(T^{(1)}_{jk}) + \beta_4(S^{(1)}_k) + \left. \begin{array}{l} \text{\textit{fixed}} \\ u_{jk} + \varepsilon_{ijk} \end{array} \right\} \text{\textit{random}}$$

MathGain_{ijk} represents the value of the dependent variable for student i nested in teacher j nested in school k . β_0 through β_4 represent the fixed intercept and the fixed effects of the student ($X^{(1)}$, $X^{(2)}$), teacher ($T^{(1)}$), and school level predictors ($S^{(1)}$). u_{jk} is the random effect associated with the intercept for teacher j nested in school k , and ε_{ijk} represents the residual. The assumed distribution of the random effects associated with teachers nested in schools is:

$$u_{jk} \sim N(0, \sigma^2_{teacher}). \text{ The assumed distribution of the residuals of student scores is:}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2). \text{ The assumption is made that } u_{jk} \text{ and } \varepsilon_{ijk} \text{ are mutually independent.}$$

For this example, only student level predictors will be in the model:

$$\text{MathGain}_{ijk} = \beta_0 + \beta_1(X^{(1)}_{ijk}) + \beta_2(X^{(2)}_{ijk}) + \left. \begin{array}{l} \text{\} \text{ fixed} \\ u_{jk} + \varepsilon_{ijk} \end{array} \right\} \text{ random}$$

Sections 4.2.3 and 4.3.2 describe the process of selecting model predictors at the student, school, and teacher level. The PROC MIXED procedure in SAS is used to calculate the solutions for the fixed and random effects of the model with 3rd Grade Reading and Mathematics Test Scores predicting the MathGain obtained after completing 4th grade. The partial output of the solution for random effects is shown in Table 3.2.

Table 3.2: Partial Output of Solution for Random Effects

Effect	Teacher	School	Estimate
u11,1	11	1	1.0885
u128,1	128	1	-2.5025
u134,1	134	1	1.9121
u227,1	227	1	3.9484
u6290,1	6290	1	-0.2378

A result of modeling teacher effects as random is that the estimates for the teacher intercepts are Empirical Best Linear Unbiased Predictors (EBLUPs). EBLUPs are linear, unbiased, and have minimum variance among all linear estimators. They are also known as “shrinkage estimators” because the random teacher effects are smaller in value than if teacher effects were modeled as fixed (West et al., 2007, p. 45). The estimated teacher effect is a statistical numerical prediction of the relative value of a particular teacher and a “direct measure of the teacher deviation from the [district] mean” of the corresponding TestGain for a particular grade (Sanders et al., 1997, p. 156).

3.2.2 Linear Mixed Model Value-Added Measure

In either the general or hierarchical specification of the LMM, a final model can be developed to predict MathGain scores for all students using data from the previous cohort of

students. The difference between the prediction of the student’s score and the student’s actual performance becomes the teacher’s value-added measure for that student. Averaging the value-added measures for a teacher’s students becomes his/her final value-added measure.

Analyzing data from the previous cohort of students following their completion of 4th grade yields the following solution and model:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	30.2771	2.2158	40	13.66	<.0001
read3	0.07181	0.02170	676	3.31	0.0010
math3	-0.1922	0.02508	676	-7.66	<.0001

$$MathGain_{ijk} = 30.2771 + .07181(3rd\ Gr\ Read_{ijk}) - .1922(3rd\ Gr\ Math_{ijk})$$

Applying this model to the following cohort of students leads to predictions for MathGain, which can be compared to the students’ actual performance, and the calculation of teachers’ value-added measures. Partial output of the LMM Value-Added Measure is contained in Table 3.3.

Each row in Table 3.3 represents a child with a LMM MathGain prediction.

Table 3.3: Partial Output of LMM Value-Added Measure

CHILDDID	Teacher	School	3rd Grade Read	3rd Grade Math	4th Grade Math	MathGain	3rd Grade Math School Mean	LMM MathGain Prediction	LMM Value Added	Overall LMM Value Added
0011002C	11	1	155.81	98.47	135.40	36.93	96.0192	22.5399	14.3901	2.8233
0011004C	11	1	131.13	112.68	142.86	30.18	96.0192	18.0364	12.1436	2.8233
0011006C	11	1	119.44	87.24	105.69	18.45	96.0192	22.0866	-3.6366	2.8233
0011007C	11	1	121.18	101.59	118.06	16.47	96.0192	19.4534	-2.9834	2.8233
0011008C	11	1	119.68	88.27	109.84	21.57	96.0192	21.9058	-0.3358	2.8233
0011009C	11	1	111.56	76.05	107.94	31.89	96.0192	23.6714	8.2186	2.8233
0011010C	11	1	98.41	76.52	115.36	38.84	96.0192	22.6368	16.2032	2.8233
0011011C	11	1	103.43	87.73	119.23	31.50	96.0192	20.8427	10.6573	2.8233
0011012C	11	1	103.93	85.12	114.84	29.72	96.0192	21.3802	8.3398	2.8233
0011013C	11	1	139.73	128.77	145.61	16.84	96.0192	15.5615	1.2785	2.8233

The difference between the prediction and the student’s actual performance becomes the teacher’s value-added measure for that student. As discussed in Section 2.2.2 averaging the value-added measures for a teacher’s students becomes the teacher’s final value-added measure.

In Table 3.3 Teacher 11 obtained an overall LMM Value-Added Measure of 2.82. Therefore,

Teacher 11 contributed to the growth of his/her students that was, on average, 2.82 test score units more than predicted without teacher effects.

3.2.3 Median Student Growth Percentile

Based on the discussion in Section 2.3.4 Quantile Regression cubic B-spline models are developed for percentiles 1-99 using the dataset of scaled 3rd Grade Mathematics and Reading test scores. The variables of 3rd Grade Mathematics and Reading test scores predict the MathGain quantiles for all students. Individual Student Growth Percentiles (SGP) are calculated for all students by determining the quantile prediction closest to the actual MathGain for each student. The closest quantile prediction provides a reference of the student with respect to the population. For example, a student has scores of 140.86, 116.66, and 132.7 on the 3rd Grade Reading test, 3rd Grade Mathematics test, and 4th Grade Mathematics test, respectively. Therefore, MathGain is 16.04. The quantile closest to the student’s actual MathGain score is the 0.31 quantile. Therefore, the student obtains a SGP of 31. Upon determining the growth percentiles for all students, the median of a teacher’s aggregated SGPs is calculated and compared to other teachers within the same population. Partial output of the Median SGP metric is contained in Table 3.4.

Table 3.4: Partial Output of Median SGP Metric

Teacher	School	Median SGP
11	1	60.0
1119	4	68.0
1165	4	70.5
1211	4	75.0
1212	4	36.0

In Table 3.4 Teacher 11 from School 1 obtained a Median SGP of 60. Teacher 1212 from School 4 obtained a Median SGP of 36. Therefore, Teacher 11 contributed to the growth of

his/her students that was 10 percentile points greater than typical growth placing him/her near the top of the effectiveness ratings. Conversely, Teacher 1212 contributed to the growth of his/her students that was 14 percentile points less than typical growth placing him/her near the bottom of the effectiveness ratings.

3.2.4 Quantile Regression Value-Added Measure

Similar to calculating the LMM Value-Added Measure, Quantile Regression can use a single cohort of students to develop a model and make a prediction of MathGain for all students. For example, 3rd and 4th grade data from cohort 2019 can be used to develop a model to make a 4th grade prediction using QR. The model is applied to the successive cohort (class that just finished 3rd grade, cohort 2020). The goal is to predict each student’s MathGain along the 50th growth percentile using the student’s 3rd Grade Reading and 3rd Grade Mathematics scores. When cohort 2020 finishes the 4th grade, the students’ predicted MathGain can be compared against their actual mathematics gain. If a student’s actual gain is greater than the predicted gain, then the teacher’s value added measure (actual-predicted) is positive. If a student’s actual gain is less than the predicted gain, then the teacher’s value added measure (actual-predicted) is negative. Averaging the value-added measures for a teacher’s students becomes his/her final value added measure for the year. Partial output of the Overall QR Teacher Value-Added Measure is shown in Table 3.5.

Table 3.5: Partial Output of Overall QR Teacher Value-Added Measure

Teacher	School	Overall QR Value Added
11	1	-0.1691
1119	4	2.8532
1165	4	4.4563
1211	4	5.5380
1212	4	-2.7933

Each row in Table 3.5 represents a teacher with an Overall QR Value Added Measure. For example Teacher 1212 from School 4 contributed to the growth of his/her students that was, on average, 2.79 test score units less than expected.

3.3 Phase II: Principal Component Analysis (PCA)

Phase II of the process to determine the Risk-Mitigated Teacher Effectiveness Index consists of determining the principal components of the four metrics determined in Phase I. The Phase I metrics are shown for all 33 teachers in Table 3.6.

Table 3.6: Output of Phase I Metrics

Teacher	School	LMM Teacher Effect (test score units)	Overall LMM Value- Added Measure (test score units)	Median SGP (%)	Overall QR Value- Added Measure (test score units)
11	1	1.0885	2.8233	60	-0.1691
1119	4	2.3605	4.7360	68	2.8532
1165	4	2.6653	4.4240	71	4.4563
1211	4	2.4209	4.1959	75	5.5380
1212	4	-3.4491	-4.6155	36	-2.7933
128	1	-2.5025	-2.4597	34	-4.8604
134	1	1.9121	3.7594	67	2.5402
2092	2	4.8811	8.7511	86	5.3194
227	1	3.9484	7.1050	81	7.0392
242	2	-1.2985	-0.7180	42	-2.1503
288	2	0.3101	2.2470	46	-0.3239
337	2	-0.8301	-0.0814	39	-0.7286
349	2	-0.1022	0.8481	52	-0.6635
351	2	-1.8090	-1.6324	30	-2.4911
362	2	-0.2003	1.4775	44	2.0656
367	2	0.5780	1.6677	53	0.5973
379	2	3.7560	6.1880	73	4.4010
407	2	0.8735	2.7905	52	1.1864
430	2	-3.1356	-3.2973	27	-5.9083
442	2	-0.8623	-0.7168	38	-1.4145
452	2	-1.9112	-1.4598	31	-2.4287
455	2	-3.4295	-4.0439	34	-5.5067
621	3	-1.4355	-0.9905	40	-1.2516
6290	1	-0.2378	0.6780	49	0.4912
7393	2	-4.6901	-5.1979	24	-4.2607
947	3	0.6009	1.8247	53	-0.2030
966	4	4.8194	8.7399	82	7.3248
34	2	-0.1977	0.9397	51	0.3432
89	3	-0.2547	0.7611	41	-1.3759
97	2	-3.2805	-2.5730	27	-4.1427
130	3	-0.4213	0.4751	52	-0.2832
206	2	-0.3947	0.8868	33	-1.5541
247	2	0.2278	1.3363	48	-0.4403

PCA needs correlated responses to be effective and will transform input variables into a smaller set of uncorrelated variables without losing much information. The literature suggests that if the transformed space accounts for 80-90% of the variance, then PCA achieved its objective of dimension reduction. Based on the construction of the four Phase I metrics, which all measure teacher effectiveness with TestGain scores as the dependent variable, the expectation is that the metrics will be highly correlated. Not only is the expectation that the metrics will be highly

correlated, but also that the metrics will be transformed into a single variable with almost as much information as the original four. In fact this research hypothesizes this outcome to be a general conclusion applicable to all academic datasets.

In the case of the Chapter 3 example, the four Phase I metrics in Table 3.6 are highly correlated and will be projected onto a lower-dimensional surface in order to discover relationships between the effectiveness metrics. The correlation matrix of the Phase I metrics is shown in Table 3.7.

Table 3.7: Correlation Matrix of Phase I Metrics

	Phase I Metric Pearson Correlation Coefficient (p-value)		
Metric	Overall LMM Value- Added Measure	Median SGP	Overall QR Value- Added Measure
LMM Teacher Effect	.994 (.000)	.954 (.000)	.946 (.000)
Overall LMM Value-Added Measure		.939 (.000)	.939 (.000)
Median SGP			.937 (.000)

As stated in Section 2.5, the different units of measurement across all four of the metrics listed in Table 3.6 triggers calculating the eigenvalues and eigenvectors from the correlation matrix instead of the covariance matrix. The PROC PRINCOMP procedure in SAS accounts for this and produces the principal components of the Phase I metrics (LMM Teacher Effect, Overall LMM Value-Added Measure, Median SGP, and Overall QR Value-Added Measure).

The eigenvectors of the correlation matrix provide the coefficients of the newly constructed uncorrelated components. Let random vector $X' = [X_1, X_2, \dots, X_p]$ have eigenvectors e_1, e_2, \dots, e_p . The principal component is given by:

$Y_i = e_i' X = e_{i1} X_1 + e_{i2} X_2 + \dots + e_{ip} X_p \quad i = 1, 2, \dots, p$. The eigenvectors of the Phase I metrics are shown in Table 3.8.

Table 3.8: Eigenvectors of the Correlation Matrix

	Eigenvectors			
	Prin 1	Prin 2	Prin 3	Prin 4
LMM Teacher Effect	0.5051	-0.4186	-0.0045	-0.7547
Overall LMM Value- Added Measure	0.5023	-0.5589	-0.1293	0.6470
Median SGP	0.4967	0.3953	0.7650	0.1086
Overall QR Value- Added Measure	0.4958	0.5968	-0.6309	0.0046

Principal Component 1, 2, 3, and 4 then become:

$$\text{Prin 1} = .5051(\text{LMM Effect}) + .5023(\text{LMM Value}) + .4967(\text{Med SGP}) + .4958(\text{QR Value})$$

$$\text{Prin 2} = -.4186(\text{LMM Effect}) - .5589(\text{LMM Value}) + .3953(\text{Med SGP}) + .5968(\text{QR Value})$$

$$\text{Prin 3} = -.0045(\text{LMM Effect}) - .1293(\text{LMM Value}) + .7650(\text{Med SGP}) + -.6309(\text{QR Value})$$

$$\text{Prin 4} = -.7547(\text{LMM Effect}) + .6470(\text{LMM Value}) + .1086(\text{Med SGP}) + .0046(\text{QR Value})$$

Principal component scores for all 33 teachers are shown in Table 3.9.

Table 3.9: Output of Principal Component Scores

Teacher	School	Prin1	Prin2	Prin3	Prin4
11	1	0.7316	-0.2461	0.4369	0.0267
1119	4	1.9324	-0.0642	0.1630	0.0337
1165	4	2.2536	0.2667	-0.0096	-0.1009
1211	4	2.4541	0.6345	-0.0010	-0.0368
1212	4	-2.3370	0.7177	0.1430	-0.0720
128	1	-2.1869	-0.1907	0.3527	0.0120
134	1	1.6262	0.0911	0.2129	-0.0117
2092	2	3.8985	-0.3004	0.3512	0.0976
227	1	3.5741	0.3058	-0.1213	0.0585
242	2	-1.0695	-0.0240	0.1424	0.0086
288	2	0.0652	-0.3648	-0.1276	0.0767
337	2	-0.7765	-0.0392	-0.2958	-0.0411
349	2	-0.0976	0.0055	0.2502	-0.0138
351	2	-1.6977	-0.1224	-0.2883	-0.0749
362	2	0.1356	0.2136	-0.6237	0.0858
367	2	0.3567	-0.0124	0.0102	-0.0713
379	2	2.7958	-0.1542	0.0438	-0.1028
407	2	0.6486	-0.1506	-0.1612	0.0397
430	2	-2.8031	-0.3002	0.2466	0.0081
442	2	-1.0008	-0.0746	-0.1909	-0.1546
452	2	-1.6570	-0.0985	-0.2619	-0.0050
455	2	-2.6989	0.1087	0.5306	0.0104
621	3	-1.0647	0.1530	-0.0998	-0.0099
6290	1	-0.0697	0.1876	-0.0864	-0.0198
7393	2	-3.2327	0.4982	-0.0923	0.1320
947	3	0.2830	-0.1684	0.1729	-0.0476
966	4	4.0440	-0.0430	-0.2130	0.0894
34	2	-0.0026	0.1477	-0.0031	0.0248
89	3	-0.5725	-0.3388	-0.1214	-0.0548
97	2	-2.4764	-0.0840	-0.1027	0.1884
130	3	-0.1769	0.1742	0.1729	0.0149
206	2	-0.8235	-0.5356	-0.4227	-0.0351
247	2	-0.0555	-0.1925	-0.0069	-0.0552

The objective of PCA, however, is to provide a lower-dimensional surface as the input for Phase III, Cluster Analysis. Recall in Section 2.5, the eigenvalues of the correlation matrix provide the variance of the corresponding principal components with Principal Component 1 having the largest eigenvalue, Principal Component 2 having the second largest eigenvalue, etc. As shown in Table 3.10, PCA 1 accounts for 96.63% of the variance within the system.

Table 3.10: Eigenvalues of the Correlation Matrix

Eigenvalues of the Correlation Matrix			
Prin Comp	Eigenvalue	Proportion	Cumulative
1	3.8545	0.9636	0.9636
2	0.0774	0.0193	0.9830
3	0.0630	0.0157	0.9987
4	0.0052	0.0013	1.0000

The structure of the correlation matrix and the ability of PCA to reduce dimensionality efficiently confirm that the teacher effectiveness metrics are highly correlated. Furthermore, the scree plot in Figure 3.1 identifies the distinct bend at $i = 2$.

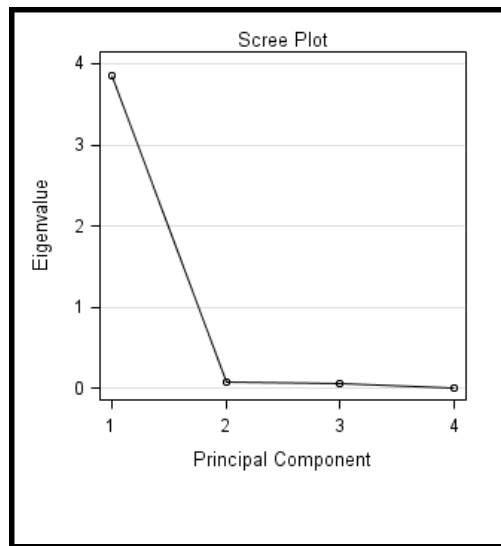


Figure 3.1: Scree Plot

There is clearly one dominant principal component as the remaining eigenvalues are relatively small and about the same size.

LMM Teacher Effect and Overall LMM Value-Added Measure are the largest contributors of the first principal component due to their coefficients. Both of the QR regression metrics, Median SGP and Overall QR Value-Added Measure, follow closely behind. LMM Teacher Effect and Overall LMM Value-Added Measure also share the largest correlations with the first principal component as displayed in Table 3.11.

Table 3.11: Correlation of Phase I Metrics with Principal Component 1

Correlation		
	Prin 1	Prin 2
LMM Teacher Effect	0.9917	-0.1165
Overall LMM Value-Added Measure	0.9862	-0.1555
Median SGP	0.9752	0.1100
Overall QR Value-Added Measure	0.9733	0.1660

The correlation of the metrics to the first principal component “measure only the univariate contribution of an individual [metric] to [the] component” (Johnson & Wichern, 2007, p. 433). It does not capture the importance of the metric to the principal component while considering the presence of the other metrics. Johnson and Wichern (2007) contend, however, that using coefficients, which is a multivariate evaluation of importance to the principal component, or correlations, tends to yield similar results when evaluating the contributions of the metrics to the principal components (2007, p. 434). As a result, the first principal component is a weighted sum of all Phase I metrics.

This dominant principal component represents the axis of greatest variability in a transformed space. For example, the highly correlated data provided by the four Phase I metrics are projected onto the two-dimensional surface in Figure 3.2.

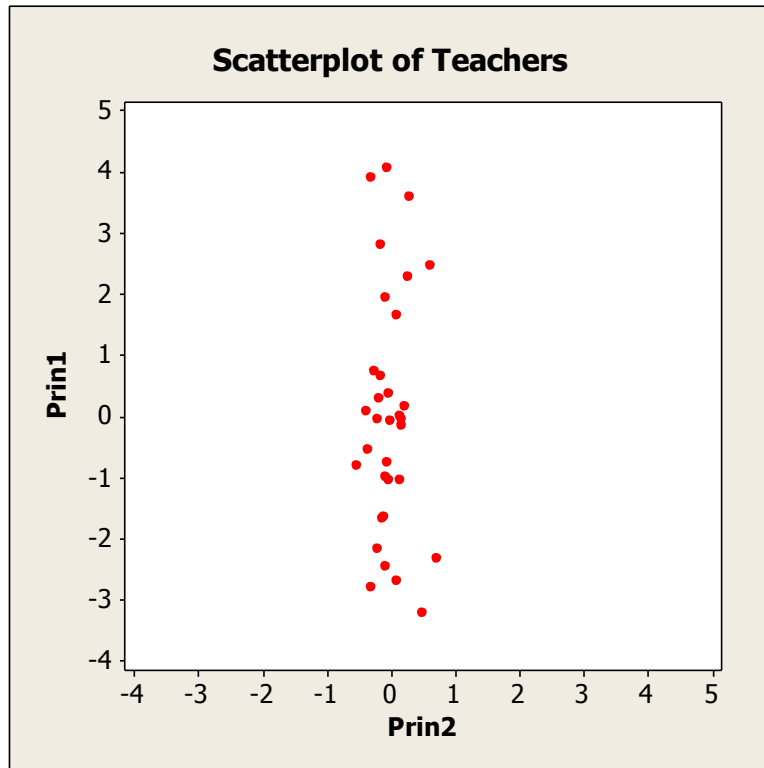


Figure 3.2: Subject-specific RMTEI Value along Dominant Principal Component Axis

Principal Component 1 is dominant and proceeds along the long axis of the data cloud in the transformed space. Effectively, one can represent all of the Phase I metrics and the majority of the variance within the system with a lone principal component. Therefore, teachers' principal component scores along this dominant principal component will become their subject-specific RMTEI values. If teachers instruct mathematics and reading, then an overall teacher effectiveness index will be obtained by taking the mean of their two subject indexes. Otherwise, the single, subject index will become teachers' overall values.

3.4 Phase III: Cluster Analysis

The purpose of Phase III, Cluster Analysis, is to illuminate teachers with similar characteristics (principal components) in the data in order to place them into effectiveness

categories. Figure 3.2, a simple scatterplot of teachers in two dimensions, provides the starting point to extract clusters of teachers with similar attributes. The scatterplot suggests that the clusters could be compact and elliptical in nature with the long axis of an ellipse in the direction of principal component 1 due to its greater range of values compared to principal component 2. No assumption is made, however, regarding the number of clusters to form. Therefore, hierarchical clustering techniques will be analyzed in an effort to derive the appropriate number of effectiveness categories to retain.

Ward's clustering method "is based on the notion that the clusters of multivariate observations are expected to be roughly elliptically shaped" (Johnson & Wichern, 2007, p. 693). It is a hierarchical, agglomerative clustering method as described in Section 2.6 that is based on minimizing the increase in an error sum of squares criterion (sum of squared deviations of every item in the cluster to the centroid). The scatter plot in Figure 3.2 suggests elliptical clusters. Also, a feature of the RMTEI process, as described in Section 3.3, is its ability to produce a dominant principal component that will continually lead to elliptical shaped clusters in two dimensions. Therefore, Ward's Method should be employed as a general prescription of the RMTEI process.

Ward's Method generated the pseudo F statistic, pseudo t^2 statistic, and the Cubic Clustering Criterion (CCC) in accordance with Section 2.6 after execution. The evidence directs fixing the number of clusters at $K = 5$. In Figure 3.3 local peaks of the CCC and pseudo F statistic as well as the point (number of clusters) prior to a large jump in value when viewing the pseudo t^2 plot from right to left are desired.

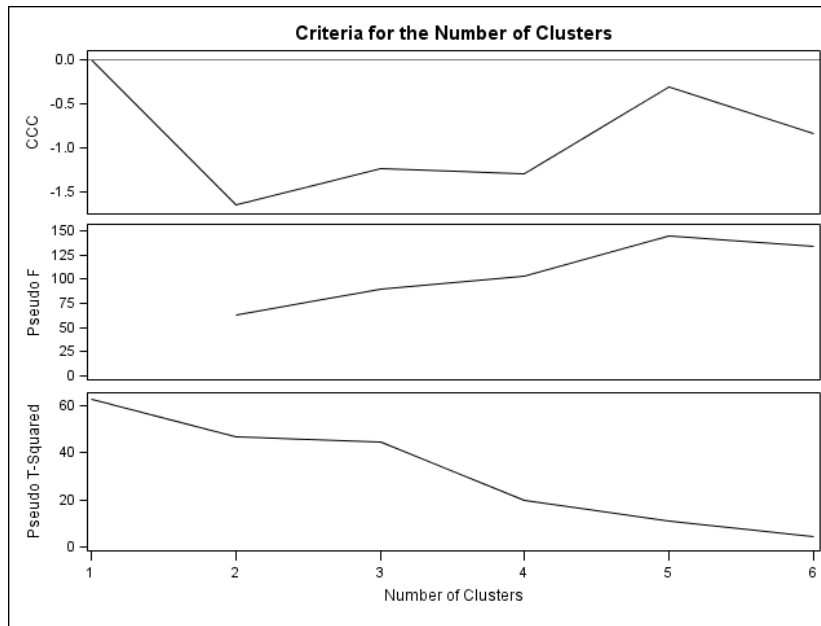


Figure 3.3: Ward’s Clustering Method Statistics for Determining Number of Clusters

Ward’s Method produced the Effectiveness Clusters in Figure 3.4 for $K = 5$ with the variables, principal component 1 and 2.

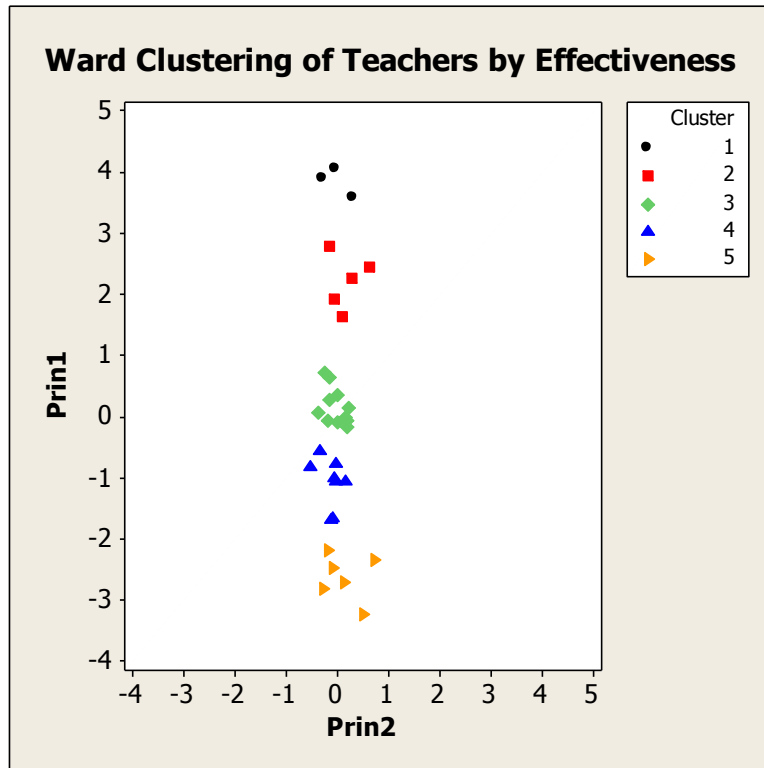


Figure 3.4: Ward’s Clustering of Teachers by Effectiveness

Having illuminated teachers with similar principal components in the data, the development of the Risk-Mitigated Teacher Effectiveness Index concludes with determining whether Ward’s Method produced clusters with statistically dissimilar characteristics.

Multivariate Analysis of Variance (MANOVA) confirms there is a statistical difference between the clusters at the $\alpha = 0.05$ level. The null hypothesis of equal mean vectors,

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, for all clusters is rejected where X_{11}, X_{12} are principal component 1 and 2 scores for cluster $l = 1, 2, \dots, 5$:

- Cluster 1: X_{11}, X_{12}
- Cluster 1: X_{21}, X_{22}
- \vdots
- \vdots
- Cluster 5: X_{51}, X_{52}

The alternative is accepted which states there is at least one cluster that has a statistically different mean vector. Conducting Analysis of Variance (ANOVA) on the variables, Principal Component 1 and 2, demonstrates that only Principal Component 1 is statistically different between the clusters at the $\alpha = 0.05$ level. Computing 95% Bonferroni simultaneous confidence intervals validates that each cluster is statistically different than every other cluster for Principal Component 1. For example, Cluster 1 is statistically different than Clusters 2, 3, 4, and 5. Cluster 2 is statistically different than Clusters 3, 4, and 5. Cluster 3 is statistically different than Clusters 4 and 5. The pattern concludes with Cluster 4 being statistically different than Cluster 5.

For the example dataset Ward’s clustering method produced five statistically different clusters of teachers while considering their principal component 1 scores. In Table 3.12, for example, teachers 2092 and 966 had comparable values for the Phase I metrics and demonstrated they had extraordinary gains in student growth. Likewise in Table 3.13, teachers 430 and 7393 had comparable values and demonstrated they had poor gains in student growth.

Table 3.12: Sample of Teachers with Extraordinary Gains in Student Growth

Teacher	LMM Teacher Effect	Overall LMM Value-Added Measure	Median SGP	Overall QR Value-Added Measure	Cluster
2092	4.8811	8.7511	86	5.3194	1
966	4.8194	8.7399	82	7.3248	1

Table 3.13: Sample of Teachers with Poor Gains in Student Growth

Teacher	LMM Teacher Effect	Overall LMM Value-Added Measure	Median SGP	Overall QR Value-Added Measure	Cluster
430	-3.1356	-3.2973	27	-5.9083	5
7393	-4.6901	-5.1979	24	-4.2607	5

3.4.1 Comparison of Clustering Results with Phase I Metrics

Figure 3.4 captures the Effectiveness Clusters produced by Ward’s Method for $K = 5$ with the variables, principal component 1 and 2. A natural question emerges; do these clusters provide better information than the clusters formed using only a single Phase I metric? The answer is “yes”. Ward’s clustering method produced the results contained in Figure 3.5 in which clustering was performed for each of Phase I metrics and the inputs from PCA (principal component 1 and 2).

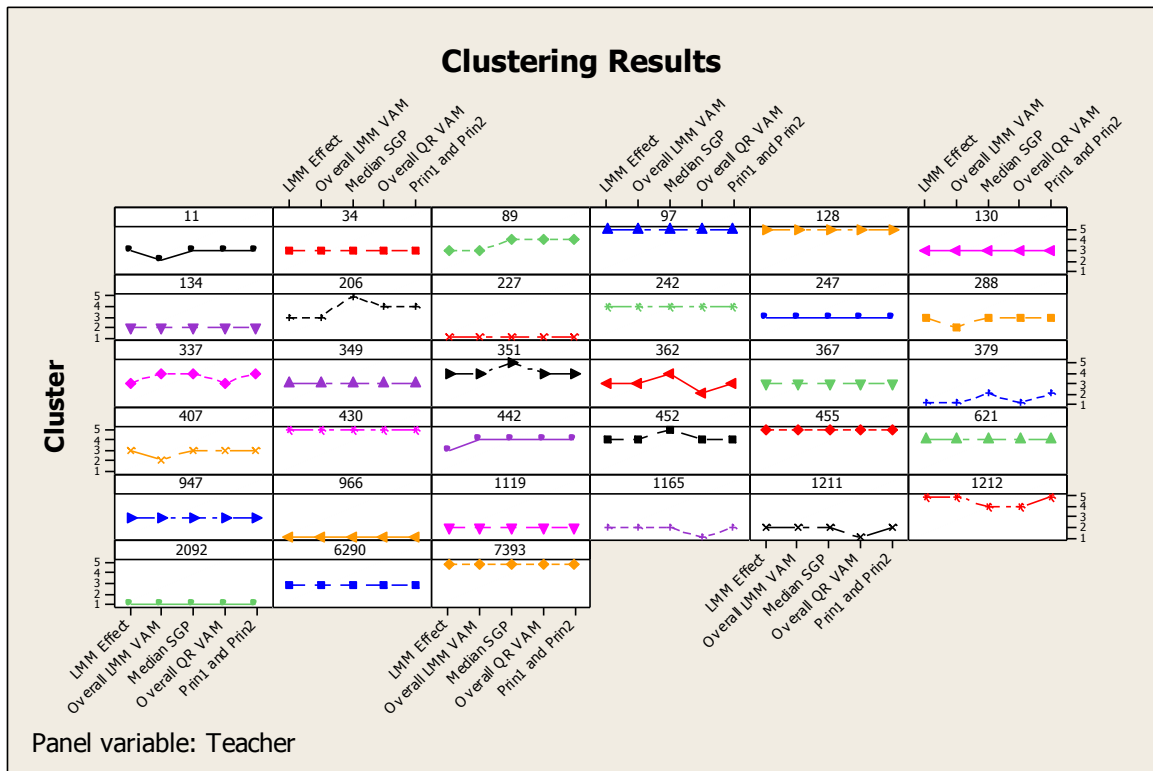


Figure 3.5: Comparison of Clustering Results with Phase I Metrics

The data highlights two thoughts. Firstly, the clusters are similar across the variables. This is expected due to the correlational structure presented in Table 3.7. Secondly, the clustering results using principal component 1 and 2 successfully represent the collection of the metrics’ results with no one metric being identical to the clustering results using the input from PCA.

Therefore, the final clustering results are not based on a single measure, and the risk associated with crafting accurate effectiveness categories has been mitigated.

3.5 Summary

Principal component scores for all 33 teachers are shown in Table 3.9. As shown in Table 3.10, PCA 1 accounts for 96.36% of the variance within the system. Teachers' PCA 1 scores are their Mathematics RMTEI values. If teachers also taught reading, then an overall teacher effectiveness index would be obtained by taking the mean of their two subject indexes. Otherwise, the Mathematics RMTEI values are the teachers' overall values.

Ward's clustering method produced five statistically different clusters of teachers while considering teachers' principal component 1 scores. Confidence in crafting clusters of teachers with similar attributes is paramount. Therefore, future work must consider the results of Ward's Method in determining the assignment of teachers to teacher effectiveness categories.

The Alabama State Department of Education desires to place teachers into "at least" four categories following the development of an objective measure of teacher effectiveness based on student growth that augments its presently used formative assessment, EDUCATEAlabama ("Alabama's Race," 2010, p. 89). Based on Ward's clustering method statistics with the example dataset, the evidence suggests fixing the number of clusters at five. A modification of the initially proposed categories in Section 1.4 complements the clusters found in Figure 3.4:

1. Extraordinary gains in student growth.
2. Significant gains in student growth.
3. Meets student growth.
4. Did not meet student growth.
5. Poor gains in student growth.

The Phase I metrics presented in this chapter are normative. The LMM Teacher Effect is a measure of the teacher deviation from the district mean. The LMM Value-Added Measure is based on the difference between actual scores and predicted scores had students been taught by the average teacher in the district. The QR Value-Added Measure is based on the difference between actual scores and typical gains within a district. The Median SGP is based on determining for each student in a district the percentage of other students that had less gain in testing achievement. Since everything is normative, how can one be sure, for example, that teachers at the bottom compared to their peers have poor gains such that the pursuit of college and workplace readiness has been negatively impacted? Regrettably, no assurances regarding adequate progress toward college and workplace readiness can be established following the RMTEI process. The only statement that can be made is that teacher gains are relative to their peers within the evaluated population.

As stated in Section 3.1 the results of Chapter 3 are truly measuring school effectiveness while being called “teacher effectiveness”. The literature supports the study of any two-level hierarchical data structure. Therefore, the practicality of labeling the data in this manner provided an instrument to showcase the methodology of the RMTEI.

In order to place the methodology of Chapter 3 into practice, longitudinal student and teacher data will be examined in Chapter 4. Success depends on the following:

- 1) Develop techniques with Alabama’s database infrastructure to streamline the process of establishing longitudinal data from existing yearly data
- 2) Develop techniques with Alabama’s database infrastructure to streamline the process of linking student achievement data with teachers

Further analysis will be undertaken in Chapter 4 to confirm the desired outcome of a precise and stable teacher effectiveness index with a discussion of the results. Chapter 5 provides an

assessment of Alabama's educator evaluation system, and then proposes an observational, summative assessment for Alabama that is a predictor of student achievement gains and correlated with the Risk-Mitigated Teacher Effectiveness Index. Lastly, Chapter 6 concludes the dissertation and provides recommendations for future study.

Chapter 4 : Data Analysis

4.1 Introduction

Implementation of the methodology discussed in Chapter 3 began with the acquisition of student achievement and teacher data. A significant amount of energy was put forth to obtain quality student and teacher data from the Alabama State Department of Education (ALSDE) and Alabama Local Education Agencies (LEA). This consisted of extensive written and verbal communication with individuals from these organizations who have the authority and ability to provide the data. Despite receiving approval from the ALSDE and several LEAs at the beginning of the process, action to provide the data rarely took place despite repeated requests at all levels of the organizational structure. In the end only a single district provided the requisite data to fully implement the RMTEI process. The district is classified as medium-sized with a student population of approximately 4,000 students (“System Profile,” 2009). It is situated in a suburban area with approximately 50% of its students available for free/reduced lunch (“System Profile,” 2009).

Examining only this single Alabama school district would limit the applicability of this research due to the lack of a large, urban school district. A large, urban school district would provide more breadth in the analysis by presenting an additional learning environment to the RMTEI process. In order to remedy this limitation, data was obtained from the National Center for Education Statistics (NCES), which “is the primary federal entity for collecting and analyzing data related to education” (“National Center for Education Statistics,” n.d.). The NCES falls

under the Institute of Education Sciences, which “is the research arm of the U.S. Department of Education” (“Institute of Education Sciences,” n.d., para. 1)

The NCES data consisted of student achievement and teacher data from 418 elementary schools (Kindergarden-5th Grade) in 17 urban districts from across the United States. “Although the districts selected for the study did not form a statistically representative sample of the nation, they were drawn from 13 states with a variety of regulatory, administrative, and demographic contexts” (Glazerman et al., 2010, p. 3). The data began with testing in 2005 and concluded in 2008. The data’s original use consisted of evaluating the benefits of comprehensive teacher induction training for beginning teachers compared to the existing, less intensive induction services provided by the district (Glazerman et al., 2010, p. vi). The benefits of comprehensive teacher induction could be positive impacts on classroom practices, a positive impact on teacher retention, and a statistically significant impact on student achievement (Glazerman et al., 2010, p. viii). Observational evaluations were conducted in 2006 by trained observers to score the effective teaching practices of both teacher groups (Glazerman et al., 2010, p. xiv). These observational evaluations will be discussed and compared to the teacher effectiveness outcomes obtained with the RMTEI in Chapter 5.

4.2 Alabama Data Analysis

Upon acquiring five years of student and teacher data from an Alabama district, methods were developed to accomplish two of this research’s sub-objectives:

- 1) Develop techniques with Alabama’s database infrastructure to streamline the process of establishing longitudinal data from existing yearly data
- 2) Develop techniques with Alabama’s database infrastructure to streamline the process of linking student achievement data with teachers

As stated in Section 1.4, Alabama claims to have the ability to link student data with teachers, yet it does not provide the linked testing data to the districts or provide analysis as a result of the linkage. Districts only receive yearly student achievement data via distributed compact disks from the Alabama State Department of Education. The lack of effective data extraction methods limits the ability of districts to analyze testing data.

This research developed procedures to create longitudinal student data linked with current year teachers. For example, analysis for 2009 required merging student scaled scores from the 2006-07, 2007-08, and 2008-09 Alabama Reading and Mathematics Test (ARMT) with the students' respective mathematics and reading teachers for the 2008-09 school year. Much of this pre-analysis work consisted of extracting students and their teachers from a schedule dataset using the desired grade and course (mathematics or reading). Care was taken, especially in older grades, to find the appropriate courses/teachers that would ultimately be linked with the reading and mathematics student achievement data of the current year. See Appendix 1 for the SAS code and its description that accomplishes these two research sub-objectives.

Similar to the example in Chapter 3, every teacher and student within the district had a unique identification code. Teachers were already linked to students in the schedule dataset. Student achievement files consisted of a matrix with column vectors of Student Identification Number, grade, school, and a single year of ARMT Reading and Mathematics Scores. The student identification number from the student achievement files provides the linkage to teachers/students in the schedule dataset. Upon merging the files, the desire is to extract what teachers contribute to their students obtaining a year's growth in a year's time as measured by the difference in the current year score and the previous year's score. This difference is captured in the new variables, ReadGain and MathGain, to allow further analysis to focus on the current

teacher. An excerpt from the 6th grade, 2009 dataset is contained in Table 4.1 that illustrates the structure described above.

Table 4.1: Excerpt from 6th grade, 2009 Dataset

Student ID Number	Grade SY2009	School	ARMT_Read_2007	ARMT_Math_2007	ARMT_Read_2008	ARMT_Math_2008	ARMT_Read_2009	ARMT_Math_2009	MTeacher	RTeacher	ReadGain	MathGain
1525747585	6	40	653	652	663	712	686	689	4260	5336	23	-23
1747572060	6	40	583	614	566	646	635	657	4260	6294	69	11
1013406288	6	40	711	652	695	706	683	689	4260	6294	-12	-17
1818385928	6	40	698	647	684	664	698		6124	8023	14	
1299184676	6	40	632	660	656	690		699	6124	3245	13	9
1940094310	6	40	662	602	667	632	672	634	6124	3245	5	2
1201820789	6	40	644	619	663	658	666	677	4260	6294	3	19
1171259616	6	40	662	660	717	700	725	732	4260	6294	8	32
1354083103	6	40	567	607	570	606	616	613	5948	8815	46	7
1600810714	6	40			570	554			6124	8023		

4.2.1 Alabama Reading and Mathematics Test

The ARMT is a criterion-referenced test administered to 3-8 Grade students. It is built using portions of the Stanford Achievement Test (Stanford 10) and additionally developed subtests. This combination ensures that all state-level content standards are appropriately tested. A student must take the items in Table 4.2 to receive ARMT reading and mathematics scores.

Table 4.2: Components of Alabama Reading and Mathematics Test

ARMT Reading	ARMT Mathematics
Stanford 10 Word Study Skills (Grade 3)	Stanford 10 Mathematics Procedures (Grades 3-8)
Stanford 10 Reading Vocabulary (Grades 3-8)	Stanford 10 Mathematics Problem Solving (Grades 3-8)
Stanford 10 Reading Comprehension (Grades 3-8)	ARMT Part 2 Mathematics Subtest
ARMT Part 2 Reading Subtest	

In addition to providing the points possible, points earned, percent correct, and achievement levels, the ARMT provides scaled scores that can be used to determine the amount of progress that is earned between years. These scaled scores are a focus of this research and contribute to studying change in performance over time, which is one of the primary purposes of the ARMT. Additionally, ARMT results for grades 3-8 are used to meet No Child Left Behind legislation

requirements with Adequate Yearly Progress determinations (“Alabama Reading and Mathematics Test: Interpreting the Student Report,” 2009, p. I–1; Pugh, 2008).

4.2.2 Testing Histories

To prepare for calculating the Phase I metrics, mutually exclusive groups of students, each with a different testing history, were created in order to compare similar students. The students of a typical school district have different testing histories based on their grade and multiple, personal circumstances. Students in grades 3-5 have fewer testing opportunities, thus they have immature testing histories. Although students in grades 6-8 should have robust testing histories, there are many reasons why that is not always the case. For example, new students from outside of the State and absences during testing contribute to a lack of data for students. Even new students from outside of the district, but within the State, contribute to the issue as there is not a systemic process at the district level to electronically capture the testing data from the previous district.

Testing histories consist of consecutive, yearly mathematics and readings scores in the ARMT with the minimum testing history being two consecutive years. For example, analysis in 2009 creates two testing histories for grades 5-8 (4th grade has a single testing history of Mathematics and Reading scores in 2008 and 2009):

- 1) Mathematics and Reading scores in 2007, 2008, and 2009
- 2) Mathematics and Reading scores in 2008 and 2009

Linear Mixed Model Phase I metrics were calculated within the testing history groups and then placed together in order to calculate an overall value for each metric. In the case of the Linear Mixed Model Teacher Effect, teachers received an effect for the testing histories in which they were present. A weighted sum for each teacher was then calculated based on the number of

students in each testing history group. Secondly, Linear Mixed Model Value Added Measures were calculated for each student within a testing history group. The testing history groups were placed back together with their new value added measures. The mean was then calculated for each teacher to determine the final Value Added Measure.

Student Growth Percentiles on the other hand were calculated for each student within a single testing history group of two years. This ensured one large testing history group to create 99 quantiles within the data. The median was calculated for each teacher to determine the Median Student Growth Percentile metric. When multiple testing history groups were considered, the medium-sized district would produce small testing history groups that did not accommodate the number of parameters to be estimated. For example analysis of 6th grade in 2011 resulted in a student testing history group of 17 students that had mathematics and reading scores in 2009, 2010, and 2011. This testing history group encountered errors with all previous year scores in the predicted subject and the prior year score in the non-predicted subject as model predictors (i.e., MathGain would have Mathematics scores in 2009 and 2010 and Reading scores in 2010 as predictors); see Section 4.2.3 for discussion of model predictors. Quantile Regression attempted to calculate seven parameters (a_0, a_1, a_2, a_3 and b_1, b_2, b_3) for each of the three independent variables in order to calculate the independent variable's component of the QR prediction:

$$\hat{y}_i = \sum_{j=0}^k a_j x_i^j + \sum_{j=1}^n b_j (x_i - t_j)_+^k \quad \text{where} \quad (x_i - t_j)_+^k = \begin{cases} (x_i - t_j)^k & \text{if } x_i \geq t_j \\ 0 & \text{if } x_i < t_j \end{cases}$$

$k = 3$ (degree), $n = 3$ (total number of knots), $t_j = \text{knot location}$

In this example the QR prediction for each student would be the sum of the three independent variable components plus an intercept. The number of parameters was greater than the number of observations, thus the model lacked the requisite degrees of freedom and did not provide a QR

prediction for each student. As a result, individual SGPs could not be calculated to allow for the calculation of teachers' Median SGP. Lastly, Quantile Regression Value-Added Measures were also calculated for each student within the minimum, two year testing history. The mean was calculated for each teacher to determine the final Quantile Regression Value Added Measure.

4.2.3 Student, Teacher, and School Level Predictors for Alabama Data

Any future effort to implement the RMTEI process at the district level should undergo an evaluative procedure to assess the adequacy of model predictors. This research examined predictors at all levels (student, teacher, and school) in an attempt to improve the predictive capability of the models used to generate the four Phase I metrics. The process began with using a student's entire testing history to predict MathGain and ReadGain. For example, a testing history in 2009 consists of Mathematics and Reading scores in 2007, 2008, and 2009. Therefore, ARMT Reading 2007 - ARMT Reading 2008 and ARMT Mathematics 2007 - ARMT Mathematics 2008 predict ReadGain and MathGain in separate models. This method of analysis produced similar results for all grades and years in which the scores of all previous years in the predicted subject were significant at an $\alpha = .05$, and only the prior year score in the non-predicted subject were statistically significant. For example, only the prior year reading score was significant for MathGain in addition to all previous math scores.

Adding predictors at the teacher level led to varying results. Teacher data provided by the districts included teaching experience expressed in months teaching in district, state (not including the district), public (outside the district or state), and private systems. It also included the teacher's highest degree obtained, expressed as a bachelor's, master's, greater than 6 years (working toward a doctorate), or doctorate degree. In many instances, teaching experience and highest degree obtained were not statistically significant.

Adding predictors at the school level also led to varying results. This research applied the methods of the Texas Projection Measure, discussed in Section 2.2.3, by considering campus mean of the predicted subject as a predictor of student achievement. In many instances, MathMean and ReadMean for a given campus were not significant predictors of MathGain and ReadGain respectively. Secondly, this research obtained the Reduced/Free lunch percentages of the different campuses within the two districts and considered them as predictors of student achievement. Obtaining results similar to using the campus mean of the predicted subject as a predictor, schools' Reduced/Free lunch percentages were not significant predictors of student achievement.

Output follows for the solution of fixed effects from 4th grade, 2011, where the predictor, myrs_employ, is a math teacher's experience expressed in years. Included are student, teacher, and school level predictors as described above. All of the teacher and school level predictors for the minimum testing history group failed to reject the null hypotheses,

$H_0 : \beta_3 = 0, H_0 : \beta_4 = 0, H_0 : \beta_5 = 0, H_0 : \beta_6 = 0$ at an $\alpha = .05$ significance level:

Solution for Fixed Effects						
Effect	mhighest_degree	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		784.97	422.11	9	1.86	0.0959
ARMT_Math_2010		-0.4637	0.04811	256	-9.64	<.0001
ARMT_Read_2010		0.2700	0.05393	256	5.01	<.0001
Percent_Free_Reduced		8.7894	28.4892	256	0.31	0.7579
MathMean		-1.0404	0.6834	256	-1.52	0.1291
myrs_employ		0.2164	0.2985	256	0.73	0.4691
mhighest_degree	6	-0.6414	11.5877	256	-0.06	0.9559
mhighest_degree	B	0.2089	5.5794	256	0.04	0.9702
mhighest_degree	M	0

The result of examining model predictors for the Alabama district led to an approach similar to the Chapter 3 example in which only student-level predictors were included in the models. The only difference, however, is that multiple years of test scores in the predicted subject were included for LMMs that utilized testing histories as discussed in Section 4.2.2.

Table 4.3 summarizes the outcome with each check mark representing the inclusion of the independent variable(s) in the corresponding model for each Phase I metric.

Table 4.3: Summary of Model Predictors for Alabama Data

Metric	LMM Teacher Effect		Overall LMM Value-Added Measure		Median SGP		Overall QR Value-Added Measure	
	MathGain	ReadGain	MathGain	ReadGain	MathGain	ReadGain	MathGain	Readgain
Dep Var Indep Var								
Math Score (previous year only)		✓		✓	✓	✓	✓	✓
Math Score (all previous years)	✓		✓					
Reading Score (previous year only)	✓		✓		✓	✓	✓	✓
Reading Score(all previous years)		✓		✓				
Teacher-Level Predictors								
School-Level Predictors								

4.3 National Center for Education Statistics Data Analysis

The data from the National Center for Education Statistics is significant in that it captures student achievement data linked to 1009 beginning teachers from 17 urban districts over 4 years.

Several items of interest with regard to the data are the following:

- 1) the data's original use consisted of evaluating the benefits of comprehensive teacher training for beginning teachers compared to the existing, less intensive training services provided by the district;
- 2) each district chose one of two providers for the comprehensive training services;
- 3) comprehensive training services lasted one or two years;
- 4) schools within districts were randomly assigned to receive comprehensive training services; and
- 5) student achievement data are recorded as z-scores to account for the difficulty of comparing 1009 teachers across 43 different tests that use scaled scores, normal curve equivalents, percent correct, or percentile rankings (Glazerman et al., 2010, p. 33)

The student achievement data is comprised of three files (2006, 2007, 2008) each representing a cohort of students who are taught by the same beginning teachers. Each file has a student pre-test and post-test score for mathematics and reading. The pre-test score from 2006 is obtained from testing at the conclusion of the 2004-2005 school year. Students' post-test scores from 2006 are the same scores as the pre-test scores from 2007. This pattern continues through 2008; therefore, the data contains four years of student achievement scores.

Student identification codes are not recorded in the student files. To account for students, the files contain teacher identification codes which are repeated with each observation representing a single student. Teachers are linked with their students' achievement scores unlike the methods presently employed by the Alabama State Department of Education. The nature of the data, however, precludes the ability to track students over time and create additional testing histories as discussed in Section 4.2.2. Therefore, all four Phase I metrics were calculated within a single two year testing history for each year as shown in Table 4.4.

Table 4.4: NCEES Testing Histories

Year	2006	2007	2008
Variables			
Mathematics and Reading Test Scores	2005, 2006	2006, 2007	2007, 2008

"Each district was assigned to one of the two providers of treatment services, either Educational Testing Service (ETS) or New Teacher Center (NTC), based primarily on district preferences" (Glazerman et al., 2010, p. 8). Districts then received either one or two years of treatment services determined, principally, by the availability of mentors for the second year (Glazerman et al., 2010, p. 3). This research's approach was to analyze districts individually by grade. As a result one does not encounter issues related to comparing teachers who may be

receiving similar induction services but from different providers and for different lengths of time. Furthermore, “the preference-based method of assigning districts to providers does not allow for and should not be used to make direct comparisons of one provider to the other. Observed differences in impacts between ETS and NTC districts may be due to the programs or to the set of districts each provider worked with; those effects cannot be separated” (Glazerman et al., 2010, p. 8).

Analysis within a district did not concern itself with treatment status. All district teachers (treatment and control) by grade were compared with each other. However, the knowledge of teachers’ treatment status precipitated answering the following research question related to the RMTEI results that will be addressed in Section 4.8:

Are the RMTEI results related to whether teachers received induction services (treatment vs. control) while considering length of induction? Glazerman et al. (2010) did observe a statistically significant difference in student achievement with teachers who received two years of induction services (2010, p. 92).

4.3.1 NCES Student Achievement Data Recorded as Z-Scores

Student achievement data from the NCES are recorded as z-scores, mean subtracted from the test score divided by the standard deviation. Population mean and standard deviation of a particular grade, subject, and test were approximated by a state or national norm sample depending on the administered test (Glazerman et al., 2010, p. A–13). A single z-score represents the percentage of the standard deviation away from the mean. The natural question then becomes, can a student’s growth be measured with z-scores? In fact, a student’s growth can be measured with z-scores. For example, a teacher can move a student to a better position by either obtaining a smaller percentage of the standard deviation away from the mean if the student’s test score is less than the mean or obtaining a larger percentage of the standard

deviation away from the mean if the test score is greater than the mean. The difference of the pre-test and post-test z-scores was captured in the new variables, ReadGain and MathGain, to allow analysis to focus on the current teacher who provided instruction leading to the administration of the post-test . This difference represents the movement of a student “relative to their local reference group” (Glazerman et al., 2010, p. A–14). A larger, positive movement compared to other students would be considered more growth.

In order to ensure the adequacy of using z-scores with the NCES data, the teacher values obtained with the RMTEI process using scaled scores from the Alabama district were compared to the teacher values obtained with the RMTEI process after converting the scaled scores to z-scores. The z-scores were computed using the mean and standard deviation of the scaled scores by grade and year. The RMTEI results using both types of scores were nearly identical, perfectly correlated, and not different with statistical significance. The results confirm the ability to safely complete the RMTEI process with z-scores for the NCES data.

4.3.2 Student, Teacher, and School Level Predictors for NCES Data

As with the Alabama district data, this research examined predictors with the NCES data in an attempt to improve the predictive capability of the models used to generate the four Phase I metrics. The process of selecting predictors began with using a student’s pre-test in mathematics and reading to predict MathGain and ReadGain. For example, the 2006 models started with the Pre-Read and Pre-Math scores from the spring of 2005 as predictors for ReadGain and MathGain in separate models. This method of analysis produced similar results for all grades, districts, and years in which the pre-test scores were significant at an $\alpha = .05$ for both ReadGain and MathGain models. Therefore, Pre-Read and Pre-Math scores were included as predictors in all models for 2006, 2007, and 2008.

Adding predictors at the teacher level led to inconsistent results. NCES teacher data included many variables that received evaluation as predictors for student reading and mathematics growth:

- 1) Route into teaching
- 2) Highest degree
- 3) Holds a degree in an education–related field
- 4) Hired after the school year began
- 5) Attended a competitive college¹
- 6) Not a beginning teacher
- 7) Held a nonteaching job for five or more years
- 8) Type of teaching certificate/licensure/credential currently held
- 9) How teacher entered the teaching profession

In most instances teacher level predictors were not statistically significant. Therefore, teacher level predictors were not included in models for 2006, 2007, and 2008.

Adding predictors at the school level also led to unreliable results. This research applied the methods of the Texas Projection Measure, discussed in Chapter 2, by considering campus mean of the predicted subject as a predictor of student achievement. In many instances, MathMean and ReadMean for a given campus were not significant predictors of MathGain and ReadGain respectively. Secondly, the NCES data contained the Reduced/Free lunch percentages of the different campuses within the districts and considered them as predictors of student achievement. Obtaining results similar to using the campus mean of the predicted subject as a

¹Glazer et al. (2010) provide the definition: “A “highly selective” college or university is one that is rated as “most competitive,” “highly competitive,” or “very competitive” by the 2003 edition of the *Barron’s Profile of American Colleges*.” (2010, p. 13).

predictor, schools' Reduced/Free lunch percentages were not significant predictors of student achievement. Therefore, school level predictors were not included in models for 2006, 2007, and 2008.

An example follows for the solution of fixed effects from a district, 4th grade, 2006, where the predictors, COMPCOLLEGE² and EDFIELD, are categorical variables for whether the math teacher attended a competitive college and holds a degree in an education-related field respectively. Also included are student, teacher, and school level predictors as described above. All of the teacher and school level predictors for the minimum testing history group failed to reject the null hypotheses, $H_0 : \beta_3 = 0, H_0 : \beta_4 = 0, H_0 : \beta_5 = 0, H_0 : \beta_6 = 0, H_0 : \beta_7 = 0$ at an $\alpha = .05$ significance level:

Solution for Fixed Effects								
Effect	highest_degree	COMPCOLLEGE. Attended Most, Highly, or Very competitive college	EDFIELD. Has a degree in education related field	Estimate	Standard Error	DF	t Value	Pr > t
Intercept				0.3388	0.3005	5	1.13	0.3107
PRE_MATH				-0.5598	0.05764	157	-9.71	<.0001
PRE_READ				0.3270	0.07742	157	4.22	<.0001
MathMean				-0.01909	0.2165	157	-0.09	0.9298
percent_free_reduced				-0.00534	0.004399	157	-1.21	0.2265
COMPCOLLEGE		0 = No		0.2000	0.1257	157	1.59	0.1136
COMPCOLLEGE		1 = Yes		0
EDFIELD			0 = No	-0.04655	0.09863	157	-0.47	0.6376
EDFIELD			1 = Yes	0
highest_degree	B			0

The result of examining model predictors for the NCES data also led to an approach similar to the Chapter 3 example in which only student-level predictors were included in the models. Table 4.5 summarizes the outcome with each check mark representing the inclusion of the independent variable in the corresponding model for each Phase I metric.

² Glazer et al. (2010) provide the definition: “A “highly selective” college or university is one that is rated as “most competitive,” “highly competitive,” or “very competitive” by the 2003 edition of the *Barron’s Profile of American Colleges*.” (2010, p. 13).

Table 4.5: Summary of Model Predictors for NCES Data

Metric	LMM Teacher Effect		Overall LMM Value-Added Measure		Median SGP		Overall QR Value-Added Measure	
	MathGain	ReadGain	MathGain	ReadGain	MathGain	ReadGain	MathGain	Readgain
Pre-Math (previous year score)	✓	✓	✓	✓	✓	✓	✓	✓
Pre-Read (previous year score)	✓	✓	✓	✓	✓	✓	✓	✓
Teacher-Level Predictors								
School-Level Predictors								

4.4 Checking Model Assumptions for the Final Linear Mixed Models

Analysis of the final models began by examining the EBLUPs for the random teacher effects. The values yielded consistent results of failing to reject normality with no outliers. An example is shown in Figure 4.1 for the Alabama mathematics teacher effects for 4th grade, 2009.

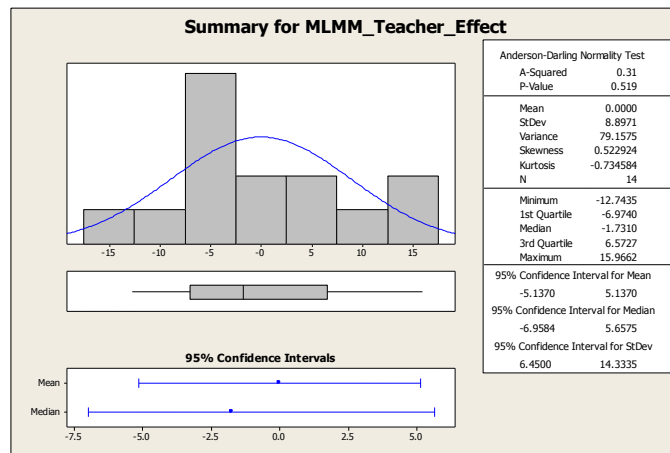


Figure 4.1: Distribution Analysis of Mathematics LMM Teacher Effects

Secondly, analysis of the conditional studentized residuals validated the model assumptions of normality and constant variance. “The conditional studentized residual for an observation is the difference between the observed value and the predicted value, based on both

the fixed and random effects in the model, divided by its estimated standard error” (West et al., 2007, p. 104). An example is shown in Figure 4.2 from the MathGain model using Alabama data from the cohort preceding 4th grade, 2009.

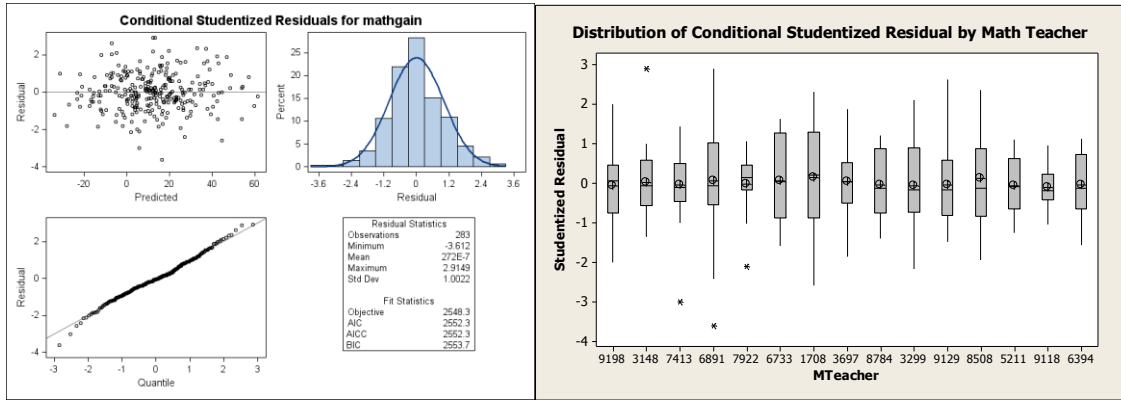


Figure 4.2: Distribution Analysis of Studentized Residuals from Mathematics LMM

Lastly, a paired comparison of MathGain and its conditional predicted value yielded inferences about their difference in means. The 95% confidence interval for the difference in means contains zero. Therefore, one would fail to reject the null hypotheses: $H_0 : \mu_i - \mu_j = 0$. The difference in means of the two variables is not significant. An example is shown in Figure 4.3 from the MathGain model using Alabama data from the cohort preceding 4th grade, 2009.

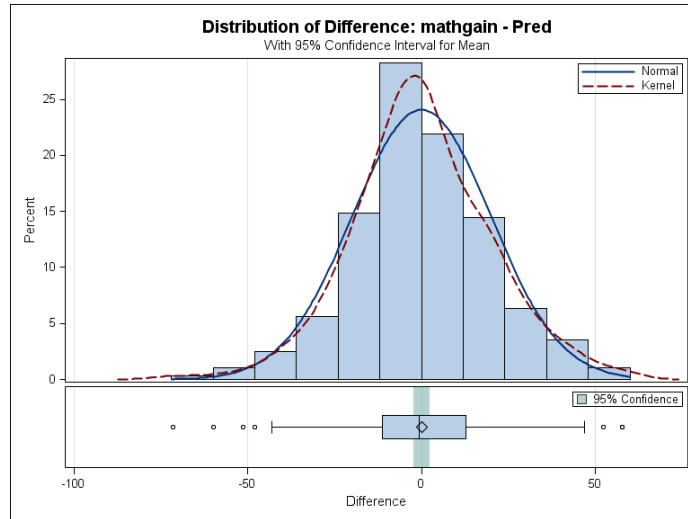


Figure 4.3: Paired Comparison of MathGain and Conditional Predicted Value

The paired comparison of MathGain and its conditional predicted value also yielded inferences about their difference in means for each teacher. The 95% confidence intervals for the differences in means for each teacher contain zero; therefore, the differences in means are not significant.

4.5 Diagnostics for the Final Quantile Regression Models

Analysis of the final model for the QR Value-Added Measure metric began by examining some diagnostic plots from the predicted 0.5 quantile of MathGain and ReadGain, “which expresses the conditional median of a response variable given predictor variables” (Hao & Naiman, 2007, p. 56). Recall that the median can be a more suitable measure of central location than the conditional mean if the conditional mean’s model suffers from inadequacy due to assumption violations.

Although not unexpected, the plot of predicted values versus standardized residuals in Figure 4.4 produces a double bow effect indicating the variance of the errors is not constant

(West et al., 2007, p. 131). This is also seen in Figure 4.4 when the distribution of standardized residuals is not homogeneous across mathematics teachers (West et al., 2007, p. 106). Both graphs in Figure 4.4 originate from analysis of the medium-sized Alabama district for 4th grade, 2009.

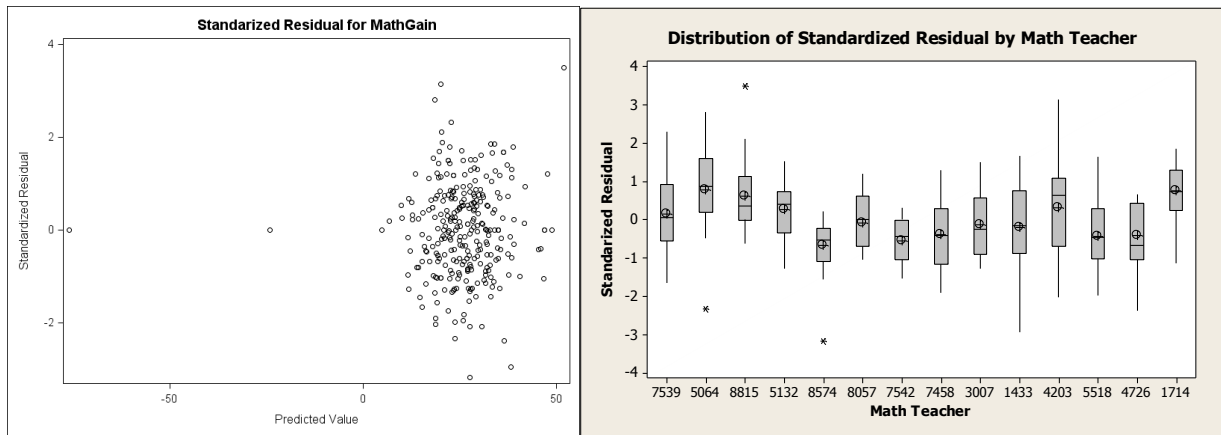


Figure 4.4: Variance Analysis of Standardized Residuals from 0.5 QR Mathematics Model

One would expect the points to be symmetrically distributed around a diagonal line in a plot of observed versus predicted values for a Linear Mixed Model. Since the median is often used to indicate the center of skewed distributions, it is not unexpected for the symmetry to be absent in Figure 4.5 for the plot of MathGain versus the 0.5 quantile prediction of MathGain for 4th grade, 2009, in the Alabama district (Hao & Naiman, 2007, pp. 12–13).

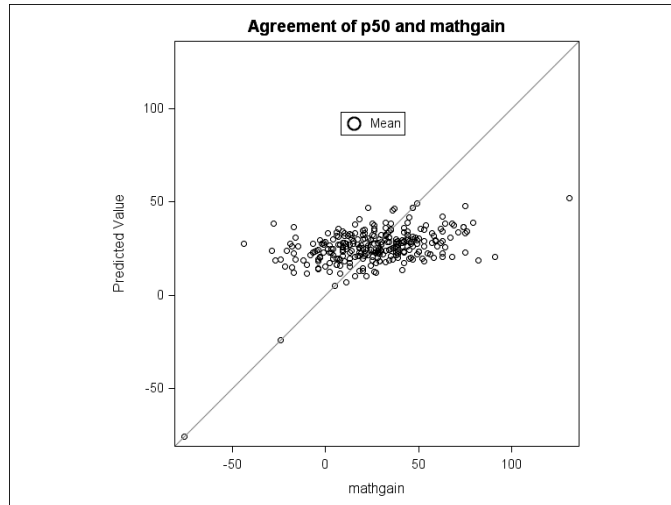


Figure 4.5: Agreement Plot of MathGain versus 0.5 QR Prediction of MathGain

The plot in Figure 4.6, however, illustrates the normal nature of the standardized residuals from the 0.5 quantile prediction of MathGain for 4th grade, 2009, in the Alabama district.

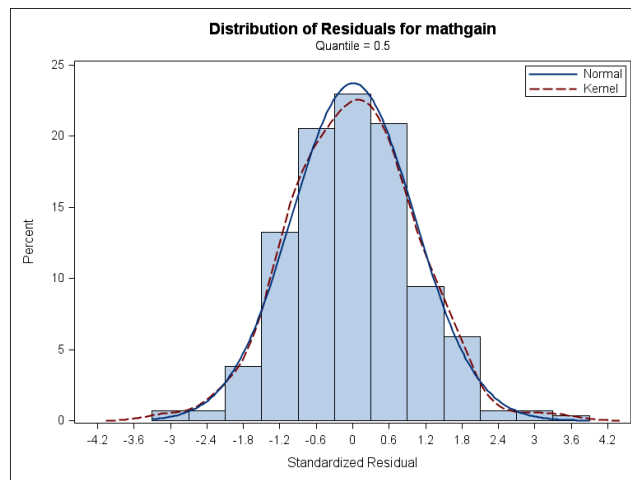


Figure 4.6: Distribution Analysis of Standardized Residuals from 0.5 QR Mathematics Model

4.6 Principal Component Analysis (PCA)

After the applying the methods of the previous sections, Phase I concluded with calculating the four teacher effectiveness metrics: Linear Mixed Model Teacher Effect, Overall Linear Mixed Model Value-Added Measure, Median Student Growth Percentile, and Overall

Quantile Regression Value-Added Measure; see Appendix 3 for the Alabama district’s results and Appendix 4 for a district from the NCES dataset. Phase II consists of determining the principal components of the four metrics determined in Phase I. Recall from Chapter 2 that the correlation matrix of the four metrics is used to extract the eigenvalue-eigenvector pairs, $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$, due to the dissimilar units of measurement for the Phase I metrics.

The high correlation between metrics and the ability of PCA to reduce dimensionality efficiently often led to one dominant principal component. For the Alabama district in a given grade and year, the Phase I metrics with MathGain as the dependent variable produced a single, large eigenvalue and principal component that absorbed the preponderance of the variance in the system. ReadGain, however, would clearly have a dominant eigenvalue but without the near total absorption of the variance. Table 4.6 summarizes the results for the Alabama district.

Table 4.6: Proportion of the Total Variance for First Principal Component in the Alabama District

Mathematics:							
Proportion of the Total Variance for 1st Principal Component							
Year\Grade	4	5	6	7	8	Average	
2009	0.9832	0.9853	0.8587	0.9967	0.6659	0.8980	
2010	0.9710	0.9981	0.9313	0.9797	0.9379	0.9636	
2011	0.9445	0.9982	0.9994	0.9959	0.9284	0.9733	
Average	0.9662	0.9938	0.9298	0.9908	0.8441		
						Overall	0.9449
Reading:							
Proportion of the Total Variance for 1st Principal Component							
Year\Grade	4	5	6	7	8	Average	
2009	0.8043	0.9637	0.9703	0.8294	0.7884	0.8712	
2010	0.9723	0.7401	0.9781	0.5540	0.9232	0.8335	
2011	0.9497	0.8577	0.9624	0.6651	0.7661	0.8402	
Average	0.9088	0.8538	0.9703	0.6829	0.8259		
						Overall	0.8483

Similarly, a district in the NCES dataset produced the results in Table 4.7.

Table 4.7: Proportion of the Total Variance for First Principal Component in a NCES District

Mathematics:						
Proportion of the Total Variance for 1st Principal Component						
Year/Grade	3	4	5	Average		
2006	0.9522	0.9428	0.9616	0.9522		
2007	0.9445	0.9293	1.1569	1.0102		
2008	0.9448	0.8142	1.1963	0.9851		
Average	0.9472	0.8954	1.1049			
					Overall	0.9825
Reading:						
Proportion of the Total Variance for 1st Principal Component						
Year/Grade	3	4	5	Average		
2006	0.9587	0.9714	0.9654	0.9652		
2007	0.9254	0.8944	0.9484	0.9228		
2008	0.9544	0.8358	0.8854	0.8919		
Average	0.9461	0.9005	0.9331			
					Overall	0.9266

Principal component analysis allowed this research to project the four metrics into a single dimension to obtain a teacher effectiveness index by subject. If a teacher taught mathematics and reading, then an overall teacher effectiveness index was obtained by taking the mean of the two subject indexes. Otherwise, the single subject index became the teacher’s overall teacher effectiveness index.

In addition to reducing the dimensionality of the Phase I metrics, PCA discovered the relationship between the effectiveness metrics. In every instance the first principal component was a weighted sum of all Phase I metrics. Based on the grade, year, and subject, the weights (coefficients) for the metrics were different; however, the prevailing theme was that each metric was highly correlated with the first principal component and contributed nearly the same proportion (compared to the other metrics) of its value to the principal component score.

Lastly, PCA provided the inputs for Phase III, Cluster Analysis, in which clustering took place based on principal component scores in a single subject. Clustering was also performed

based on overall teacher effectiveness index values. An example of the input for Cluster Analysis is contained in Table 4.8.

Table 4.8: Results from PCA for 4th Grade, 2011, in the Alabama District

Teacher	Mathematics		Reading		Overall Index Value
	prin1	prin2	prin1	prin2	
5064	0.7205	0.5362	-0.0079	-0.2183	0.3563
1714	4.2248	0.0932	5.1338	-0.0124	4.6793
8815	0.8364	0.2912	0.0527	-0.2900	0.4446
4203	0.6840	0.2958	-2.2187	-0.2888	-0.7673
5132	2.2178	-0.5271	-0.1321	-0.6740	1.0428
7539	0.7081	0.3148	2.0895	0.2618	1.3988
8057	-0.0737	-0.0172	0.5666	0.1816	0.2464
3007	1.4536	0.1150	0.8888	-0.1911	1.1712
1433	-2.2769	0.6641	-1.6085	-0.2387	-1.9427
7458	-0.7233	-1.0158	0.6358	0.5562	-0.0438
5518	-1.7678	0.2399	-3.0689	0.3594	-2.4183
4726	-3.0736	0.1825	-0.5650	0.0638	-1.8193
7542	0.3562	-0.3420	-0.1776	0.8245	0.0893
8574	-2.6457	-0.4508	-1.8866	0.2219	-2.2661
5629	-0.6406	-0.3798	0.2980	-0.5560	-0.1713

4.7 Cluster Analysis

The purpose of Phase III, Cluster Analysis, is to accomplish two of this research’s sub-objectives methods:

- 1) Confirm or modify Alabama’s effectiveness rating categories based on being able to detect statistically different groups of teachers by effectiveness.
- 2) Report all teachers in accordance with the ratings categories for an Alabama school district using the newly developed objective effectiveness index.

As described in Section 3.4, the agglomerative hierarchical clustering method of J.H.

Ward produced clusters based on the first two principal component scores (simultaneously) in a single subject. The results produced statistically different clusters when considering the cluster means of the first principal component score in a single subject. Clustering was also performed based on overall teacher effectiveness values, and the results produced statistically different

clusters when considering the cluster means of the overall teacher effectiveness index value. The analysis of the medium-sized Alabama district could not replicate the result of Chapter 3, however, by producing five statistically different clusters within a specific grade, subject, and year. This can be attributed to the size of the district with the number of teachers per grade being less than the Chapter 3 example. In many instances grades 5-8 would only have three teachers for a single subject. In these situations, statistically different clusters did not provide useful information. Clustering based on the overall teacher effectiveness index values always produced greater statistically different clusters since the pool of teachers was greater (both mathematics and reading teachers are included). The number never exceeded four statistically different clusters, however. The graphs of Figure 4.7 capture the discussion above for the Alabama district for all three years.

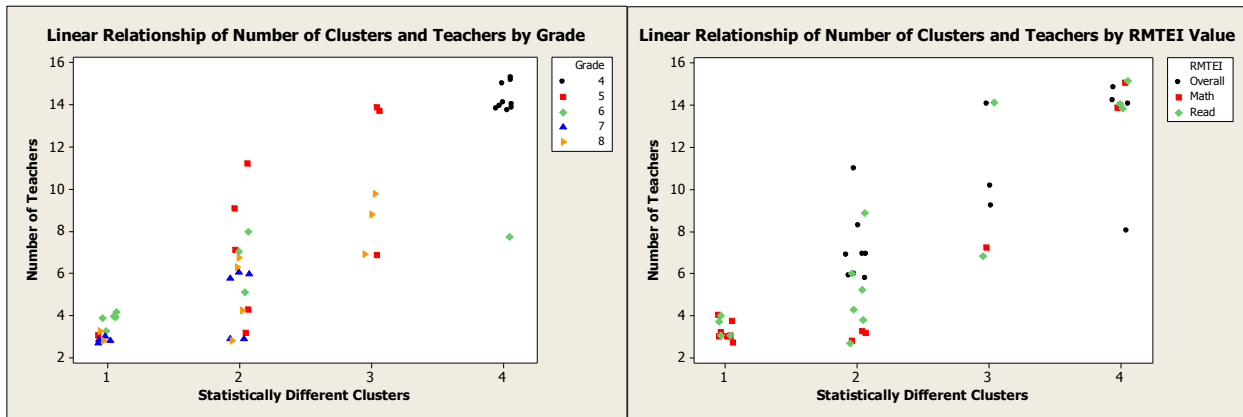


Figure 4.7: Linear Relationship of the Number of Clusters and Teachers for the Alabama District

An example follows of the Bonferroni confidence intervals for the difference of cluster means (Overall Index Value) for 4th Grade, 2011, in the Alabama school district:

```

Bonferroni (Dunn) t Tests for Overall Index Value
Alpha                                0.05
Error Degrees of Freedom              11
Error Mean Square                     0.11881
Critical Value of t                   3.20812

```

Comparisons significant at the 0.05 level are indicated by ***.

CLUSTER Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
4 - 3	3.4750	2.1982	4.7519	***
4 - 1	4.6573	3.4751	5.8394	***
4 - 2	6.7909	5.5546	8.0273	***
3 - 4	-3.4750	-4.7519	-2.1982	***
3 - 1	1.1822	0.4192	1.9453	***
3 - 2	3.3159	2.4713	4.1605	***
1 - 4	-4.6573	-5.8394	-3.4751	***
1 - 3	-1.1822	-1.9453	-0.4192	***
1 - 2	2.1337	1.4406	2.8268	***
2 - 4	-6.7909	-8.0273	-5.5546	***
2 - 3	-3.3159	-4.1605	-2.4713	***
2 - 1	-2.1337	-2.8268	-1.4406	***

Computing 95% Bonferroni simultaneous confidence intervals validates that each cluster is statistically different than every other cluster for the Overall Index Value. For example, Cluster 1 is statistically different than Clusters 2, 3, and 4. Cluster 2 is statistically different than Clusters 3 and 4. Lastly, Cluster 3 is statistically different than Cluster 4. The result of placing teachers in effectiveness categories using their overall effectiveness index values is shown in Figure 4.8 for 4th Grade, 2011, in the Alabama school district.

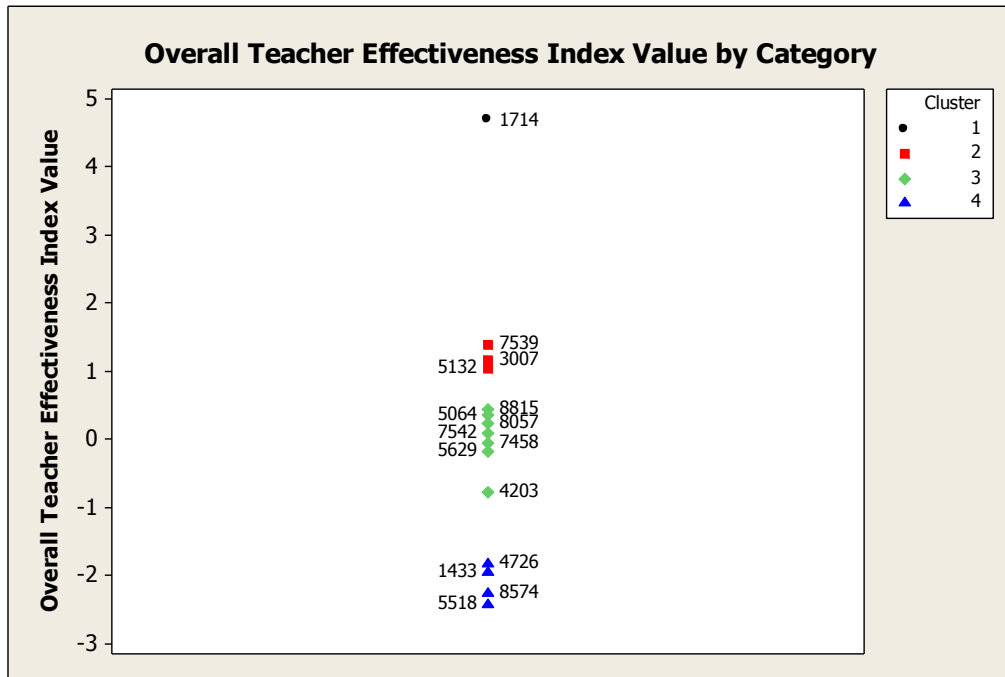


Figure 4.8: 4th Grade, 2011, Effectiveness Categories for the Alabama District³

³ Cluster values have been modified to ensure Cluster 1 represents extraordinary gains in student growth, ..., and Cluster 4 represents poor gains.

Due to the nature of the NCES study, districts were analyzed separately with only beginning teachers included in the study. In several instances districts did not have any or few beginning teachers of a specific grade, subject, and year. Similar to the medium-sized Alabama district, the analysis of the NCES data could not replicate the result of Chapter 3 by producing five statistically different clusters within a specific grade, subject, and year. The graphs of Figure 4.9 for a district in the NCES dataset also capture the linear relationship of the number of clusters and teachers for all three years.

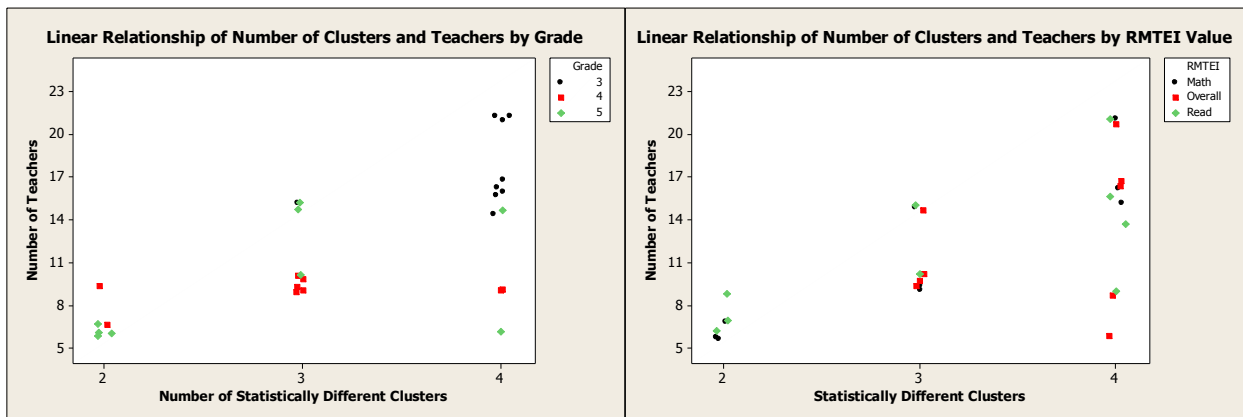


Figure 4.9: Linear Relationship of Number of Clusters and Teachers for a NCES District

4.7.1 Comparison of Clustering Results with Phase I Metrics

Ward’s clustering method produced the results in Figure 4.10 and Figure 4.11 in which clustering was performed for each of Phase I metrics and the input from PCA (principal components 1 and 2).

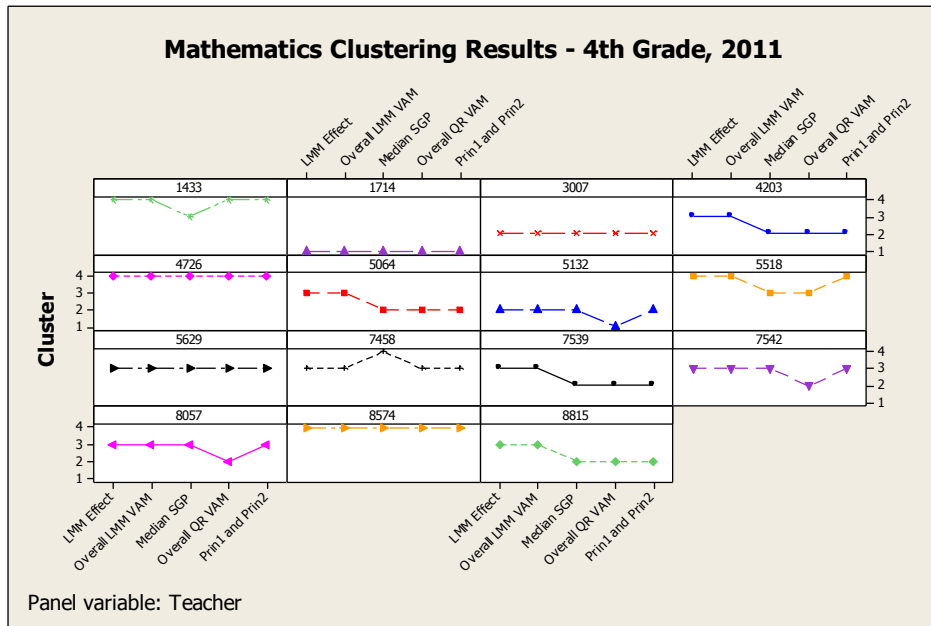


Figure 4.10: Mathematics Clustering Results for 4th Grade, 2011, in the Alabama District⁴

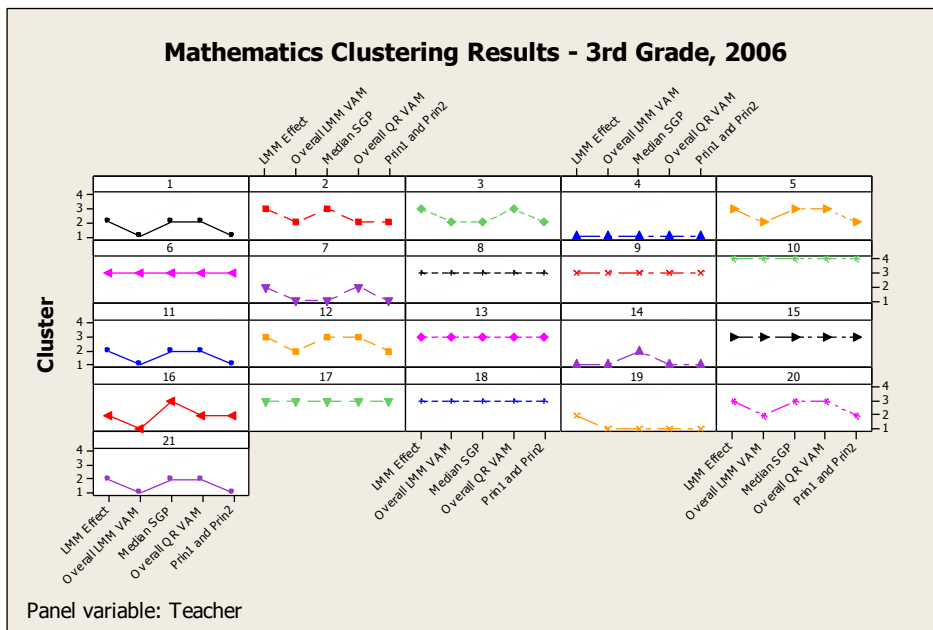


Figure 4.11: Mathematics Clustering Results for 3rd Grade, 2006, in a NCES District⁵

⁴ Values have been modified to ensure clusters across the variables are on the same scale (i.e., Cluster 1 represents extraordinary gains in student growth, ..., and Cluster 4 represents poor gains).

⁵ Values have been modified to ensure clusters across the variables are on the same scale (i.e., Cluster 1 represents extraordinary gains in student growth, ..., and Cluster 4 represents poor gains).

Once again a question must be asked; do the clusters provided by principal components 1 and 2 provide better information than the clusters formed using only a single Phase I metric? The clusters remain similar across the variables, and the clustering results using principal component 1 and 2 successfully represent the collection of the metrics' results with no one metric being identical to the clustering results using the input from PCA. Therefore, the final clustering results provide better information, are not based on a single measure, and have mitigated the risk associated with crafting accurate effectiveness categories.

4.8 Subject-Specific and Overall RMTEI Values

Five years of student and teacher data from a medium-sized Alabama district produced subject-specific and overall Risk-Mitigated Teacher Effectiveness Index values for grades 4-8 for the 2008-09, 2009-10, and 2010-11 school years; see Appendix 3 for complete results. Four years of student and teacher data from the NCES yielded subject-specific and overall Risk-Mitigated Teacher Effectiveness Index values for grades 3-5 for the 2005-06, 2006-07, 2007-08 school years; see Appendix 4 for complete results from a district in the NCES dataset. An excerpt of the final results from the medium-sized Alabama district is shown in Table 4.9 with each 4th grade teacher receiving a sparkline, “small, high-resolution graphics embedded in a context of words, numbers, images” to capture the index values over time (Tufte, 2006, p. 7).

Table 4.9: Excerpt of Final Results Sorted by the 2009 Mathematics RMTEI value

teacher	school	grade2011	Mathematics			MSparkline	Reading			RSparkline	Overall Index Value			IVSparkline
			2009	2010	2011		2009	2010	2011		2009	2010	2011	
5064	25	4	3.3457	-0.4346	0.7205		0.5328	-1.0480	-0.0079		1.9393	-0.7413	0.3563	
1714	10	4	3.0823	3.1655	4.2248		0.7264	3.2447	5.1338		1.9044	3.2051	4.6793	
8815	25	4	2.1481	1.8351	0.8364		0.1127	1.4932	0.0527		1.1304	1.6641	0.4446	
4203	20	4	1.6122	-2.0304	0.6840		-1.0756	-2.3744	-2.2187		0.2683	-2.2024	-0.7673	
5132	10	4	1.1003	4.0786	2.2178		1.2701	2.7546	-0.1321		1.1852	3.4166	1.0428	
7539	10	4	0.3079	1.3980	0.7081		-2.1642	1.6205	2.0895		-0.9281	1.5093	1.3988	
8057	20	4	-0.1194	-0.6315	-0.0737		-0.5729	-3.1174	0.5666		-0.3461	-1.8745	0.2464	
3007	15	4	-0.6870	-0.8924	1.4536		1.2367	-1.0908	0.8888		0.2749	-0.9916	1.1712	
1433	25	4	-0.9020	-0.0333	-2.2769		-0.4659	-0.4690	-1.6085		-0.6839	-0.2511	-1.9427	
7458	20	4	-1.4394	-1.6230	-0.7233		-2.3517	0.1714	0.6358		-1.8955	-0.7258	-0.0438	
5518	15	4	-1.6379	-1.5135	-1.7678		-1.6645	-2.4991	-3.0689		-1.6512	-2.0063	-2.4183	
4726	25	4	-1.7912	-2.2494	-3.0736		2.5634	1.1237	-0.5650		0.3861	-0.5628	-1.8193	
7542	15	4	-2.3511	0.5721	0.3562		2.4138	1.1369	-0.1776		0.0313	0.8545	0.0893	
8574	15	4	-2.6685	-1.6412	-2.6457		-0.5611	-0.9464	-1.8866		-1.6148	-1.2938	-2.2661	
5629	25	4			-0.6406				0.2980					

The ability to detect consistent teachers, positively or negatively, relative to their peers is readily achieved with the sparklines. A positive index value places a teacher in the top half of the population of teachers in terms of attaining student growth. The graph in Figure 4.12 allows the comparison of 4th grade mathematics teachers over the same three year span.

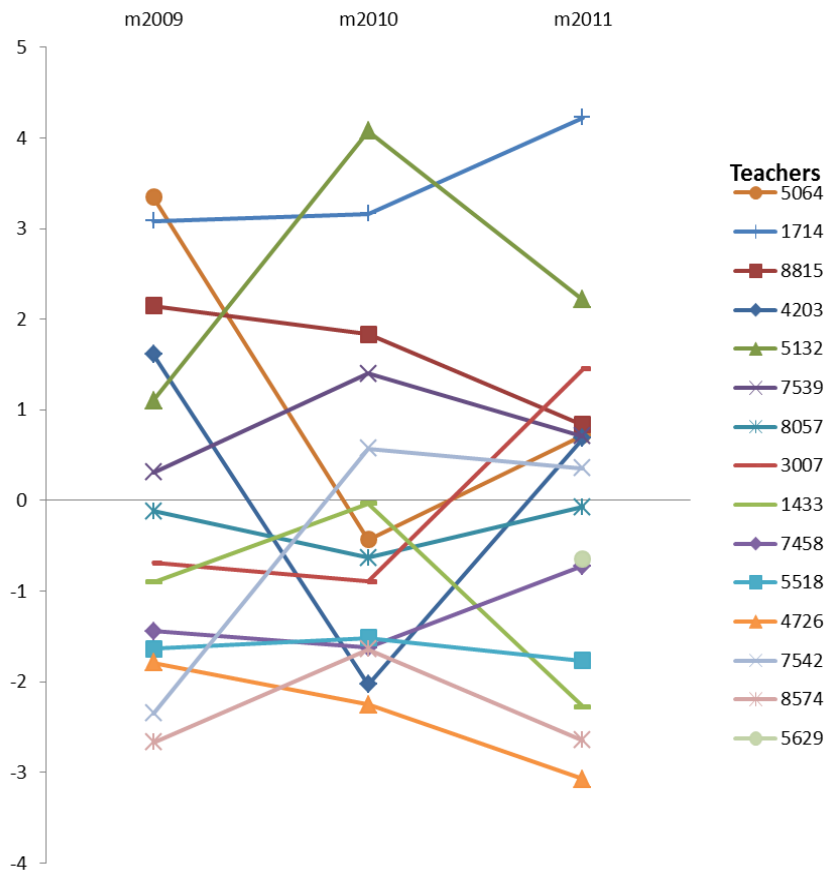


Figure 4.12: Final 4th Grade Mathematics RMTEI values

Based on Figure 4.12 the preliminary indication is that the index does not suffer from stability concerns from year to year due to the majority of teachers having modest movement of their mathematics-specific index value over time. Nine of 14 teachers would fall into this category, which does not include teacher 5629 with a single value in 2011. This notion as well as the precision of the index will be explored fully, however, in Sections 4.9 and 4.10.

Lastly, NCES analysis was undertaken to determine whether there is a statistical difference between induction groups (control vs. treatment) while considering their RMTEI values and length of induction. The analysis required merging district RMTEI results by years of induction services (one or two years) for the districts’ treatment group. Recall, Glazerman et al.

(2010) did observe a statistically significant difference in student achievement with teachers who received two years of induction services (2010, p. 92).

The grouping of ten (10) one-year districts led only to a significance of Treatment Status (one year of comprehensive induction versus the existing, less intensive training) for mathematics teachers in the third year (p-value=.0245). Reading and Overall RMTEI values failed to reject $H_0 : \mu_1 = \mu_2$ at an $\alpha = 0.05$ level for 2006, 2007, and 2008. The grouping of seven (7) two-year districts led only to a significance of Treatment Status (two years of comprehensive induction versus the existing, less intensive training) for mathematics teachers in the third year (p-value=.0084). Reading and Overall RMTEI values failed to reject $H_0 : \mu_1 = \mu_2$ at an $\alpha = 0.05$ level for 2006, 2007, and 2008. Considering the analysis above, this research concludes in the third year that there is a statistical difference between treatment and control teachers' performance while considering their RMTEI mathematics values for both one-year and two-year districts. As a result, this research asserts that comprehensive induction services for beginning teachers have a greater impact on student achievement in mathematics compared to reading. An excerpt of the results is contained in Table 4.10.

Table 4.10: Excerpt of Treatment Status Analysis

Year Districts	Dep Var: Mathematics RMTEI Source: Treatment (p-value)			Dep Var: Reading RMTEI Source: Treatment (p-value)			Dep Var: Overall RMTEI Source: Treatment (p-value)		
	2006	2007	2008	2006	2007	2008	2006	2007	2008
One-year districts	0.3643	0.2989	0.0245	0.8439	0.1091	0.8343	0.5240	0.1446	0.1315
Two-year districts	0.1399	0.9192	0.0084	0.8471	0.6842	0.8470	0.4822	0.7763	0.0631

4.9 Precision of the Risk-Mitigated Teacher Effectiveness Index

As stated in Section 2.4 small sample sizes present challenges for measuring teacher effectiveness. Precision of the results still remain an issue, however, for middle school teachers who may teach a greater number of students compared to an elementary school teacher who may teach a single class multiple subjects. Lockwood, Louis, and McCaffrey (2002) investigated the precision of teacher effects from value-added models and found that variation in most scenarios caused “estimated rankings [to] be sufficiently imprecise to preclude distinguishing among all but the most extreme teachers” (McCaffrey et al., 2004, p. 107). Therefore, one of the primary objectives of this research is to overcome this lack of precision when measuring teacher effectiveness.

Since teachers’ Phase I metrics for a given subject, grade, and year have varying units of measurement, the Teacher Effectiveness Index values were obtained from the standardized version of those metrics by calculating the eigenvectors of the metrics’ correlation matrix. The correlation matrix captures the variation of the Phase I metrics. The eigenvectors of the correlation matrix form the four distinct principal components. The variance of the principal component scores from the dominant principal component is the largest eigenvalue associated with the corresponding eigenvector. This eigenvalue absorbed the preponderance of the variation of the system, and the principal component scores from the dominant principal

component became teachers' subject-specific RMTEI values. Section 4.7 demonstrated that the results of cluster analysis consistently produced statistically different clusters when considering the cluster means of the first principal component score in a single subject. Clustering also produced statistically different clusters based on the overall teacher effectiveness index values, thereby allowing teacher effectiveness to be measured in an overall and subject-specific manner.

After compiling the final results for all subject-specific and overall Risk-Mitigated Teacher Effectiveness Index values, analysis was undertaken to ascertain the precision of the RMTEI. Considering the results following a specific year in which teachers taught both mathematics and reading, teachers earned two subject-specific index values and an overall effectiveness index value for that year. An excerpt of the final results is contained in Table 4.9. To form an analogy, the values were treated as automobile part measurements in a manufacturing process where one could determine whether the process was in a state of statistical control through the use of a control chart. With assessments over three years, an individual teacher could consist of a subgroup of three Mathematics, Reading, or Overall RMTEI values. All teachers, acting as subgroups, then contributed to being able to determine whether the RMTEI process was in a state of statistical control. "The purpose of any control chart is to identify occurrences of special causes of variation that come from outside of the usual process" since a stable process only contains variation from common causes (Johnson & Wichern, 2007, p. 239). The analysis produced the control charts in Figure 4.13 and Figure 4.14 for 4th grade and 3rd grade mathematics teachers, respectively.

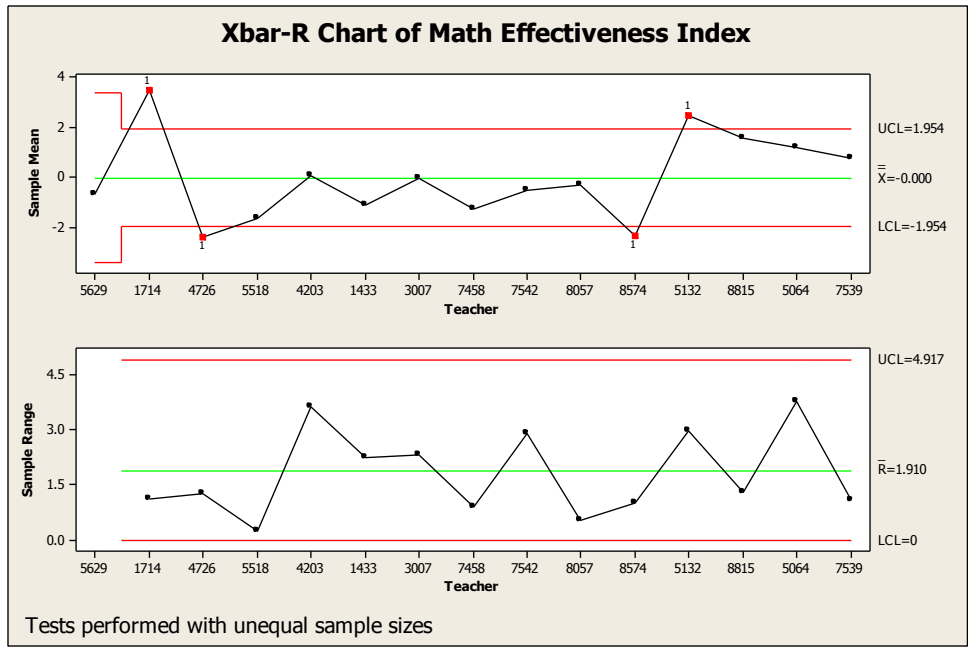


Figure 4.13: Control Chart Analysis for Alabama 4th Grade Mathematics Teachers

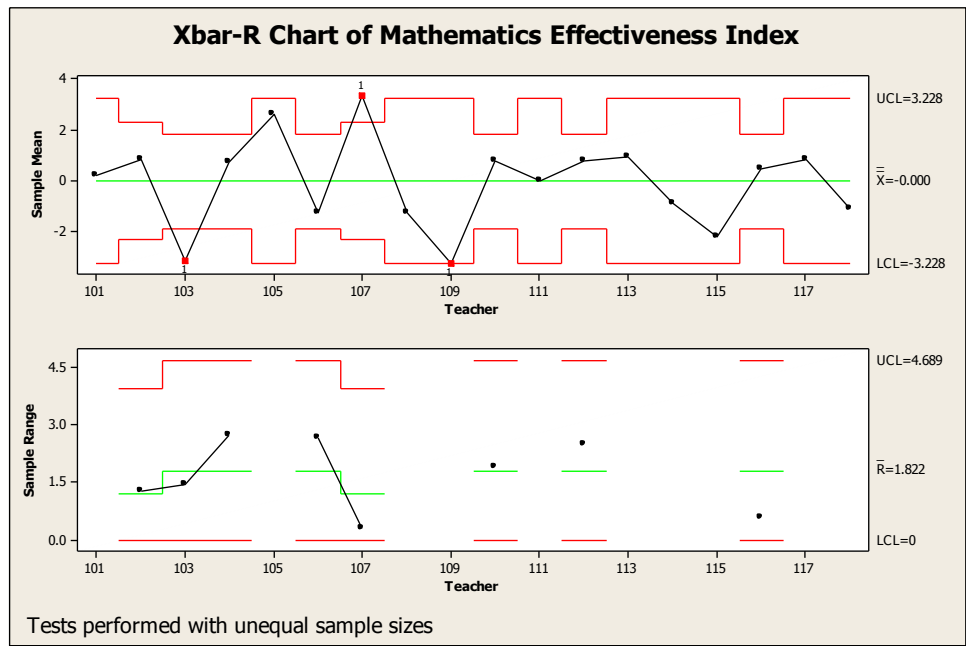


Figure 4.14: Control Chart Analysis for 3rd Grade Mathematics Teachers in a NCES District

The charts in Figure 4.13 and Figure 4.14 of the process mean and process range over the span of the mathematics teachers as subgroups reveal two noteworthy items. Firstly, the R-Chart, which measures within-subgroup variability, remains in control. Secondly, the \bar{x} chart, which measures between-subgroup variability, determines that several teachers fall outside of the control limits. The following paragraph from Trietsch (1999) articulates why the control charts demonstrate the adequacy of the RMTEI process for these grades, subject, and districts:

“As discussed in AT&T (1958), control charts may be used to verify whether a given measurement instrument is adequate for a particular job. The main idea there is that multiple measurements of the same item are used as subgroups, and if the measurement instrument is adequate, many points in the \bar{x} chart should fall outside the control limits. In other words, we expect the variation between parts to be large relative to the range of multiple measurements of the same part” (1999, p. 38).

4.10 Stability of the Risk-Mitigated Teacher Effectiveness Index

As stated in Section 2.4, McCaffrey, Sass, and Lockwood (2009) compared the teacher effectiveness results from two successive cohorts of students in four counties of Florida elementary and middle schools and found low correlations of teacher effectiveness between the two years (Braun et al., 2010, pp. 45–46). The literature suggests that teacher effectiveness measurement instruments lack stability from year to year. Therefore, one of the primary objectives of this research is to overcome this lack of stability and demonstrate high correlation between the subject-specific and overall Risk Mitigated Teacher Effectiveness Index values for the Alabama and NCES datasets in 2009-2011 and 2006-2008, respectively.

The first three years of data, 2007-2009, were used to generate the 2009 subject-specific and overall Risk-Mitigated Teacher Effectiveness Index values for the medium-sized Alabama School district. Years 2010 and 2011 were used to demonstrate the stability and precision of the Effectiveness Index values. Effectiveness index values for grades 4-8 were determined for 2010

and 2011 based on 2007-2010 and 2007-2011 data, respectively. These index values were compared to the index values obtained from 2009 for the same teachers. For example, Mr. Jones taught 7th grade mathematics to cohort 2014 in 2009 and received an effectiveness index. This value can be compared to the effectiveness index obtained following the 2010 school year in which he taught 7th grade mathematics to cohort 2015. Collectively, effectiveness index values were generated for all teachers in grades 4-8 for 2009, 2010, and 2011. Stability analysis is shown in Figure 4.15 and Table 4.11 for mathematics teachers in the Alabama district.

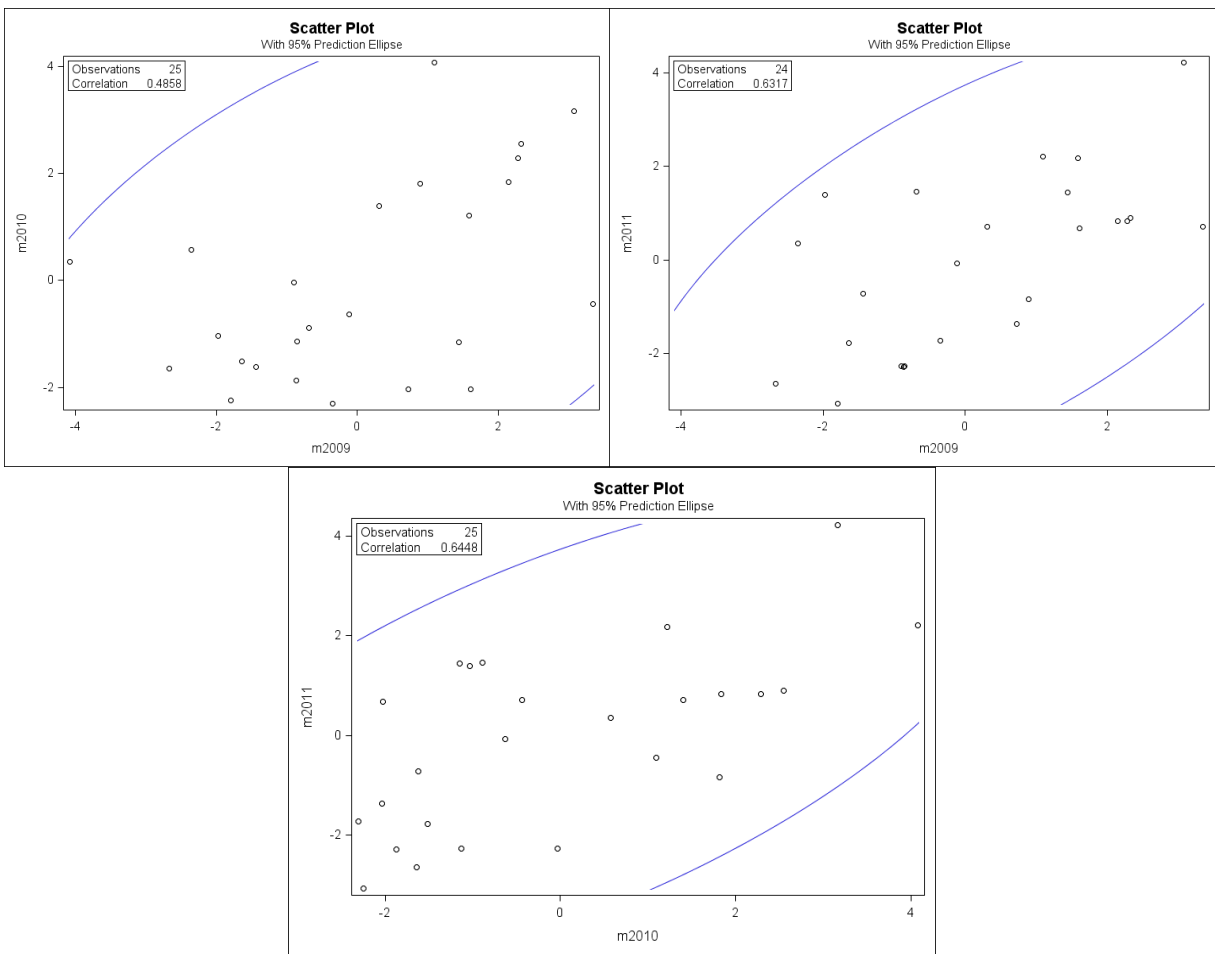


Figure 4.15: Excerpt of Final Stability Analysis for the Alabama District

Table 4.11: Correlation of Yearly RMTEI Math Values for the Alabama District

		Mathematics RMTEI Pearson Correlation Coefficient (p-value)		
	year	2009	2010	2011
Mathematics RMTEI	2009	1.0000	.4858 (.0138)	.6317 (.0009)
	2010		1.0000	.6448 (.0005)
	2011			1.000

As the analysis suggests in Table 4.11, mathematics index values remained stable with statistically significant correlation between 2009 and 2010, 2009 and 2011, and 2010 and 2011. Reading index values remained stable with statistically significant correlation between 2009 and 2010 as well as between 2010 and 2011. Reading index values, however, did not demonstrate statistically significant correlation between 2009 and 2011. The reason for this fact remains elusive although future study as discussed in Chapter 6 may suggest that the reading portion of the ARMT may suffer from faulty test-forms equating. As a result, overall RMTEI values remained stable with statistically significant correlation between 2009 and 2010 as well as between 2010 and 2011. The correlation of overall RMTEI values between 2009 and 2011 was almost statistically significant at an $\alpha = 0.05$ with a p-value = .0507. Table 4.12 summarizes the results for yearly reading and overall RMTEI values.

Table 4.12: Correlation of Yearly Reading, Overall RMTEI Values for the Alabama District

		Reading RMTEI Pearson Correlation Coefficient (p-value)			Overall RMTEI Pearson Correlation Coefficient (p-value)		
	year	2009	2010	2011	2009	2010	2011
Reading RMTEI	2009	1.0000	.3853 (.0268)	.0645 (.7491)			
	2010		1.0000	.3769 (.0481)			
	2011			1.0000			
Overall RMTEI	2009				1.0000	.4344 (.0036)	.3236 (.0507)
	2010					1.0000	.4127 (.0081)
	2011						1.0000

In addition to testing whether correlation values differed from zero with a null hypothesis of $H_0 : \rho = 0$ (depicted in Table 4.11 and Table 4.12), this research was interested in using a Fisher’s z-transformation to test differences between the correlations. For example, is there a statistical difference between the correlation of 2009-2010 and 2009-2011 for Mathematics RMTEI values? A desirable feature of the RMTEI process would be no statistical significance in the difference of correlation values. Conducting a Fisher’s z-transformation of the correlation values with the corresponding hypotheses,

$H_0 : \rho_{2009,2010} - \rho_{2009,2011} = 0, H_0 : \rho_{2009,2010} - \rho_{2010,2011} = 0, H_0 : \rho_{2009,2011} - \rho_{2010,2011} = 0$, demonstrated no statistical difference between the correlation values in each set of Mathematics, Reading, and Overall RMTEI values for the Alabama district.

The first two years of the NCES data, 2005-2006, were used to generate the 2006 subject-specific and overall Risk-Mitigated Teacher Effectiveness Index values for 17 school districts. Years 2007 and 2008 were used to demonstrate the stability and precision of the Effectiveness Index values. Effectiveness Index values for grades 3-5 were determined for 2007 and 2008 based on 2006-2007 and 2007-2008 data respectively. These index values were compared to the

index values obtained from 2006 for the same teachers. Collectively, Effectiveness Index values were generated for all teachers in grades 3-5 for 2006, 2007, and 2008. Stability analysis is shown in Figure 4.16 and Table 4.13 for mathematics teachers from the NCES districts.

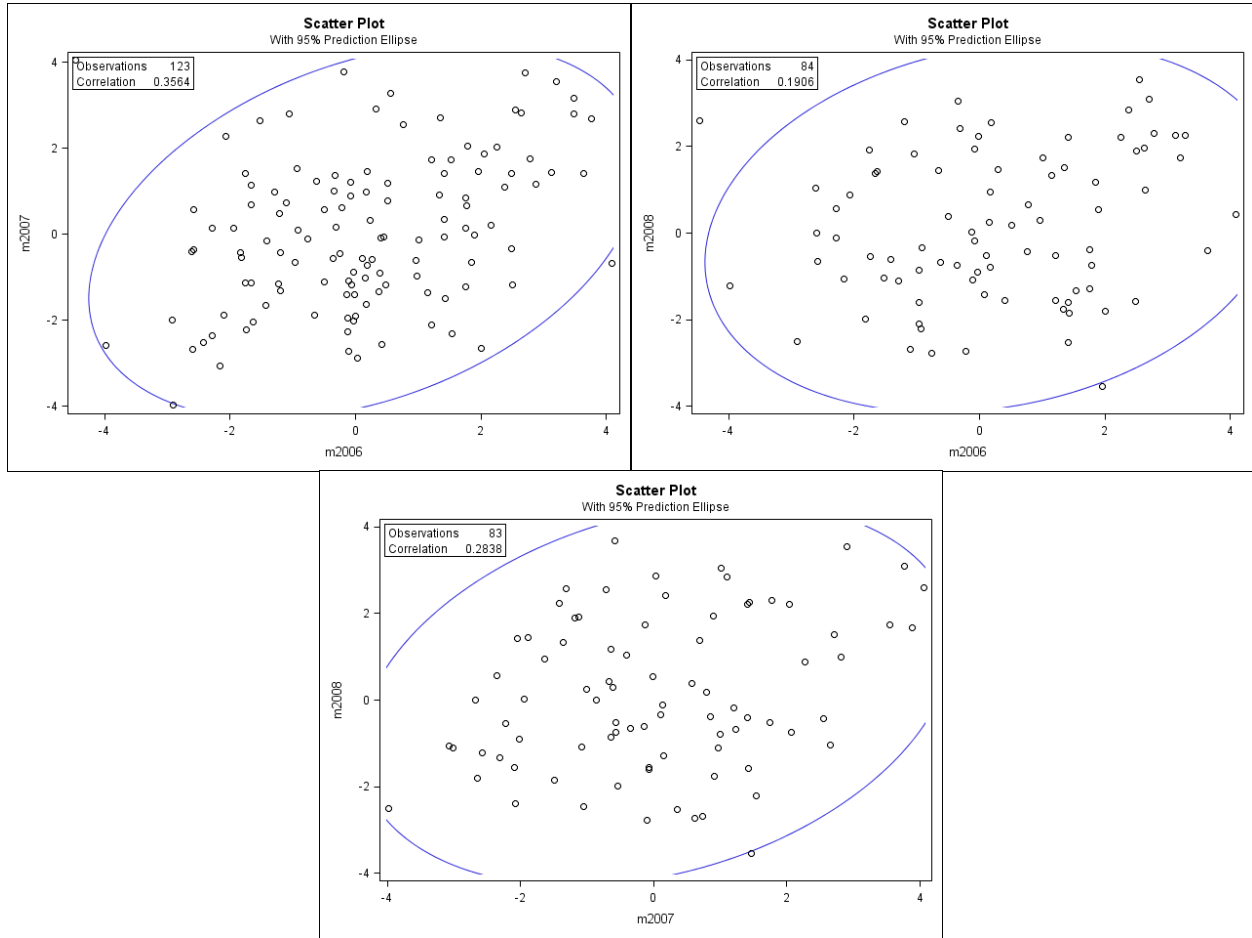


Figure 4.16: Excerpt of Final Stability Analysis for NCES Data

Table 4.13: Correlation of Yearly Mathematics RMTEI Values for NCES Data

		Mathematics RMTEI Pearson Correlation Coefficient (p-value)		
	year	2006	2007	2008
Mathematics RMTEI	2006	1.0000	.3564 (<.0001)	.1906 (.0825)
	2007		1.0000	.2838 (.0093)
	2008			1.000

As the analysis suggests in Table 4.13, mathematics index values remained stable with statistically significant correlation between 2006 and 2007 as well as between 2007 and 2008. Reading index values remained stable with statistically significant correlation between 2006 and 2007 as well as between 2007 and 2008. Reading and Mathematics index values, however, did not demonstrate statistically significant correlation between 2006 and 2008. A reason for this can likely be attributed to the two years of comprehensive induction services obtained by approximately half of the treatment teachers who generated statistically significant difference in student achievement in the third year compared to the control group of teachers (Glazerman et al., 2010, p. 92). Overall RMTEI values, however, remained stable with statistically significant correlation between all three years. Table 4.14 summarizes the results for yearly reading and overall RMTEI values.

Table 4.14: Correlation of Yearly Reading and Overall RMTEI Values for NCES Data

		Reading RMTEI			Overall RMTEI		
		Pearson Correlation Coefficient (p-value)			Pearson Correlation Coefficient (p-value)		
	year	2006	2007	2008	2006	2007	2008
Reading RMTEI	2006	1.0000	.2823 (.0016)	.1597 (.1418)			
	2007		1.0000	.2503 (.0217)			
	2008			1.0000			
Overall RMTEI	2006				1.0000	.3819 (<.0001)	.2432 (.0202)
	2007					1.0000	.3144 (.0030)
	2008						1.0000

In addition to testing whether the NCES correlation values differed from zero with a null hypothesis of $H_0 : \rho = 0$ (depicted in Table 4.13 and Table 4.14), this research used a Fisher's z-transformation to test differences between the correlations. Conducting a Fisher's z-transformation of the correlation values with the corresponding hypotheses,

$H_0 : \rho_{2006,2007} - \rho_{2006,2008} = 0, H_0 : \rho_{2006,2007} - \rho_{2007,2008} = 0, H_0 : \rho_{2006,2008} - \rho_{2007,2008} = 0$, demonstrated no statistical difference between the correlation values in each set of Mathematics, Reading, and Overall RMTEI values for the NCES dataset.

4.10.1 Comparison of Stability Results with Phase I Metrics

Correlation analysis produced the results of Figure 4.17 and Figure 4.18 in which correlation coefficients were calculated for each of the mathematics Phase I metrics and Mathematics RMTEI values. Figure 4.17 indicates the correlation of the mathematics values between 2009 and 2010 as well as between 2009 and 2011.

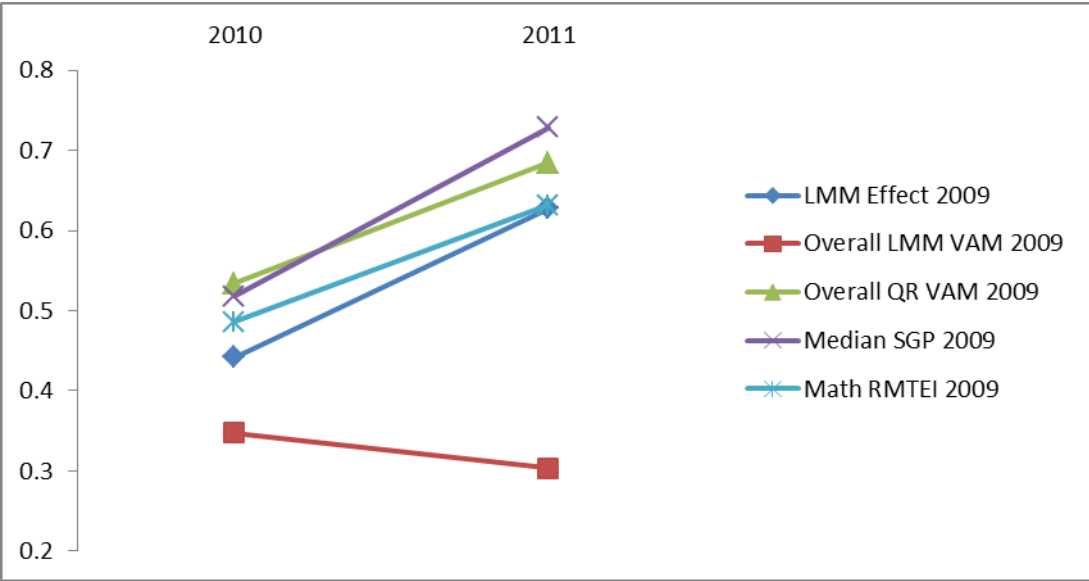


Figure 4.17: Mathematics Correlation Results in the Alabama District for 2009

Figure 4.18 indicates the correlation between the mathematics values of 2010 and 2011.

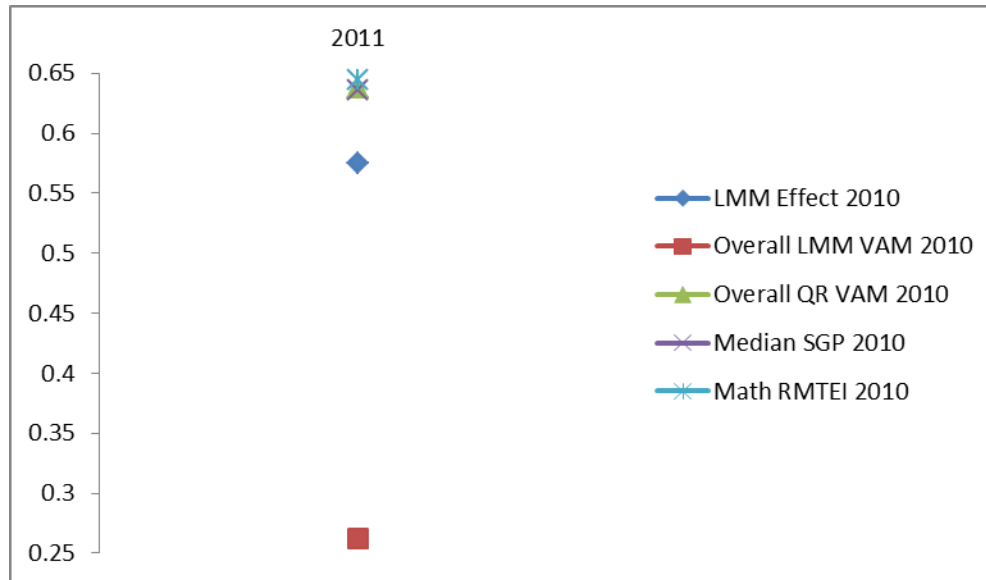


Figure 4.18: Mathematics Correlation Results in the AL District between 2010 and 2011

The purpose of the preceding analysis is to compare the stability results of the RMTEI values with the Phase I metrics; are the RMTEI values more stable than a single Phase I metric? The results are mixed. The Mathematics RMTEI values attained greater correlation than the LMM metrics but fell below both QR metrics between 2009 and 2010 as well as between 2009 and 2011. The Mathematics RMTEI values attained greater correlation than all of the mathematics Phase I metrics between 2010 and 2011. Reading values also produced mixed results. The Reading RMTEI values attained greater correlation than all of the reading Phase I metrics for the values between 2009 and 2010 as well as between 2009 and 2011. The Reading RMTEI values, however, only attained greater correlation than LMM Teacher Effect and Median SGP between 2010 and 2011.

Figure 4.19 indicates the correlation of the mathematics values between 2006 and 2007 as well as between 2006 and 2008 for a district in the NCES dataset.

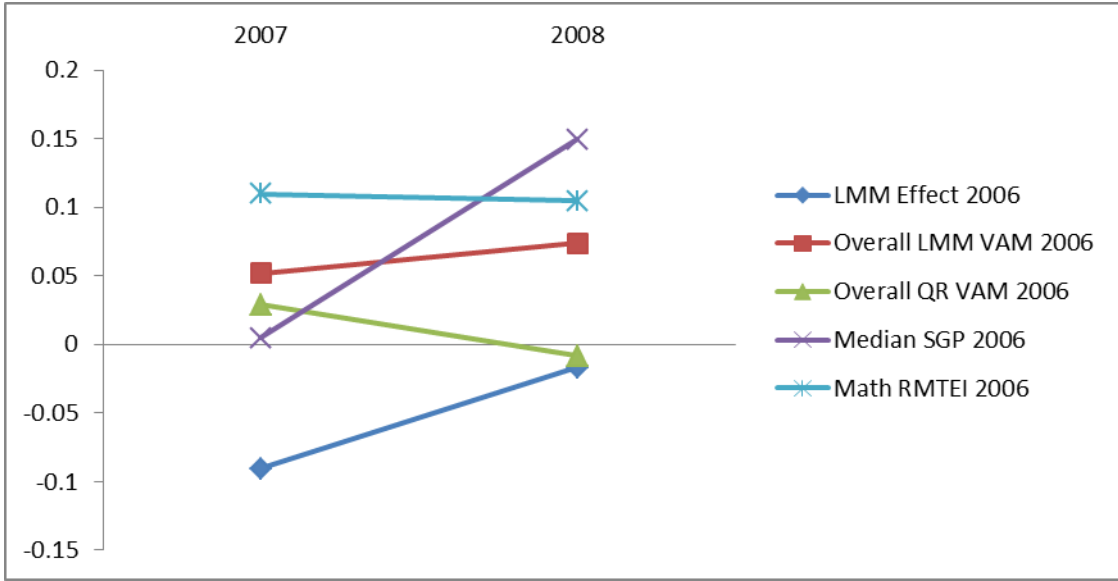


Figure 4.19: Mathematics Correlation Results for 2006 in a NCES District

Figure 4.20 indicates the correlation between the mathematics values of 2007 and 2008 for the same district in the NCES dataset.

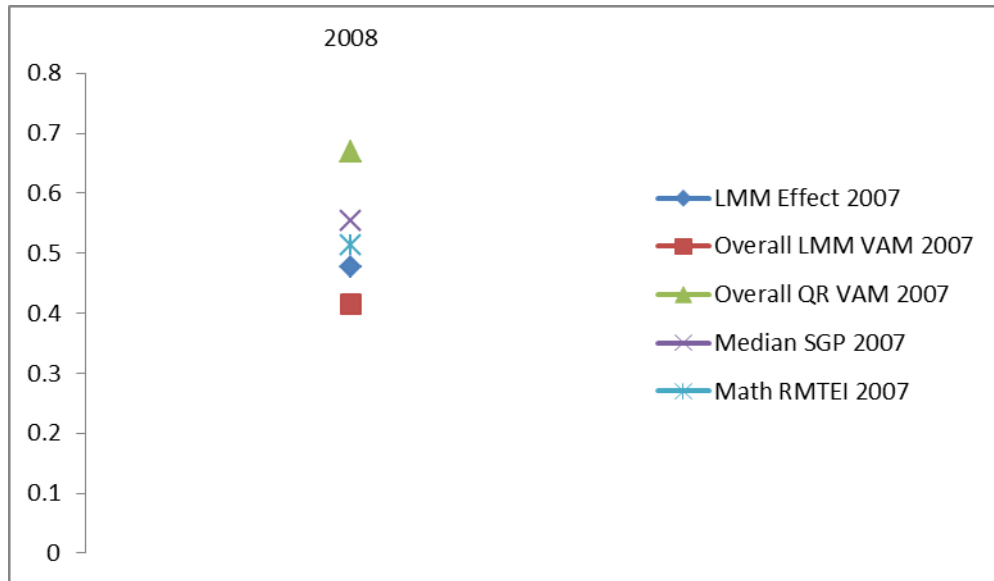


Figure 4.20: Mathematics Correlation Results between 2007 and 2008 for a NCES District

Once again, the results are mixed. The Mathematics RMTEI values attained greater correlation than all of the Phase I metrics between 2006 and 2007, but fell below only Median SGP between

2006 and 2008. The Mathematics RMTEI values, however, only attained greater correlation than the LMM metrics between 2007 and 2008. The Reading RMTEI values attained greater correlation than all of the reading Phase I metrics for the values between 2006 and 2007; however, the Reading RMTEI values fell below all of the Phase I metrics for the values between 2006 and 2008 and between 2007 and 2008.

The stability results from the Alabama and NCES districts demonstrate that the RMTEI process improved the stability of teacher effectiveness measurement while not fully solving the problem of poor stability from year to year. The RMTEI process did not fully overcome the unreliable nature of the Phase I metrics when applied individually. Testing revealed, however, that the correlation between years was statistically different from zero while not being statistically different from each other.

4.11 Summary

The methodology of Chapter 3 was placed into practice in Chapter 4 by examining five years of student and teacher data from a medium-sized Alabama school district. Overall and subject-specific Risk-Mitigated Teacher Effectiveness Index values were calculated for grades 4-8 for the 2008-09, 2009-10, and 2010-11 school years. Analysis of the NCES data yielded overall and subject-specific Risk-Mitigated Teacher Effectiveness Index values for the 2005-06, 2007-08, 2008-09 school years. Analysis was also undertaken in Chapter 4 to confirm the desired outcome of a precise and stable teacher effectiveness index.

The NCES RMTEI values for the 2005-06 school year were calculated based on a two-year process instead of three years (two years of data to generate a model from the previous cohort with application of the model in the current, third year). Linear Mixed Model and Quantile Regression Value-Added Measures were calculated based on a model generated from

the current year cohort and not the previous cohort. Linear Mixed Model Teacher Effect and Median Student Growth Percentile were calculated as previously discussed in Chapter 3. All four metrics utilized a single testing history of two consecutive years (2005 and 2006).

The NCES RMTEI values for the 2006-07 and 2007-08 schools years were calculated in the same fashion as the medium-sized Alabama district while utilizing only a single testing history group of two years. Analysis of the NCES data, which consists of 17 urban districts from across the United States, remedied the limitation of only exploring a medium-sized district in Alabama. In addition the NCES data also overcame the Alabama data limitation of only being from a suburban area.

Recall that the Alabama State Department of Education desires to place teachers into “at least” four categories following the development of an objective measure of teacher effectiveness based on student growth that augments its presently used formative assessment, EDUCATEAlabama (“Alabama’s Race,” 2010, p. 89). Teachers can be assigned to more subject-specific effectiveness categories in grades with a greater number of teachers (large districts, evaluating more than one district at a time, or grades within elementary schools). Also, teachers can be assigned to more effectiveness categories using the overall effectiveness index since the pool of teachers is greater (includes both mathematics and reading teachers). In the end the evidence suggests fixing the number of clusters at a maximum of four for a medium-sized district. A modification of the initially proposed categories in Section 1.4 complements the clusters found in Section 4.7:

1. Extraordinary gains in student growth.
2. Meets student growth.
3. Did not meet student growth.
4. Poor gains in student growth.

Chapter 5 : Assessment of Teacher Evaluation in Alabama

5.1 Introduction

The Alabama State Department of Education (ALSDE) implemented the Professional Education Personnel Evaluation (PEPE) system for administrators and teachers in 1993 and 1997 respectively (“AL PEPE for Teachers,” 1998, pp. 9–10). The ALSDE then implemented EDUCATEAlabama to replace PEPE as the educator evaluation system for typical classroom teachers prior to the 2010-2011 School Year (SY) (“EDUCATEAlabama Webinar,” 2011, p. 8). School counselors, school librarians, Alabama Reading Initiative Reading Coaches, and all special educators (pre-K, psychometrist/school psychologists, speech-language pathologists, and “special educators that teach students who generally take the Alabama Alternate Assessment”) will be integrated into the EDUCATEAlabama evaluation system for the 2011-2012 SY (“Alabama Professional,” 2011, sec. 2). School administrators will remain under the PEPE system until the ALSDE finalizes the development of a new administrator evaluation system, LEADAlabama, at a date to be published (“EDUCATEAlabama Webinar,” 2011, p. 46).

Prior to the implementation of EDUCATEAlabama, the PEPE system led school administrators to provide non-tenured teachers with annual summary scores for eight competencies. These eight competencies were then summed to provide a composite competency score to be used for summative purposes. Tenured teachers received a composite competency score every year, two years, or three years at the discretion of the local school system (“Professional Education,” 2008, p. 15).

5.2 Professional Education Personnel Evaluation (PEPE)

The PEPE system required instructional leaders to evaluate teachers' performance against eight competencies "which effective educators are known to possess" ("Professional Education," 2008, p. 1). The competencies include: Preparation for Instruction, Presentation of Organized Instruction, Assessment of Student Performance, Classroom Management, Positive Learning Climate, Communication, Professional Development and Leadership, and Performance of Professional Responsibilities ("Professional Education," 2008, p. B-3-B-8). A four point scale was used to score all eight competencies:

- 1) "Unsatisfactory - Indicates the educator's performance in this position requirement is not acceptable. Improvement activities must be undertaken immediately.
- 2) Needs Improvement - Indicates the educator's performance sometimes but not always meets expectations in this position requirement. Improvement activities are required for performance to consistently meet standards.
- 3) Area of Strength - Indicates the educator consistently meets and sometimes exceeds expectations for performance in this position requirement. Performance can be improved in the area(s) indicated, but current practices are clearly acceptable.
- 4) Demonstrates Excellence - Indicates the educator does an outstanding job in this position requirement. No area for improvement is readily identifiable" ("Professional Education," 2008, p. 19).

Multiple observations, an interview, and a review supported the final summary score for each competency. The eight competencies were then summed to provide a composite competency score for each teacher.

5.3 EDUCATEAlabama

"EDUCATEAlabama is a formative system designed to provide information about an educator's current level of practice within the Alabama Continuum for Teacher Development, which is based on the Alabama Quality Teaching Standards (AQTs)" ("EDUCATEAlabama,"

n.d.). The Alabama Continuum for Teacher Development provides benchmarks of performance for each teaching standard along the teacher continuum: pre-service, beginning, emerging, applying, integrating, and innovating (“Alabama Continuum,” 2009). The system is a means to encourage dialogue between the educator and the instructional leader.

EDUCATEAlabama begins with an educator self-assessment that is completed at the beginning of the school year based on the AQTs. The educator and instructional leader then complete a Professional Learning Plan focused on a select number of areas that demand the greatest attention. Based on observations and continued dialogue throughout the school year, evidence of educator growth is recorded for the corresponding areas of the Professional Learning Plan. Instructional leaders close educators’ evaluations following the conclusion of the school year and open a new evaluation prior to the beginning of the next school year. The EDUCATEAlabama process does not generate summative evaluation data (“EDUCATEAlabama Webinar,” 2011).

Based on the focus of U.S. Education policy to close achievement gaps and research demonstrating that teachers have the greatest impact on student learning compared to other factors controlled by school systems, Alabama has moved away from measures that offer teacher accountability and development (PEPE) and focused entirely on teacher development (EDUCATEAlabama). This dissertation research intends to provide evidence of whether NCES teacher observational data from 2006 are a stable predictor of teachers’ student achievement gains and correlated with the Risk-Mitigated Teacher Effectiveness Index. The desired outcome is evidence of quantitative, observational evaluations being a stable predictor of student achievement gains, a complement to the Risk-Mitigated Teacher Effectiveness Index, and an appropriate tool for Alabama to provide teacher accountability and development.

5.4 NCES Teacher Effectiveness Scoring Analysis

Teacher observational data obtained from 17 urban school districts from 2006 contain scoring of effectiveness practices for approximately 700 beginning teachers that received either comprehensive teacher induction or the existing, less intensive induction services provided by the district. Student Achievement data from the same 17 school districts will be used to provide evidence of whether the observational effectiveness scoring is a stable predictor of teachers' student achievement gains and correlated with the Risk-Mitigated Teacher Effectiveness Index.

Data from cohort 1, 2005-2006, were used to generate Risk-Mitigated Teacher Effectiveness Index values for all teachers with pre-test and post-test student achievement scores. The 2006 teachers who then received a classroom observation (all teachers who taught literacy but not English as a Second Language, special education, or those with one year or greater experience) were subsequently analyzed. Teachers were not separated by grade since students' z-scores allow comparison of teachers across grades. Both treatment and control groups were included. Approximately 180 teachers met all requirements.

“Observers scored teachers in each of three constructs based on a set of items that are believed to be indicators of good practice: implementation of a lesson, content of a lesson, and classroom culture” (Glazerman et al., 2010, p. 32). The three domains of good teaching practice consisted of multiple indicators that were measured on a five-point scale: (1) no evidence, (2) limited evidence, (3) moderate evidence, (4) consistent evidence, and (5) extensive evidence. Each domain then received an overall composite score consisting of the average of the domain's indicators (Glazerman et al., 2010, p. 32).

Concurrent to the development of the RMTEI values, NCES teacher effectiveness scores in 2006 were analyzed as predictors for student achievement gains in 2006. Specifically, teacher

composite scores for Content of a Lesson, Classroom Culture, and Implementation of a Lesson (CSLITIMP) were examined as predictors for student reading gains. The results were clear that the observational evaluations were a strong predictor for student achievement gains in reading with each composite score being statistically significant at the $\alpha = 0.05$ level. P-values for Content of a Lesson, Classroom Culture, and Implementation of a Lesson were .0136, < .0001, and .0003, respectively.

In order to suggest a relationship between observational evaluations and an objective measure of teacher effectiveness, a requirement exists to ascertain the correlation of NCES teacher effectiveness scores with the Risk-Mitigated Teacher Effectiveness Index computed for the 2005-2006 school year. Recall Alabama's Race to the Top Phase II application that expressed the need for correlated components of a teacher evaluation system. As a result, Figure 5.1 and Table 5.1 show a comparison of the RMTEI values for reading against the observational evaluations.

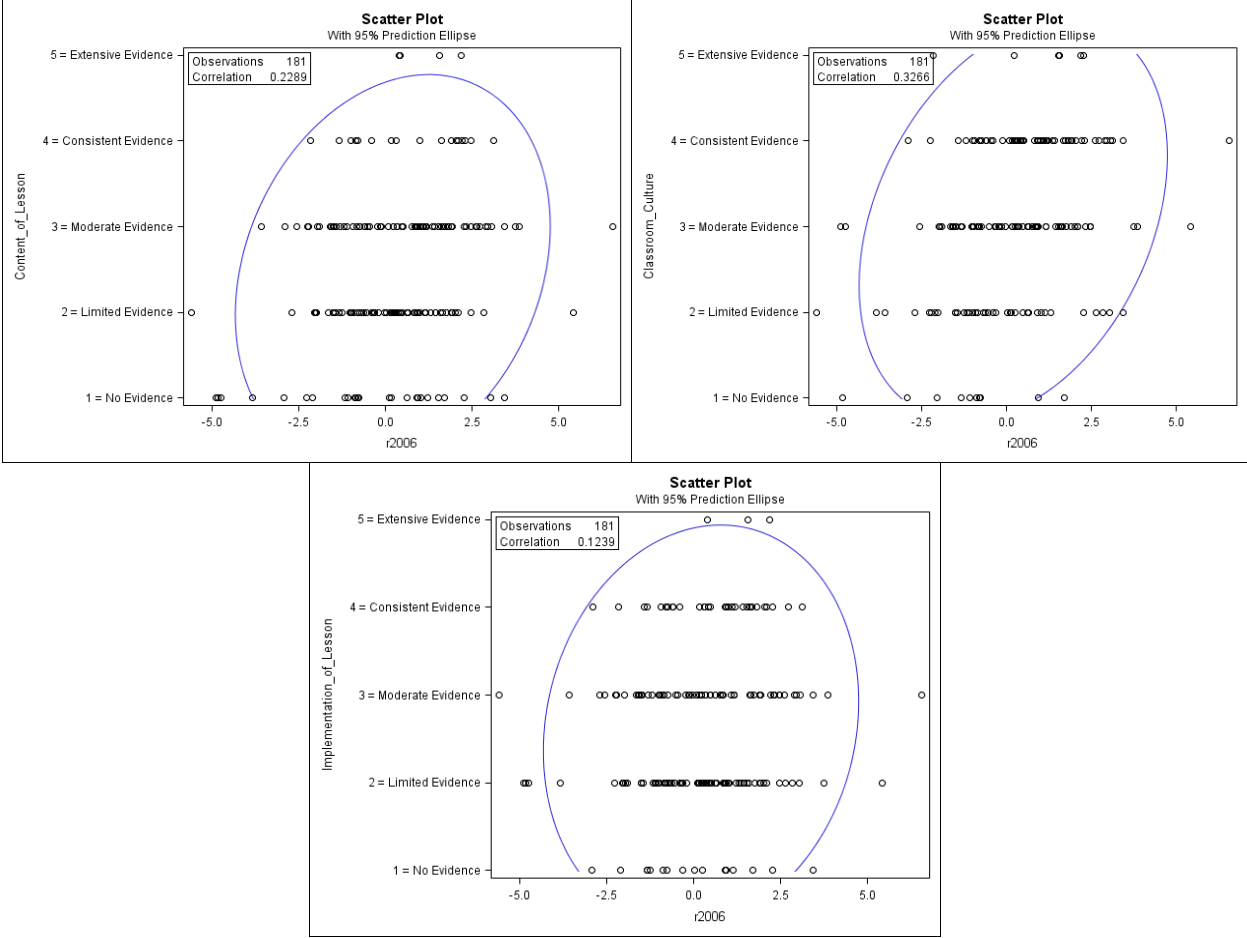


Figure 5.1: Scatter Plots of RMTEI Reading Values versus Observational Evaluations

Table 5.1: Correlation of RMTEI Reading Values and Observational Evaluations

		Observational Domains		
		Pearson Correlation Coefficient (p-value)		
Domain	RMTEI	Content of Lesson	Classroom Culture	Implementation of Lesson
	r2006	.2289 (.0019)	.3266 (<.0001)	.1239 (.0965)

The reading index values showed statistically significant correlation at an $\alpha = 0.05$ with Content of a Lesson and Classroom Culture from the 2006 classroom observations. Reading index values showed near statistically significant correlation with Implementation of a Lesson. These correlation results also confirm the precision of RMTEI values.

As a final bit of investigation for the group of literacy teachers that received observational evaluations in 2006, analysis was undertaken to determine whether there is a statistical difference between induction groups (control vs. treatment) while considering their RMTEI values. Glazerman et al. did not observe statistically significant differences between treatment and control teachers' performance on observational scoring (2010, p. xxxi). Years of induction services (one or two years) did not merit consideration for the treatment group since only one year had transpired when the observational evaluations took place. This research similarly concludes there is not a statistical difference between treatment and control teachers' performance while considering their RMTEI reading values. Conducting an ANOVA to test if teachers' Treatment Status affects the RMTEI reading values demonstrated that one fails to reject $H_0 : \mu_1 = \mu_2$ at an $\alpha = 0.05$ level.

5.5 Summary

The Alabama State Department of Education has expended significant resources developing EDUCATEAlabama as the next evolution of teacher evaluation. This process provides an approach for instructional leaders to communicate with educators, to chart a path for development, and to make a final assessment of the established goals. There is no accountability mechanism in place with this formative assessment that is directly tied to student growth, however. Alabama is heading in the opposite direction of the strategical framework established by the U.S. Department of Education in "A Blueprint for Reform".

According to Alabama's Race to the Top Phase II application, the ultimate desire is correlated components of a teacher evaluation system tied to student growth that are able to place teachers into "at least" four effectiveness categories ("Alabama's Race," 2010, p. 89). The Risk

Mitigated Teacher Effectiveness Index along with the observational assessment system in the Teacher Induction Study provide the Alabama State Department of Education with an approach to meet its stated need.

Chapter 6 : Research Summary

6.1 Conclusion

Alabama's Race to the Top Phase II application clearly stated a desire to research and develop a process to measure teacher effectiveness based on student growth with Race to the Top funds. Alabama came in last place out of the 36 states that submitted Phase II applications, thus it was not granted funds by the U.S. Department of Education. Dr. Morton, the former State Superintendent of Education, stated that despite knowing Alabama's application would not be competitive in Phase II, it would serve as a foundation for needed reforms. As a result, the primary objective of this research was to fill the need of augmenting Alabama's formative educator evaluation system, EDUCATEAlabama, with a precise and stable teacher effectiveness index based on student growth. This was accomplished.

A review of the literature to evaluate the effectiveness of teachers highlighted two general techniques. Firstly, LMMs produce estimates of the random teacher effect in the form of EBLUPs. Secondly, LMMs produce predictions of student achievement scores in order to calculate teachers' value-added measures by comparing the students' actual achievement scores with the predicted scores.

Colorado calculated SGPs using Quantile Regression to determine school effectiveness by comparing the medians of schools' student growth percentiles. This research used Quantile Regression to determine a third objective measure of teacher effectiveness by calculating a growth percentile for each student. The median of a teacher's aggregated student growth percentiles was calculated and compared to other teachers' medians within the population.

Lastly, a fourth method to measure teacher effectiveness was developed by combining Quantile Regression with the LMM practice of calculating achievement score predictions to derive teachers' value-added measures. The 0.5 quantile model generated by Quantile Regression produced predictions of student achievement scores in order to calculate teachers' value-added measures by comparing the students' actual achievement scores with the predicted scores.

The development of the Risk-Mitigated Teacher Effectiveness Index comprised three phases. Phase I entailed calculating the four teacher effectiveness metrics described above:

- 1) Linear Mixed Model Teacher Effect - statistical prediction of the relative value of a particular teacher by measuring the teacher deviation from the district mean; greater is better.
- 2) Linear Mixed Model Value-Added Measure - a teacher's average of the difference between students' actual achievement and predicted achievement had they been taught by the average teacher in the district; greater is better.
- 3) Median Student Growth Percentile - serves as an indicator of student growth associated with each teacher by calculating the median of a teacher's student growth percentiles. A student's growth percentile is obtained by determining what percentage of other students had less growth in testing achievement; greater is better.
- 4) Quantile Regression Value-Added Measure - a teacher's average of the difference between students' actual achievement and predicted achievement of a typical student within the district; greater is better.

Subject-specific and overall teacher index values were calculated in Phase II with an analytical process involving principal component analysis that used the four teacher effectiveness metrics determined in Phase I.

Since teachers' Phase I metrics for a given subject, grade, and year have varying units of measurement, the principal components were obtained from the standardized version of those metrics by calculating the eigenvectors of the metrics' correlation matrix. The variance of the principal component scores from the dominant principal component is the largest eigenvalue associated with the corresponding eigenvector. This eigenvalue absorbs the preponderance of

the variation of the system, and the principal component scores from the dominant principal component became teachers' subject-specific RMTEI values. If teachers instruct mathematics and reading, then an overall teacher effectiveness index was obtained by taking the mean of their two subject indexes. Otherwise, the single, subject index was the teachers' overall values.

The principal components served as the inputs to Phase III, Cluster Analysis. Since the RMTEI process produces a dominant principal component that continually leads to elliptically shaped clusters in two dimensions, Ward's clustering method, which expects elliptically shaped clusters, was employed as a general prescription of the RMTEI process. Ward's clustering method illuminated teachers with similar characteristics (principal components) in the data, which provided better assignments to teacher effectiveness categories compared to clustering with a single Phase I metric. Categories can be then used to identify teachers who employ pedagogical strategies or exhibit certain behaviors that positively impact student learning.

In order to accomplish the primary objective of this research, the following sub-objectives were completed:

- 1) Develop techniques with Alabama's database infrastructure to streamline the process of establishing longitudinal data from existing yearly data in order to make this information readily accessible to school districts (see Appendix 1).
- 2) Develop techniques with Alabama's database infrastructure to streamline the process of linking student achievement data with teachers in order to make this information readily accessible to school districts (see Appendix 1).
- 3) Confirm or modify Alabama's effectiveness rating categories based on being able to detect statistically different groups of teachers by effectiveness.
- 4) Report all teachers in accordance with the rating categories for an Alabama school district using the newly developed objective effectiveness index.
- 5) Provide an assessment of Alabama's educator evaluation system, compare and contrast the results of an objective effectiveness index with observational teacher evaluation data, and propose an observational, summative assessment for Alabama that is

a predictor of student achievement gains and correlated with an objective effectiveness index.

Specifically addressing sub-objective five (5), Alabama's educator evaluation system, EDUCATEAlabama, does not contain an accountability mechanism directly tied to student growth. It is not in compliance with "A Blueprint for Reform". The Risk Mitigated Teacher Effectiveness Index provides the Alabama State Department of Education (ALSDE) with an objective measure of teacher effectiveness. It is correlated with the observational assessment system in the Teacher Induction Study. The combination of these two assessment instruments that are based on student growth provides the ALSDE with a structure to meet its Race to the Top requirement of a new evaluation system with correlated components.

6.2 Limitations of the Risk Mitigated Teacher Effectiveness Index

As with any endeavor whose purpose is to improve the quality of an existing process, there are always certain aspects of that endeavor that cannot be fully addressed or improved. The Risk Mitigated Teacher Effectiveness Index is no exception; however, the Index's purpose by definition is to reduce the negative consequences that exist in the literature when measuring teacher effectiveness. Many of those negative consequences have been addressed in the preceding sections, but some limitations of the process remain and deserve discussion.

6.2.1 Limitations in Reporting

The RMTEI process places teachers into categories following the development of an objective measure of teacher effectiveness based on student growth. The range of categories extends from extraordinary to poor gains of student growth. Despite the importance of knowing such information, a limitation exists in its reporting by not being able to explicitly state why

certain teachers have appropriate growth whereas others do not. Although this is characteristic of many teacher effectiveness measuring instruments, education leaders would then have to conduct observational evaluations to ascertain the nature of a teacher’s performance that leads to the corresponding student growth.

Secondly, measuring teacher effectiveness through the use of student testing data precludes conducting assessments of teachers in certain subjects and grades. While mathematics and reading testing typically commence yearly for grades 3-8, the lack of yearly testing of other subjects does not allow student growth to be measured or linked to a teacher. Therefore, the application of the RMTEI process can rarely take place beyond the eighth grade or in subjects other than mathematics and reading due to data limitations. Alabama presently administers the tests in Table 6.1 (“Alabama Student Assessment Program Overview,” n.d.).

Table 6.1: Alabama Student Assessment Program Overview

Grades	Subject	Test
3-8	Reading and Mathematics	Stanford 10 and ARMT
9-12	Mathematics, Reading, Language, Social Studies, Science/Biology	Alabama High School Graduation Exam (AHSGE)
5, 7, and 10	Writing	Alabama Direct Assessment of Writing (ADAW)
5 and 7	Science	Alabama Science Assessment (ASA)
8	English, Mathematics, Reading, and Science	EXPLORE

Lastly, with the creation of mutually exclusive testing histories for the Alabama district as described in Section 4.2.2 for LMMs (NCES students only formed the minimum, two year testing history), all students became part of a testing history except those without a recorded test in the previous year and those who did not have a linked teacher. The lack of a linked teacher

only applies to the Alabama district as the NCES data had all teachers linked with their students' achievement scores. Therefore, the excluded students fell under one of the three classifications:

- 1) A lone test in the current year (students new to the school district);
- 2) No recorded test for the previous year (students who may have been in the district for an extended period but failed to take the previous year test); and
- 3) A teacher could not be linked to the student for the current year, which only applies to the Alabama district (for unknown reasons the scheduling database did not contain the student despite the student having recorded test scores).

All students in the testing histories then contributed to the creation of the two Phase I LMM metrics. Median Student Growth Percentiles and Quantile Regression Value-Added Measures on the other hand were calculated from students within a single testing history of two years.

Once again, select students were omitted from the analysis. In this case, however, only students without a linked teacher in the current year were omitted (Alabama district only). As a result, the research findings apply only to students likely to be tested and free of errors with regard to recording student schedules.

6.2.2 RMTEI Values and Small Populations of Teachers

Alabama's effectiveness index values in 2009 and 2010 were analyzed as predictors for student achievement gains in 2010 and 2011, respectively. The analysis paints a clear picture that the effectiveness index values are statistically significant predictors of student achievement gains for larger populations of teachers for each grade. For example, grades 5-8 would routinely have only three teachers for a single subject, thus the significance of the index values was rarely achieved. Grade 4, however, typically consisted of 14 teachers and produced subject-specific and overall effectiveness index values that were statistically significant predictors of student achievement gains for the next school year as stated above. The lone exception was the 2009 4th

grade reading index value that was not a statistically significant predictor of reading gains for 2010 with a p-value of .0843.

The precision of the RMTEI, as demonstrated with control charts in Section 4.9, was limited to larger populations of teachers. Specifically, when an individual teacher consisted of a subgroup of three or fewer effectiveness values and was compared to few teachers (e.g., two teachers per subject in grades 5-8 for the Alabama district) , the \bar{x} chart, which measures between-subgroup variability, determined that most teachers fell within the control limits and did not adequately distinguish teachers. The R-Chart, which measures within-subgroup variability, consistently remained in control as desired.

6.3 Future Study

As stated in Section 4.1 a significant amount of energy was put forth by this research to obtain quality student and teacher data from the ALSDE and Alabama Local Education Agencies. Much of the energy expenditure was of little benefit to this research as only a single medium-sized district provided the requisite data to fully implement the RMTEI process. As a result, additional data was obtained from the NCES. Alabama's Phase II Race to the Top application clearly stated a desire to create partnerships with the research institutions of the State and to make data readily available to them. The ALSDE must take action to accomplish these goals.

Future study remains whereby additional data must be obtained from Alabama to fully investigate the usefulness of the RMTEI approach for the State. The results of this research clearly allow one to offer its conclusions to the State, but the findings would have broader implications had the preponderance of the data come from Alabama. This research also proposes future study of the test-forms equating for the reading portion of the ARMT. Based on the

analysis of the previous chapters, reading RMTEI values trailed behind mathematics RMTEI values in correlation across years and as predictors for future reading gains. Additional data can illuminate whether this is a persistent issue or particular to the lone medium-sized Alabama district for the specific study years. If the future study also suffers from these issues, then analysis must be undertaken to determine if successive reading forms of the ARMT are equally difficult. For example, “if a test form is more difficult than those that precede and follow, it will systematically make the first measure of gain too low and the second too high” (Bock, Wolfe, & Fisher, 1996, p. 13).

As a means to “help campuses and educators identify individual students in need of intervention”, desirable research in Alabama includes calculating statistical projections of test scores to the next grade and evaluating those projections versus desired proficiency levels (Texas Education Agency, 2009, p. 25). Intervention methods of administrators would be employed when presented with the following two scenarios:

- 1) Students who currently meet the standard but are projected to not meet the proficiency standard.
- 2) Students who currently do not meet the standard and are projected to not meet the proficiency standard (Texas Education Agency, 2009, p. 25).

Alabama included similar language on this topic in its reform agenda as part of the Race to the Top application whereby it wants to “develop predictive trajectories for its students through graduation...[and] create a dashboard-style early warning system for teachers” (“Alabama’s Race,” 2010, p. 9). Any state’s strategic education plan would be wise to consider this method of projection “to accelerate student achievement, close achievement gaps, [and] inspire our children to excel” (“A Blueprint for Reform,” 2010, p. 2).

Lastly, the documents supporting EDUCATEAlabama's development provide an excellent structure to return teacher accountability to the evaluation. Noteworthy is the cross-walk established in the Alabama Continuum for Teacher Development that provides benchmarks of performance for the Alabama Quality Teaching Standards along the teacher continuum: pre-service, beginning, emerging, applying, integrating, and innovating ("Alabama Continuum," 2009). Therefore, future study consists of developing a summative assessment comprised of the observational effectiveness scoring in the Teacher Induction Study to be incorporated into EDUCATEAlabama to provide teacher accountability and development.

References

- A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act -- TOC. (2010, September 2). Laws; Publicity. Retrieved April 12, 2011, from <http://www2.ed.gov/policy/elsec/leg/blueprint/publicationtoc.html>
- Alabama Continuum for Teacher Development. (2009). Alabama State Department of Education. Retrieved from <http://alex.state.al.us/leadership/Alabama%20Continuum%20for%20Teacher%20Development.pdf>
- Alabama Professional Education Personnel Evaluation Program. (2011, May 12). *Alabama PEPE Program*. Retrieved May 27, 2011, from <http://www.alabamapepe.com/>
- Alabama Professional Education Personnel Evaluation Program for Teachers. (1998, August 31). Alabama State Department of Education. Retrieved from <http://pixdoc.com/doc/alabama+board+of+education+pepe+form/>
- Alabama Reading and Mathematics Test: Interpreting the Student Report. (2009, September 11). Alabama State Department of Education Student Assessment. Retrieved from <https://docs.alsde.edu/documents/91/ARMT%20Interpreting%20Student%20Group-%20Reports.pdf>
- Alabama Student Assessment Program Overview. (n.d.). Retrieved from http://www.hsv.k12.al.us/dept/merts/testing/Testing_overview.php
- Alabama's Race to the Top Application. (2010, June 1). Alabama State Department of Education. Retrieved from <http://www2.ed.gov/programs/racetothetop/phase2-applications/alabama.pdf>

- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37 –65. doi:10.3102/10769986029001037
- Betebenner, D. W. (2007, October 5). Estimation of Student Growth Percentiles for the Colorado Student Assessment Program. Colorado Department of Education. Retrieved from http://www.cde.state.co.us/cdedocs/Research/PDF/technicalsgppaper_betebenner.pdf
- Bice, T. (2010, September 21). Alabama Deputy State Superintendent of Education.
- Bock, R. D., Wolfe, R. G., & Fisher, T. H. (1996). *A review and analysis of the Tennessee Value-Added Assessment System*. Comptroller of the Treasury.
- Braun, H. I., Chudowsky, N., & Koenig, J. A. (Eds.). (2010). *Getting Value Out of Value-Added: Report of a Workshop*. Washington: National Academies Press.
- Colorado's Academic Growth Model. (2008, February 13). Colorado Department of Education. Retrieved from http://www.cde.state.co.us/cdeassess/documents/res_eval/FinalLongitudinalGrowthTAPReport.pdf
- Corcoran, S. P. (2010). Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Annenberg Institute for School Reform at Brown University. Retrieved from <http://www.annenberginstitute.org/products/Corcoran.php>
- Crouse, D. (2011, April 6). Director of Federal Programs and Professional Services, Roanoke City Schools, Roanoke, AL.

Delaware and Tennessee Win First Race to The Top Grants. (2011, January 27). Press Releases; Retrieved April 13, 2011, from <http://www2.ed.gov/news/pressreleases/2010/03/03292010.html>

DiChiara, L. (2011, April 1). Superintendent, Phenix City Public Schools, Phenix City, AL.

Edgeworth, F. Y. (1888). On a new method of reducing observations relating to several quantiles. *Philosophical magazine: a journal of theoretical, experimental and applied physics* (Vol. 25, pp. 184–191). Taylor & Francis.

EDUCATEAlabama. (n.d.).*Educator Evaluations*. Retrieved May 27, 2011, from <http://alex.state.al.us/leadership/evaluations.html>

EDUCATEAlabama Information Webinar. (2011, April 11). Alabama State Department of Education. Retrieved from <http://alex.state.al.us/leadership/evaluations.html>

Educator Effectiveness Resolution. (2010, May 27). Alabama State Board of Education. Retrieved from http://www.alsde.edu/html/boe_resolutions2.asp?id=1662

Fifth Grade Data Codebook: Early Childhood Longitudinal Study [United States]: Kindergarten Class of 1998-1999, Fifth Grade. (2006, February). United States Department of Education. National Center for Education Statistics.

Glazerman, S., Isenberg, E., Dolfen, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010, October). Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study (NCEE 2010-4028). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <http://ies.ed.gov/ncee/pubs/20104027/pdf/20104028.pdf>

Hao, L., & Naiman, D. Q. (2007). *Quantile Regression*. Quantitative applications in the social sciences. Thousand Oaks, Calif: Sage Publications.

- Holdren, J. P. (2011, January 6). America COMPETES Act Keeps America's Leadership on Target. *The White House Blog*. Retrieved from <http://www.whitehouse.gov/blog/2011/01/06/america-competes-act-keeps-american-leadership-target>
- House, D. H. (2010, March 17). Spline Curves. Retrieved from <http://www.cs.clemson.edu/~dhouse/courses/405/notes/splines.pdf>
- Institute of Education Sciences. (n.d.). *Director of IES*. Retrieved from <http://ies.ed.gov/director/biography.asp>
- Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Prentice Hall.
- Kaw, A., & Keteltas, M. (2009, November 20). Spline Method of Interpolation. Retrieved from http://numericalmethods.eng.usf.edu/mws/gen/05inp/mws_gen_inp_txt_spline.pdf
- Kincaid, D., & Cheney, W. (2002). *Numerical Analysis: Mathematics of Scientific Computing* (3rd Revised ed.). American Mathematical Society.
- Larson, J. (2010, November 16). Developer/Project Lead, Information Systems, Alabama State Department of Education.
- Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. (2010, December). Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Finding-Policy_Brief.pdf
- Lim, L. K. S., Acito, F., & Rusetski, A. (2006). Development of archetypes of international marketing strategy. *Journal of International Business Studies*, 37(4), 499–524.
doi:10.1057/palgrave.jibs.8400206

- Lissitz, B., & Doran, H. (2009, July). Modeling Growth for Accountability and Program Evaluation: An Introduction for Wisconsin Educators. Retrieved from <http://dpi.wi.gov/oea/pdf/introgrowth.pdf>
- Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems. *Journal of Educational and Behavioral Statistics*, 27(3), 255–270.
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Berkeley, CA: University of California Press.
- Mathews, J. H. (2004). Cubic Splines. Retrieved from <http://math.fullerton.edu/mathews/n2003/CubicSplinesMod.html>
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). *Evaluating Value-Added Models for Teacher Accountability* (1st ed.). RAND Corporation.
- McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009, October). The Intertemporal Variability of Teacher Effect Estimates. Retrieved from http://www.performanceincentives.org/data/files/news/PapersNews/200903_McCaffrey_etAl_TeacherEffectEstimate1.pdf
- Morton, J. B. (2010, September 1). An Open Letter. Alabama State Department of Education.
- National Center for Education Statistics. (n.d.). *Welcome to NCES*. Retrieved from <http://nces.ed.gov/>
- Pearson, M., & Stecher, B. (2004). *Organizational Improvement and Accountability: Lessons for Education from Other Sectors*. (B. Stecher & S. N. Kirby, Eds.). RAND.

- Professional Education Personnel Evaluation Program of Alabama. (2008, May 1). Alabama State Department of Education. Retrieved from <http://www.alabamapepe.com/teacher.htm>
- Pugh, J. (2008, February 25). Alabama Reading and Mathematics Test. Alabama State Department of Education Student Assessment. Retrieved from <https://docs.alsde.edu/documents/91/Alabama%20Reading%20and%20Mathematics%20Test.pdf>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Advanced quantitative techniques in the social sciences (2nd ed.). Thousand Oaks: Sage Publications.
- Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools* (pp. 137–162). Thousand Oaks, Calif: Corwin Press.
- SAS EVAAS for K-12. (2010). Retrieved from http://www.sas.com/resources/product-brief/SAS_EVAAS_for_K-12.pdf
- SAS/STAT 9.2 User's Guide Second Edition. (2009). Retrieved from <http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf>
- Schmidhammer, J. L. (2010). Agglomerative Hierarchical Clustering Methods. Retrieved from <http://bus.utk.edu/stat/stat579/Hierarchical%20Clustering%20Methods.pdf>
- Secretary Spellings Approves Additional Growth Model Pilots for 2008-2009 School Year. (2009, January 21). Press Releases; Retrieved April 13, 2011, from <http://www2.ed.gov/news/pressreleases/2009/01/01082009a.html>

- System Profile Report 2008-2009. (2009). Alabama State Department of Education. Retrieved from http://www.alsde.edu/html/school_info.asp?menu=school_info&footer=general&sort=county
- Texas Education Agency. (2009, January 12). Growth Model Pilot Application for Adequate Yearly Progress Determinations under the No Child Left Behind Act. Retrieved from <http://www.tea.state.tx.us/student.assessment/measures/archive/>
- Trietsch, D. (1999). *Statistical Quality Control: A Loss Minimization Approach*. World Scientific Pub Co Inc.
- Tufte, E. R. (2006). *Beautiful Evidence*. Graphics Pr.
- Wei, Y., & He, X. (2006). Conditional Growth Charts. *The Annals of Statistics*, 34(5), 2069–2097.
- West, B., Welch, K. B., & Galecki, A. T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton: Chapman & Hall/CRC.
- Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching. (2010, June). Bill and Melinda Gates Foundation. Retrieved from <http://www.gatesfoundation.org/highschools/Documents/met-framing-paper.pdf>

Appendix 1: Establishing Longitudinal Student Achievement Data Linked with Teacher Information

The process to create longitudinal student data linked with current year teachers began with actions at the school district level followed by the actions of this research. Appendix 1 will describe each set of actions in order to offer these methods to other districts within Alabama.

I. District Actions

The medium-sized Alabama District provided four types of files for this research: Schedule, Course Counts, Teacher Information, and Student Achievement.

A. Schedule

The schedule dataset was obtained by a Structured Query Language (SQL) program that queried the STI Education Data Management Solutions database. The SQL program requested the specific fields of student identification number, period, course number, teacher identification number, and school identification number to export to a comma delimited/Comma Separated Values (CSV) file. In order to account for the manner in which students were recorded in ‘terms’, the district had to perform a query for elementary schools and an additional query for secondary schools. Elementary school students are recorded in a single term for a year whereas secondary school students are recorded in two terms for a single year. This accounts for secondary students possibly having different schedules in the second half of the school year.

B. Course Counts

The ‘course counts’ dataset was obtained by a SQL program that queried the STI Education Data Management Solutions database. The SQL program requested the specific fields of course number, short name, long name, school identification number, and the number of students to export to a CSV file. The course counts information was queried separately from the

schedule information due to the applicable data being in a different STI table. The district technology coordinator acknowledged that the schedule and course counts information could possibly be obtained together in a single query.

C. Teacher Information

The teacher information dataset was obtained from the yearly Local Education Agency Personnel System (LEAPS) report. This report is generated from a query of the McAleer Accounting System database. The district provided this research with a portion of the CSV LEAPS report, specifically the fields of school year, teacher identification number, gender, ethnicity, highest degree obtained, school, and teaching experience.

D. Student Achievement

School districts receive yearly student achievement files via distributed compact disks from the Alabama State Department of Education. The student achievement files are CSV in nature and consist of a matrix with over 200 column vectors reporting the nature of students and their respective ARMT scores. The district provided this research with a portion of the CSV student achievement files, specifically Student Identification Number, grade, school, and ARMT Reading and Mathematics Scores.

II. Researcher Actions

The process to create longitudinal student data linked with current year teachers began with extracting students and their teachers from the schedule dataset using the desired grade and course (mathematics or reading). The schedule datasets were yearly in nature and consisted of elementary and secondary school files. For example, two Rectangular format files (columns represent variables and rows represent observations) with comma separated values contained the schedules of all students in 2009:

- 1) 2008-2009 Elementary School (kindergarten – 4th grade)
- 2) 2008-2009 Secondary School (5th grade – 8th grade)

The files contained the column headings of student identification number, period, course number, teacher identification number, and school identification number. For each student, his/her student identification code was repeated to capture the different courses and corresponding periods. The teacher identification code also repeated if that teacher taught all subjects for that student. The school identification number also repeated for the student. An excerpt of the schedule dataset is presented in Table 6.2.

Table 6.2: Excerpt of Schedule Dataset

Student ID Number	Period	Course Number	Teacher	School
1305648914	6	300.02	5085	20
1305648914	1	301.02	5085	20
1305648914	2	302.02	5085	20
1305648914	4	303.02	5085	20
1305648914	11	305.02	5085	20
1305648914	7	307.02	5085	20
1305648914	3	308.02	5085	20
1305648914	5	311.02	5085	20
1864909372	6	300.04	8299	20
1864909372	1	301.04	8299	20

The ‘course counts’ dataset for each year was used to discern the course name for each course number. Column headings consisted of course number, short name, long name, school identification number, and the number of students in that particular course number. The course numbers were devised to represent not only a particular course in a particular grade, but also the particular section. The course number along with the school identification number depicted an actual classroom of students. An excerpt of the ‘course counts’ dataset is presented in Table 6.3.

Table 6.3: Excerpt of Course Counts Dataset

Course Number	SName	LName	School	Count
6510.01	READING	READING	40	28
6510.02	READING	READING	40	30
6510.03	READING	READING	40	24
6520.01	READING	READING	40	30
6520.02	READING	READING	40	29
6520.03	READING	READING	40	26
6530.01	READING	READING	40	29
6530.02	READING	READING	40	29
6530.03	READING	READING	40	23
6550.01	READING INV	READING INTERVENTION	40	14

The reading and mathematics course numbers for each grade and year were then recorded to create the requisite extraction from the schedule dataset.

After importing the schedule dataset into SAS, sorting by teacher, and specifying the year and grade for analysis, the appropriate template found the appropriate courses/teachers in the schedule dataset that would be linked with the reading and mathematics student achievement data of the current year. For example, the user specifies the year and grade to be 2009 and 6th grade respectively. As a result, the required schedule dataset is retrieved (secondary school dataset) followed by the extraction of reading and mathematics courses/teachers for that particular grade and year using a range of course numbers:

```
%let year=2009;
%let SY='2008-2009';
%let grade=6;

%macro teachers (year=);
data mathtchrs;
if &grade = 4 then set schedule.elemtchrs&year;
else if &grade=5 and &year=2007 then set schedule.elemtchrs&year;
else if &grade=5 and &year=2008 then set schedule.elemtchrs&year;
else if &grade=5 and &year>2008 then set schedule.sectchrs&year;
else if &grade>=6 then set schedule.sectchrs&year;

if &grade= 6 and &year=2009 then
if school ne 40 or cnum < 6300.01 then delete;
else if cnum > 6350.02 then delete;
MTeacher=teacher;
run;
```

```

data readtchrs;
if &grade = 4 then set schedule.elemtchrs&year;
else if &grade=5 and &year=2007 then set schedule.elemtchrs&year;
else if &grade=5 and &year=2008 then set schedule.elemtchrs&year;
else if &grade=5 and &year>2008 then set schedule.sectchrs&year;
else if &grade>=6 then set schedule.sectchrs&year;

if &grade=6 and &year=2009 then
if school ne 40 or cnum < 6100.01 then delete;
else if cnum > 6550.03 then delete;
else if 6130.03<cnum<6500.01 then delete;
RTeacher=teacher;
run;
%mend teachers;

```

The medium-sized Alabama district also provided a CSV teacher information file. It contained the column headings of school year, teacher identification number, gender, ethnicity, highest degree obtained, school, and teaching experience expressed in months teaching in the district, state (not including the district), public (outside the district or state), and private systems. An excerpt of the teacher information file is presented in Table 6.4.

Table 6.4: Excerpt of Teacher Information File

Year	Teacher	Gender	Ethnicity	Highest Degree	School	District	State	Public	Private
2005-06	6345	F	1	M	35	36	0	0	0
2006-07	6345	F	1	M	35	48	0	0	0
2007-08	6345	F	1	M	45	48	0	0	0
2008-2009	6345	F	1	M	45	72	0	0	0
2005-06	7216	F	1	M	25	96	0	0	0
2006-07	7216	F	1	M	10	108	0	0	0
2007-08	7216	F	1	M	10	108	0	0	0
2008-2009	7216	F	1	M	10	132	0	0	0
2009-2010	7216	F	1	M	10	144	0	0	0
2010-11	7216	F	1	M	10	156	0	0	0

The teacher information for the current year was imported into SAS, sorted by teacher, and merged with the teachers/students extracted from the schedule dataset.

```

data District_mathtchrs;
merge mathtchrs teacher_info;
by teacher;
if year = ' ' then year= &SY ;
if Student_ID_Number = ' ' then delete;
myrs_employ=(system+state+public+private)/12;
mhighest_degree=highest_degree;
if myrs_employ=' ' then myrs_employ =0;

```



```

run;

data District_readtchrs;
merge readtchrs teacher_info;
by teacher;
if year = ' ' then year= &SY;
if Student_ID_Number = ' ' then delete;
ryrs_employ=(system+state+public+private)/12;
rhighest_degree=highest_degree;
if ryrs_employ=' ' then ryrs_employ =0;
run;

```

Following the merge, the new datasets were then sorted by student identification number instead of teacher in preparation for the merge with the student achievement files.

```

proc sort data=District_mathtchrs nodupkey;
by student_ID_number;
run;

proc sort data=District_readtchrs nodupkey;
by student_ID_number;
run;

```

The nodupkey option precludes multiple teachers for the same student in a particular subject.

The goal is a single teacher for a single student in a particular subject. The option selects the first teacher listed in the dataset for a particular student. This action is desired as the teachers had been previously extracted from the schedule dataset by a priority system. For instance, a student may have a reading teacher as well as reading intervention teacher. The reading teacher was extracted first from the schedule dataset followed by the reading intervention teacher. The nodupkey then removed the reading intervention teacher. Therefore, the student has the appropriate reading teacher linked for future analysis.

Analysis for 6th grade in 2009, for example, required merging student scaled scores from the 2006-07, 2007-08, and 2008-09 Alabama Reading and Mathematics Test (ARMT) with the students' respective mathematics and reading teachers for the 2008-09 school year. Districts only receive yearly student achievement files via distributed compact disks from the Alabama State Department of Education. The student achievement files are CSV in nature and consist of

a matrix with column vectors of Student Identification Number, grade, school, and a single year of ARMT Reading and Mathematics Scores. The files were imported into SAS and sorted by student identification number. Thereafter, three files of student achievement data, a reading teachers' file, and a mathematics teachers' file were merged to allow subsequent longitudinal analysis. The student identification number provides the linkage of student data to teacher data.

```

data District_students;
if &year=2009 then merge student.studata&year2 student.studata&year3
student.studata&year4 District_mathtchrs District_readtchrs;
by Student_ID_Number;

if &currentgrade_SY ne &grade then delete;
if Student_ID_Number = ' ' then delete;
run;

```

An excerpt from the 6th grade, 2009 dataset is shown in Table 6.5 that displays the outcome of the process described in this appendix.

Table 6.5: Excerpt from 6th grade, 2009 Dataset

Student ID Number	Grade SY2009	School	ARMT_Read_2006	ARMT_Math_2006	ARMT_Read_2007	ARMT_Math_2007	ARMT_Read_2008	ARMT_Math_2008	ARMT_Read_2009	ARMT_Math_2009	MTeacher	RTeacher
1052237657	6	40	643	657	653	652	663	712	686	689	4260	5336
1914798360	6	40	578	611	583	614	566	646	635	657	4260	6294
1076229848	6	40	643	675	711	652	695	706	683	689	4260	6294
1656276879	6	40	663	636	698	647	684	664	698		6124	8023
1508806089	6	40	652	652	632	660	656	690	669	699	6124	3245
1111256901	6	40		597	662	602	667	632	672	634	6124	3245
1145074252	6	40	605	620	644	619	663	658	666	677	4260	6294
1695736851	6	40	681	639	662	660	717	700	725	732	4260	6294
1781892488	6	40	542	575	567	607	570	606	616	613	5948	8815
1668075284	6	40					570	554			6124	8023

A flowchart to illustrate and summarize the entire process is contained in Figure 6.1.

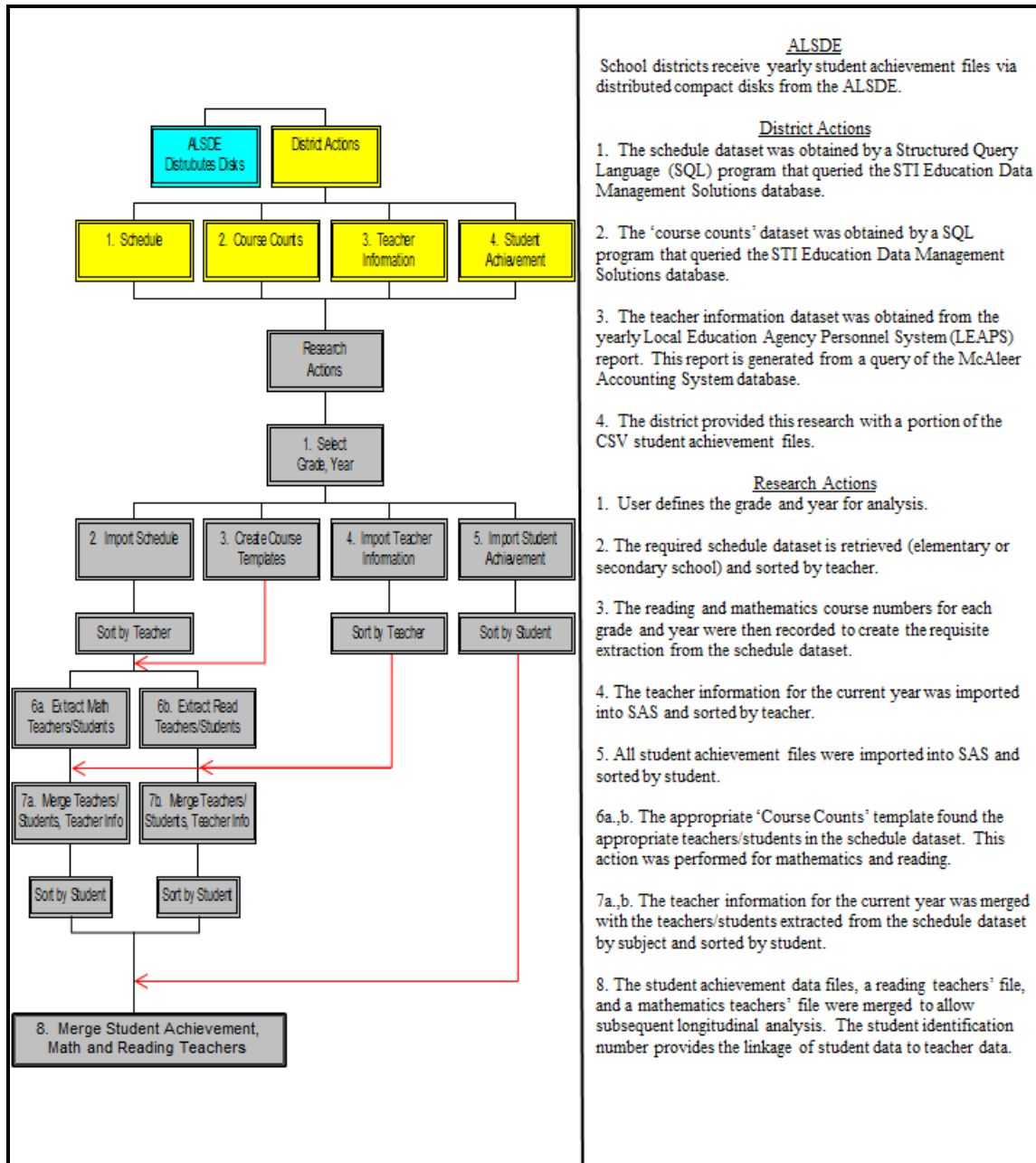


Figure 6.1: Establishing Longitudinal Student Achievement Data Linked with Teacher Information

Appendix 2: SAS Code

Input File:

```
LIBNAME student 'F:\Auburn\PhD Research\District A\Student Data\';
LIBNAME lunch 'F:\Auburn\PhD Research\District A\Free Reduced Lunch\';
LIBNAME appen3 'F:\Auburn\PhD Research\INSY 8990\Dissertation
Defense\Chapters 4-6\appen3';

%let myear1=ARMT_Math_2006;
%let myear2=ARMT_Math_2007;
%let myear3=ARMT_Math_2008;
%let myear4=ARMT_Math_2009;
%let myear5=ARMT_Math_2010;
%let myear6=ARMT_Math_2011;

%let ryear1=ARMT_Read_2006;
%let ryear2=ARMT_Read_2007;
%let ryear3=ARMT_Read_2008;
%let ryear4=ARMT_Read_2009;
%let ryear5=ARMT_Read_2010;
%let ryear6=ARMT_Read_2011;

%let mcheck=0;
%let rcheck=0;
%let ivcheck=0;

%let j=2007; /*initiates count for datasets. remains the same regardless
of year below */
/*-----*/
/* need to change with a different year */

%let year=2009;
%let model_year=2008;

%let k = 2008; /*= model_year*/
%let k_1=2007; /* k-1 with minimum of 2007; used to accommodate 5th
grade's testing hist: 1,2,2,2 for years 2008,09,10,11 */
%let k_2=2007; /* k-2 with a minimum of 2007; used to accommodate 6th
grade's testing hist: 1,2,3,3 for years 2008,09,10,11*/
%let currentgrade_SY=Grade_SY2009;
%let modelgrade_SY=Grade_SY2008;

%let SY='2008-2009'; /*current year; 2006-07, 2007-08, 2008-2009, 2009-
2010, 2010-11 */
%let Model_SY='2007-08'; /*model year; 2006-07, 2007-08, 2008-2009, 2009-
2010, 2010-11 */
```

Phase I, II, III:

```
%Macro RMTEI (grade=);

proc datasets lib=work kill nolist memtype=data;
quit;

/*-----*/
/*Organize data for current cohort of students*/

data District_RMTEI;
set student.district_students&grade&year;
int=1;
if &ryear3 ne ' ' and &myear3 ne ' ' and &ryear2 ne ' ' and &myear2 ne ' '
then indicator=2007;
else if &ryear3 ne ' ' and &myear3 ne ' ' then indicator=2008;
readgain=&ryear4-&ryear3;
mathgain=&myear4-&myear3;
run;

proc sort data=District_RMTEI;
by school;
run;

proc means data=District_RMTEI;
by school;
var ARMT_Math_&model_year;
output out=campus_MathMean mean=MathMean/autolabel;
run;

proc means data=District_RMTEI;
by school;
var ARMT_Read_&model_year;
output out=campus_ReadMean mean=ReadMean/autolabel;
run;

%macro LMM_datasets; /*creation of groups of students based on testing
histories*/
%do i = &j %to &k;
data LMM&grade&i ;
merge District_RMTEI campus_MathMean campus_ReadMean lunch.ses&model_year;
by school;
if &i ne indicator then delete;
if mteacher=' ' or rteacher=' ' then delete;
run;
%end;
%mend LMM_datasets;
%LMM_datasets

data SGP&grade; /*all students included without consideration of testing
histories*/
merge District_RMTEI campus_MathMean campus_ReadMean;
by school;
if mteacher=' ' or rteacher=' ' then delete;
run;
```

```

/*-----*/
/*Organize data to develop model from previous cohort of students*/

data Model_RMTEI;
set student.model_students&grade&year;
if &ryear2 ne ' ' and &myear2 ne ' ' and &ryear1 ne ' ' and &myear1 ne ' '
then indicator=2007;
else if &ryear2 ne ' ' and &myear2 ne ' ' then indicator=2008;
readgain=&ryear3-&ryear2;
mathgain=&myear3-&myear2;
run;

proc sort data=Model_RMTEI;
by school;
run;

%let prioryr_model = %eval(&model_year-1);
proc means data=Model_RMTEI;
by school;
var ARMT_Math_&prioryr_model;
output out=Modelcampus_MathMean mean=MathMean/autolabel;
run;

proc means data=Model_RMTEI;
by school;
var ARMT_Read_&prioryr_model;
output out=Modelcampus_ReadMean mean=ReadMean/autolabel;
run;

%macro model_datasets;
%do i = &j %to &k;
data model&grade&i;
merge Model_RMTEI Modelcampus_MathMean Modelcampus_ReadMean
lunch.ses&prioryr_model;
by school;
if &i ne indicator then delete;
if mteacher=' ' or rteacher=' ' then delete;
run;
%end;
%mend model_datasets;
%model_datasets

/*-----*/
/*Calculate LMM Teacher Effects for Math and Reading*/

%macro teacher_effects;
%do i = &j %to &k;
ods output solutionr=meffects&grade&i ;
proc mixed data=LMM&grade&i covtest ;
class mteacher school ;
model mathgain = ARMT_Math_&i-ARMT_Math_&k ARMT_Read_&k-ARMT_Read_&k /
solution ;
random int/ solution subject= mteacher(school);
run;

ods output solutionr=reffects&grade&i ;
proc mixed data=LMM&grade&i covtest ;

```

```

class rteacher school ;
model readgain = ARMT_Math_&k-ARMT_Math_&k ARMT_Read_&i-ARMT_Read_&k /
solution;
random int/ solution subject= rteacher(school);
run;

data mtcheffects&grade&i;
set meffects&grade&i;
MLMM_Teacher_Effect&grade&i=estimate;
DF&grade&i=df;
estimate&grade&i=MLMM_Teacher_Effect&grade&i*DF&grade&i;
run;

proc sort data=mtcheffects&grade&i;
by mteacher;
run;

data rtcheffects&grade&i;
set reffects&grade&i;
RLMM_Teacher_Effect&grade&i=estimate;
DF&grade&i=df;
estimate&grade&i=RLMM_Teacher_Effect&grade&i*DF&grade&i;
run;

proc sort data=rtcheffects&grade&i;
by rteacher;
run;
%end;
%mend teacher_effects;
%teacher_effects

%macro check;
    %if &grade= 4 %then %do;
data mtcheffects&grade;
merge mtcheffects&grade&k-mtcheffects&grade&k;
by mteacher;
Est_total=sum(of estimate&grade&k-estimate&grade&k);
DFtotal= sum(of DF&grade&k-DF&grade&k);
MLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then MLMM_Teacher_Effect=0;
run;

data rtcheffects&grade;
merge rtcheffects&grade&k-rtcheffects&grade&k;
by rteacher;
Est_total=sum(of estimate&grade&k-estimate&grade&k);
DFtotal= sum(of DF&grade&k-DF&grade&k);
RLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then RLMM_Teacher_Effect=0;
run;
    %goto exit;
    %end;

    %if &grade= 5 %then %do;
data mtcheffects&grade;
merge mtcheffects&grade&k_1-mtcheffects&grade&k;
by mteacher;

```

```

Est_total=sum(of estimate&grade&k_1-estimate&grade&k);
DFtotal= sum(of DF&grade&k_1-DF&grade&k);
MLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then MLMM_Teacher_Effect=0;
run;

data rtcheffects&grade;
merge rtcheffects&grade&k_1-rtcheffects&grade&k;
by rteacher;
Est_total=sum(of estimate&grade&k_1-estimate&grade&k);
DFtotal= sum(of DF&grade&k_1-DF&grade&k);
RLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then RLMM_Teacher_Effect=0;
run;

    %goto exit;
    %end;

    %if &grade=6 %then %do;
data Mtcheffects&grade;
merge mtcheffects&grade&k_2-mtcheffects&grade&k;
by mteacher;
Est_total=sum(of estimate&grade&k_2-estimate&grade&k);
DFtotal= sum(of DF&grade&k_2-DF&grade&k);
MLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then MLMM_Teacher_Effect=0;
run;

data rtcheffects&grade;
merge rtcheffects&grade&k_2-rtcheffects&grade&k;
by rteacher;
Est_total=sum(of estimate&grade&k_2-estimate&grade&k);
DFtotal= sum(of DF&grade&k_2-DF&grade&k);
RLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then RLMM_Teacher_Effect=0;
run;

    %goto exit;
    %end;

    %if &grade>=7 %then %do;
data Mtcheffects&grade;
merge mtcheffects&grade&j-mtcheffects&grade&k;
by mteacher;
Est_total=sum(of estimate&grade&j-estimate&grade&k);
DFtotal= sum(of DF&grade&j-DF&grade&k);
MLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then MLMM_Teacher_Effect=0;
run;

data rtcheffects&grade;
merge rtcheffects&grade&j-rtcheffects&grade&k;
by rteacher;
Est_total=sum(of estimate&grade&j-estimate&grade&k);
DFtotal= sum(of DF&grade&j-DF&grade&k);
RLMM_Teacher_Effect=Est_total/DFtotal;
If DFtotal = " " then RLMM_Teacher_Effect=0;
run;

    %goto exit;

```



```

        %end;

%exit: %mend check;
%check

/*-----*/
/*Calculate Median SGP for Math and Reading*/

%macro QR_SGP;
proc quantreg data=SGP&grade ci=resampling ;
effect sp=spline (ARMT_Math_&k ARMT_Read_&k /details );
model mathgain = sp/quantile=.01 .02 .03 .04 .05 .06 .07 .08 .09 .1
.11 .12 .13 .14 .15 .16 .17 .18 .19 .2
.21 .22 .23 .24 .25 .26 .27 .28 .29 .3
.31 .32 .33 .34 .35 .36 .37 .38 .39 .4
.41 .42 .43 .44 .45 .46 .47 .48 .49 .5
.51 .52 .53 .54 .55 .56 .57 .58 .59 .6
.61 .62 .63 .64 .65 .66 .67 .68 .69 .7
.71 .72 .73 .74 .75 .76 .77 .78 .79 .8
.81 .82 .83 .84 .85 .86 .87 .88 .89 .9
.91 .92 .93 .94 .95 .96 .97 .98 .99 ;
output out=chp3_2_3m&grade pred=p quantile=q ;
run;

data msgpbase;
set chp3_2_3m&grade;
keep mteacher school mathgain p1-p99;
run;

data mconvert;
set msgpbase;
array test{*}_numeric_;
DO i = 4 to dim(test);
test(i)=abs(test(i)-test(3));
end;
drop i;
mindev=min(of p1-p99);/* p1-p99 become the abs deviations due to array test*/
/* columns 4 102 */
run;

data msgp&grade;
set mconvert;
sgp=0;
array test{*}_numeric_;
do i=4 to 102;
if test{i}=mindev then sgp=i-3;/*will give the highest sgp if multiple values
are mindev*/
end;
if mindev= ' ' then sgp=' ' ;
drop i;
run;

/* Read SGP */
proc quantreg data=SGP&grade ci=resampling ;
effect sp=spline (ARMT_Math_&k ARMT_Read_&k /details );
model readgain = sp/quantile=.01 .02 .03 .04 .05 .06 .07 .08 .09 .1
.11 .12 .13 .14 .15 .16 .17 .18 .19 .2

```

```

.21 .22 .23 .24 .25 .26 .27 .28 .29 .3
.31 .32 .33 .34 .35 .36 .37 .38 .39 .4
.41 .42 .43 .44 .45 .46 .47 .48 .49 .5
.51 .52 .53 .54 .55 .56 .57 .58 .59 .6
.61 .62 .63 .64 .65 .66 .67 .68 .69 .7
.71 .72 .73 .74 .75 .76 .77 .78 .79 .8
.81 .82 .83 .84 .85 .86 .87 .88 .89 .9
.91 .92 .93 .94 .95 .96 .97 .98 .99 ;
output out=chp3_2_3r&grade pred=p quantile=q ;
run;

data rsgpbase;
set chp3_2_3r&grade;
keep rteacher school readgain p1-p99;
run;

data rconvert;
set rsgpbase;
array test{*}_numeric_;
DO i = 4 to dim(test);
test(i)=abs(test(i)-test(3));
end;
drop i;
mindev=min(of p1-p99);/* p1-p99 become the abs deviations due to array test*/
/*          4 102 */
run;

data rsgp&grade;
set rconvert;
sgp=0;
array test{*}_numeric_;
do i=4 to 102;
if test{i}=mindev then sgp=i-3;/*will give the highest sgp if multiple values
are mindev*/
end;
if mindev= ' ' then sgp=' ' ;
drop i;
run;
%mend QR_SGP;
%QR_SGP

data msgp;
set msgp&grade;
run;

proc sort data=msgp;
by mteacher;
run;

data rsgp;
set rsgp&grade;
run;

proc sort data=rsgp;
by rteacher;
run;

```

```

proc means data=msgp;
by mteacher;
var sgp;
output out=msgrowthper&grade median=MTeacher_Median_SGP n=Mnumber_students;
run;

proc means data=rsgp ;
by rteacher;
var sgp;
output out=rsgrowthper&grade median=RTeacher_Median_SGP n=Rnumber_students;
run;

/*-----*/
/*LMM Value added measure */

%macro LMM_model_build; /*creating the model from previous cohort of
students*/
%do i = &j %to &k ;
%let m= %eval(&i-1);
%let n=%eval(&k-1);

ods output solutionf=mfixed&grade&i ;
ods graphics on;
proc mixed data=model&grade&i boxplot covtest;
class mteacher school ;
model mathgain = ARMT_Math_&m-ARMT_Math_&n ARMT_Read_&n-ARMT_Read_&n
/solution ;
random int/subject= mteacher(school) solution;
run;
ods graphics off;

ods output solutionf=rfixed&grade&i;
proc mixed data=model&grade&i covtest;
class rteacher school ;
model readgain = ARMT_Math_&n-ARMT_Math_&n ARMT_Read_&m-ARMT_Read_&n
/solution;
random int/subject= rteacher(school) solution;
run;
%end;
%mend LMM_model_build;
%LMM_model_build

/* apply created models */
/* use data (indep vars) from cohort requiring predictions*/
%macro LMM_apply_model;
%if &grade= 4 %then %do;
%do i = &k %to &k ;
Proc IML;
Reset NoLog; /* send output to the listing file */
/* Read the data into the matrix X */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&i":"ARMT_Math_&k") into x2;
read all var("ARMT_Read_&k":"ARMT_Read_&k") into x3;
close LMM&grade&i;
x=x1||x2||x3;

```

```

use mfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close mfixed&grade&i;
pred=x*y;
create MLMMpredicted&grade&i var{pred};
append var{pred};
close MLMMpredicted&grade&i;
Reset log;

/* Reading LMM Value Added Measure */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&k":"ARMT_Math_&k") into x2;
read all var("ARMT_Read_&i":"ARMT_Read_&k") into x3;
close LMM&grade&i;
x=x1||x2||x3;
use rfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close rfixed&grade&i;
pred=x*y;
create RLMMpredicted&grade&i var{pred};
append var{pred};
close RLMMpredicted&grade&i;
Reset log;
Quit;

data chp3_2_2_1&grade&i;
merge LMM&grade&i MLMMpredicted&grade&i;
vam=mathgain-pred;
run;
data chp3_2_2_2&grade&i;
merge LMM&grade&i RLMMpredicted&grade&i;
vam=readgain-pred;
run;
%end;

data chp3_2_2_1;
set chp3_2_2_1&grade&k-chp3_2_2_1&grade&k;
run;

data chp3_2_2_2;
set chp3_2_2_2&grade&k-chp3_2_2_2&grade&k;
run;
%end;

%if &grade= 5 %then %do;
%do i = &k_1 %to &k ;
Proc IML;
Reset NoLog; /* send output to the listing file */
/* Read the data into the matrix X */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&i":"ARMT_Math_&k") into x2;
read all var("ARMT_Read_&k":"ARMT_Read_&k") into x3;
close LMM&grade&i;

```

```

x=x1||x2||x3;
use mfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close mfixed&grade&i;
pred=x*y;
create MLMMpredicted&grade&i var{pred};
append var{pred};
close MLMMpredicted&grade&i;
Reset log;

/* Reading LMM Value Added Measure */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&k":"ARMT_Math_&k") into x2;
read all var("ARMT_Read_&i":"ARMT_Read_&k") into x3;
close LMM&grade&i;
x=x1||x2||x3;
use rfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close rfixed&grade&i;
pred=x*y;
create RLMMpredicted&grade&i var{pred};
append var{pred};
close RLMMpredicted&grade&i;
Reset log;
Quit;

data chp3_2_2_1&grade&i;
merge LMM&grade&i MLMMpredicted&grade&i;
vam=mathgain-pred;
run;
data chp3_2_2_2&grade&i;
merge LMM&grade&i RLMMpredicted&grade&i;
vam=readgain-pred;
run;
%end;

data chp3_2_2_1;
set chp3_2_2_1&grade&k_1-chp3_2_2_1&grade&k;
run;

data chp3_2_2_2;
set chp3_2_2_2&grade&k_1-chp3_2_2_2&grade&k;
run;
%end;

%if &grade=6 %then %do;
%do i = &k_2 %to &k ;
Proc IML;
Reset NoLog; /* send output to the listing file */
/* Read the data into the matrix X */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&i":"ARMT_Math_&k") into x2;
read all var("ARMT_Read_&k":"ARMT_Read_&k") into x3;

```

```

close LMM&grade&i;
x=x1||x2||x3;
use mfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close mfixed&grade&i;
pred=x*y;
create MLMMpredicted&grade&i var{pred};
append var{pred};
close MLMMpredicted&grade&i;
Reset log;

/* Reading LMM Value Added Measure */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&k":"ARMT_Math_&k") into x2;
read all var("ARMT_Read_&i":"ARMT_Read_&k") into x3;
close LMM&grade&i;
x=x1||x2||x3;
use rfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close rfixed&grade&i;
pred=x*y;
create RLMMpredicted&grade&i var{pred};
append var{pred};
close RLMMpredicted&grade&i;
Reset log;
Quit;

data chp3_2_2_1&grade&i;
merge LMM&grade&i MLMMpredicted&grade&i;
vam=mathgain-pred;
run;
data chp3_2_2_2&grade&i;
merge LMM&grade&i RLMMpredicted&grade&i;
vam=readgain-pred;
run;
%end;

data chp3_2_2_1;
set chp3_2_2_1&grade&k_2-chp3_2_2_1&grade&k;
run;

data chp3_2_2_2;
set chp3_2_2_2&grade&k_2-chp3_2_2_2&grade&k;
run;
%end;

%if &grade>=7 %then %do;
%do i = &j %to &k ;
Proc IML;
Reset NoLog; /* send output to the listing file */
/* Read the data into the matrix X */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&i":"ARMT_Math_&k") into x2;

```

```

read all var("ARMT_Read_&k":"ARMT_Read_&k") into x3;
close LMM&grade&i;
x=x1||x2||x3;
use mfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close mfixed&grade&i;
pred=x*y;
create MLMMpredicted&grade&i var{pred};
append var{pred};
close MLMMpredicted&grade&i;
Reset log;

/* Reading LMM Value Added Measure */
use LMM&grade&i;
read all var{int} into x1;
read all var("ARMT_Math_&k":"ARMT_Math_&k") into x2;
read all var("ARMT_Read_&i":"ARMT_Read_&k") into x3;
close LMM&grade&i;
x=x1||x2||x3;
use rfixed&grade&i var{estimate}; /* coefficients calculated from model data
*/
read all var{estimate} into y;
close rfixed&grade&i;
pred=x*y;
create RLMMpredicted&grade&i var{pred};
append var{pred};
close RLMMpredicted&grade&i;
Reset log;
Quit;

data chp3_2_2_1&grade&i;
merge LMM&grade&i MLMMpredicted&grade&i;
vam=mathgain-pred;
run;

data chp3_2_2_2&grade&i;
merge LMM&grade&i RLMMpredicted&grade&i;
vam=readgain-pred;
run;
%end;

data chp3_2_2_1;
set chp3_2_2_1&grade&j-chp3_2_2_1&grade&k;
run;

data chp3_2_2_2;
set chp3_2_2_2&grade&j-chp3_2_2_2&grade&k;
run;
%end;
%mend LMM_apply_model;
%LMM_apply_model

proc sort data=chp3_2_2_1;
by mteacher;
run;

```

```

proc sort data=chp3_2_2_2;
by rteacher;
run;

proc means data = chp3_2_2_1;
by mteacher;
var vam;
output out=MLMMvam&grade mean=Overall_MLMM_Teacher_VAM ;
run;

proc means data = chp3_2_2_2;
by rteacher;
var vam;
output out=RLMMvam&grade mean=Overall_RLMM_Teacher_VAM;
run;

/*-----*/
/*QR value added measure */

%macro QR_VAM;
ods graphics on;
proc quantreg data=SGP&grade ci=resampling plots=all ;
effect sp=spline (ARMT_Math_&k ARMT_Read_&k /details);
model mathgain = sp/quantile=.5 diagnostics;
output out=chp3_2_4m&grade pred=p50 sresidual=sresid;
run;
ods graphics off;

proc quantreg data=SGP&grade ci=resampling ;
effect sp=spline (ARMT_Math_&k ARMT_Read_&k /details );
model readgain = sp/quantile=.5;
output out=chp3_2_4r&grade pred=p50;
run;

/*QR estimates */
data mqrestimate;
set chp3_2_4m&grade;
mqrvamstud=mathgain-p50;
run;

data rqrestimate;
set chp3_2_4r&grade;
rqrvamstud=readgain-p50;
run;
%mend QR_VAM;
%QR_VAM

proc sort data=mqrestimate;
by mteacher;
run;

proc sort data=rqrestimate;
by rteacher;
run;

proc means data = mqrestimate;
by mteacher;

```



```

var mqrvamstud;
output out=mqrteacher&grade mean=Overall_MQR_Teacher_VAM;
run;

proc means data = rqrestimate;
by rteacher;
var rqrvamstud;
output out=rqrteacher&grade mean=Overall_RQR_Teacher_VAM;
run;

/*-----*/
/* Chapter 3_2table creation */

data chp3_2mtable;
merge mtcheffects&grade MLMMvam&grade mqrteacher&grade msgrowthper&grade;
by mteacher;
keep mteacher school
MLMM_Teacher_Effect
Overall_MLMM_Teacher_VAM
MTeacher_Median_SGP
Overall_MQR_Teacher_VAM
Mnumber_students;
if Mnumber_students < 5 then delete;
run;

data chp3_2rtable;
merge rtcheffects&grade RLMMvam&grade rqrteacher&grade rsgrowthper&grade;
by rteacher;
keep rteacher school
RLMM_Teacher_Effect
Overall_RLMM_Teacher_VAM
RTeacher_Median_SGP
Overall_RQR_Teacher_VAM
Rnumber_students;
if Rnumber_students < 5 then delete;
run;

/*-----*/
/* Phase II and III for Math teachers */

/* Principal Component Analysis */
ods graphics on;
proc princomp data=chp3_2mtable out=mteacher_prn1 PLOTS=(SCORE(NCOMP=2
ellipse)patternprofile ) ;
var MLMM_Teacher_Effect
Overall_MLMM_Teacher_VAM
MTeacher_Median_SGP
Overall_MQR_Teacher_VAM;
run;
ods graphics off;

Proc IML;
Reset NoLog; /* send output to the listing file */
/* Read the data into the matrix X */
use mteacher_prn1 var{MLMM_Teacher_Effect
Overall_MLMM_Teacher_VAM
MTeacher_Median_SGP

```

```

Overall_MQR_Teacher_VAM};
read all var{MLMM_Teacher_Effect
Overall_MLMM_Teacher_VAM
MTeacher_Median_SGP
Overall_MQR_Teacher_VAM} into x;
close mteacher_prn1;
/* Number of Observations and Variables */
n=nrow(x);
p=ncol(x);
/* Compute sample mean, covariance, and inverse */
one=J(n,1,1);
xbar=(X`*one)/n;
print "Sample Means", xbar;
xstar=X-one*xbar`;
s=xstar`*xstar/(n-1);
print "Sample Covariance Matrix", s;
rho=corr(x);
print "Sample Correlation Matrix", rho;
detrho=det(rho);
print detrho;
call eigen(lambda,Evecs,rho);
print lambda, Evecs;
d=sqrt(diag(lambda));
print d;
/* compute correlation of metric with prin comp */
corr=evecs*d;
print corr;
Reset log;
Quit;

proc sort data=mteacher_prn1;
by prin1;
run;

Proc GPlot Data=mteacher_prn1;
Plot Prin1*Prin2=1 / HRef=0 VRef=0 VAxis=Axis1 HAxis=Axis2;
Axis1 Label=(A=90 "Principal Component 1");
Axis2 Label=("Principal Component 2");
Symbol1 C=Black V=Dot H=0.7 I=None PointLabel=(C=Black "#mteacher");
title "Mathematics Teachers";
Run; symbol1;Quit;
title;

/* Cluster Analysis */
/* Ward's Method */
ods graphics on;
proc cluster data=mteacher_prn1 outtree=mtree
method=ward plots=all;
id mteacher;
var prin1 prin2;
run;
ods graphics off;

%macro mcluster_check;
%do y = 4 %to 2 %by -1;
proc tree data=mtree noprint out=mout&y n=&y;
copy prin1 prin2;

```

```

run;

ods output CLDiffs=mbonci&y&grade;
proc glm data=mout&y;
  class cluster;
  model prin1 = cluster ;
  means cluster/bon cldiff;
run;
quit;

proc means data=mbonci&y&grade;
var significance;
output out=msigmean&y&grade mean=sigmean;
run;

data mcheck&y&grade;
set msigmean&y&grade;
if sigmean = 1 then CALL SYMPUT('mcheck',1) ;
run;

%if &mcheck =1 and &y =4 %then %do;
data appen3.mbonci&grade&year;
set mbonci&y&grade;
run;
%global m_numclust;
%let m_numclust=4;
%let mcheck=0;
%goto exit;
%end;

%if &mcheck=1 and &y =3 %then %do;
data appen3.mbonci&grade&year;
set mbonci&y&grade;
run;
%global m_numclust;
%let m_numclust=3;
%let mcheck=0;
%goto exit;
%end;

%if &mcheck=1 and &y=2 %then %do;
data appen3.mbonci&grade&year;
set mbonci&y&grade;
run;
%global m_numclust;
%let m_numclust=2;
%let mcheck=0;
%goto exit;
%end;

%if &mcheck=0 and &y =2 %then %do;
data appen3.mbonci&grade&year;
set mbonci&y&grade;
run;
%global m_numclust;
%let m_numclust=1;

```

```

%goto exit;
%end;

%end;

%exit: %mend mcluster_check;
%mcluster_check

proc tree data=mtree noprint out=mout n=&m_numclust;
copy prin1 prin2;
run;

proc sgplot data=mout;
scatter y=prin1 x=prin2 /group=Cluster ;
title "Ward Clustering of Mathematics Teachers ";
run;
title;

/* MANOVA, ANOVA, and Bon CI's to determine if clusters are statistically
different */
proc sort data=mout;
by _name_;
run;

data mward;
set mout;
ident=_n_;
keep _name_ cluster ident;
run;

proc sort data=mteacher_prn1;
by mteacher;
run;

data mteacher_prn3;
set mteacher_prn1;
ident=_n_;
run;

data mcombine;
merge mteacher_prn3 mward;
by ident;
mprin1=prin1;
mprin2=prin2;
mcluster=cluster;
run;

proc glm data=mcombine;
class cluster;
model prin1 prin2 = cluster;
means cluster/bon cldiff;
lsmeans cluster / out=Mmeansout;
manova h=cluster ;
run;
quit;

```

```

/*-----*/
/* Phase II and III for Reading teachers */

/* Principal Component Analysis */
ods graphics on;
proc princomp data=chp3_2rtable out=rteacher_prn1 plots=screes ;
var RLMM_Teacher_Effect
Overall_RLMM_Teacher_VAM
RTeacher_Median_SGP
Overall_RQR_Teacher_VAM;
run;
ods graphics off;

Proc IML;
Reset NoLog; /* send output to the listing file */
/* Read the data into the matrix X */
use rteacher_prn1 var{RLMM_Teacher_Effect
Overall_RLMM_Teacher_VAM
RTeacher_Median_SGP
Overall_RQR_Teacher_VAM};
read all var{RLMM_Teacher_Effect
Overall_RLMM_Teacher_VAM
RTeacher_Median_SGP
Overall_RQR_Teacher_VAM} into x;
close rteacher_prn1;
/* Number of Observations and Variables */
n=nrow(x);
p=ncol(x);
/* Compute sample mean, covariance, and inverse */
one=J(n,1,1);
xbar=(X`*one)/n;
print "Sample Means", xbar;
xstar=X-one*xbar`;
s=xstar`*xstar/(n-1);
print "Sample Covariance Matrix", s;
rho=corr(x);
print "Sample Correlation Matrix", rho;
detrho=det(rho);
print detrho;
call eigen(lambda,Evecs,rho);
print lambda, Evecs;
d=sqrt(diag(lambda));
print d;
/* compute correlation of metrics with prin comps */
corr=evecs*d;
print corr;
Reset log;
Quit;

proc sort data=rteacher_prn1;
by prin1;
run;

Proc GPlot Data=rteacher_prn1;
Plot Prin1*Prin2=1 / HRef=0 VRef=0 VAxis=Axis1 HAxis=Axis2;
Axis1 Label=(A=90 "Principal Component 1");
Axis2 Label=("Principal Component 2");

```

```

    Symbol1 C=Black V=Dot H=0.7 I=None PointLabel=(C=Black "#rteacher");
    title "Reading Teachers";
Run; Symbol1;
title;
Quit;

/* Cluster Analysis */
/* Ward's Method */
ods graphics on;
proc cluster data=rteacher_prn1 outtree=rtree
method=ward plots=all;
id rteacher;
var prin1 prin2;
run;
ods graphics off;

%macro rcluster_check;
%do y = 4 %to 2 %by -1;
proc tree data=rtree noprint out=rout&y n=&y;
copy prin1 prin2;
run;

ods output CLDiffs=rbonci&y&grade;
proc glm data=rout&y;
class cluster;
model prin1 = cluster ;
means cluster/bon cldiff;
run;
quit;

proc means data=rbonci&y&grade;
var significance;
output out=rsigmean&y&grade mean=sigmean;
run;

data rcheck&y&grade;
set rsigmean&y&grade;
if sigmean = 1 then CALL SYMPUT('rcheck',1) ;
run;

%if &rcheck =1 and &y =4 %then %do;
data appen3.rbonci&grade&year;
set rbonci&y&grade;
run;
%global r_numclust;
%let r_numclust=4;
%let rcheck=0;
%goto exit;
%end;

%if &rcheck=1 and &y =3 %then %do;
data appen3.rbonci&grade&year;
set rbonci&y&grade;
run;
%global r_numclust;
%let r_numclust=3;
%let rcheck=0;

```

```

%goto exit;
%end;

%if &rcheck=1 and &y=2 %then %do;
data appen3.rbonci&grade&year;
set rbonci&y&grade;
run;
%global r_numclust;
%let r_numclust=2;
%let rcheck=0;
%goto exit;
%end;

%if &rcheck=0 and &y =2 %then %do;
data appen3.rbonci&grade&year;
set rbonci&y&grade;
run;
%global r_numclust;
%let r_numclust=1;
%goto exit;
%end;

%end;

%exit: %mend rcluster_check;
%rcluster_check

proc tree data=rtree noprint out=rout n=&r_numclust;
copy prin1 prin2;
run;

/* MANOVA, ANOVA, and Bon CI's to determine if clusters are statistically
different */
proc sort data=rout;
  by _name_;
  run;

data rward;
set rout;
ident=_n_;
keep _name_ cluster ident;
run;

proc sort data=rteacher_prn1;
by rteacher;
run;

data rteacher_prn3;
set rteacher_prn1;
ident=_n_;
run;

data rcombine;
  merge rteacher_prn3 rward;
  by ident;
  rprin1=prin1;

```

```

rprin2=prin2;
rcluster=cluster;
run;

proc glm data=rcombine;
  class cluster;
  model prin1 prin2 = cluster ;
  means cluster/bon cldiff;
  lsmeans cluster / out=Rmeansout;
  manova h=cluster ;
  run;
  quit;

/*-----*/
/* combining prin1 from both math and reading processes to calculate an
overall value*/

data RMTEI;
merge mcombine rcombine;
by _name_;
if mprin1 = " " then IndexValue = rprin1;
else if rprin1= " " then IndexValue = mprin1;
else IndexValue=(mprin1+rprin1)/2;
keep _name_ school mprin1 mprin2 rprin1 rprin2 indexvalue
MLMM_Teacher_Effect Overall_MLMM_Teacher_VAM MTeacher_Median_SGP
Overall_MQR_Teacher_VAM
RLMM_Teacher_Effect Overall_RLMM_Teacher_VAM RTeacher_Median_SGP
Overall_RQR_Teacher_VAM
mteacher rteacher mcluster rcluster Mnumber_students Rnumber_students;
run;

/*Cluster Analysis of overall value */
/* Ward's Method */
ods graphics on;
proc cluster data=RMTEI outtree=tree
method=ward plots=all;
id _name_;
var IndexValue;
run;
ods graphics off;

%macro ivcluster_check;
%do y = 4 %to 2 %by -1;
proc tree data=tree noprint out=ivout&y n=&y;
copy indexvalue;
run;

ods output CLDiffs=ivbonci&y&grade;
proc glm data=ivout&y;
  class cluster;
  model indexvalue = cluster ;
  means cluster/bon cldiff;
run;
quit;

proc means data=ivbonci&y&grade;
var significance;

```



```

output out=ivsigmean&y&grade mean=sigmean;
run;

data ivcheck&y&grade;
set ivsigmean&y&grade;
if sigmean = 1 then CALL SYMPUT('ivcheck',1) ;
run;

%if &ivcheck =1 and &y =4 %then %do;
data appen3.ivbonci&grade&year;
set ivbonci&y&grade;
run;
%global iv_numclust;
%let iv_numclust=4;
%let ivcheck=0;
%goto exit;
%end;

%if &ivcheck=1 and &y =3 %then %do;
data appen3.ivbonci&grade&year;
set ivbonci&y&grade;
run;
%global iv_numclust;
%let iv_numclust=3;
%let ivcheck=0;
%goto exit;
%end;

%if &ivcheck=1 and &y=2 %then %do;
data appen3.ivbonci&grade&year;
set ivbonci&y&grade;
run;
%global iv_numclust;
%let iv_numclust=2;
%let ivcheck=0;
%goto exit;
%end;

%if &ivcheck=0 and &y =2 %then %do;
data appen3.ivbonci&grade&year;
set ivbonci&y&grade;
run;
%global iv_numclust;
%let iv_numclust=1;
%goto exit;
%end;

%end;

%exit: %mend ivcluster_check;
%ivcluster_check

proc tree data=tree noprint out=out n=&iv_numclust;
copy IndexValue;
run;

proc sgplot data=out;

```

```

scatter y=indexvalue x=_name_ / group=cluster;
axis label="Teacher";
title "Ward's Cluster Analysis for Teachers by Overall Index Value";
run;

Proc GPlot Data=out;
  Plot indexvalue*_name_=cluster/HRef=0 VRef=0 VAxis=Axis1 HAxis=Axis2 ;
  Axis1 Label=(A=90 "Index Value");/*A=90 rotates title 90 degrees*/
  Axis2 Label=("Teacher");
  Symbol H=1 PointLabel=(C=Black "#_name_");
Run;
Symbol1;
title;
Quit;

/* MANOVA, ANOVA, and Bon CI's to determine if clusters are statistically
different */
proc glm data=out;
  class cluster;
  model Indexvalue = cluster ;
  means cluster/bon cldiff;
  lsmeans cluster / out=meansout;
  manova h=cluster ;
  run;
  quit;

/*-----*/
/* create final file */

proc sort data=out;
by _name_;
run;

data final;
merge RMTEI out;
by _name_;
drop clusname;
teacher=input(_name_,best12.);
run;

data appen3.appen3&grade&year;
set final;
run;

%MEND RMTEI;

%macro reports;
%do p = 4 %to 8;
%RMTEI(grade=&p);
%end;
%mend reports;

%reports

```

Appendix 3: Alabama District Results

2009:

Teacher	School	grade	MLMM Teacher Effect	Overall MLM Teacher VAM	Overall MQR Teacher VAM	Mteacher Median SGP	Number students	mpin1	mpin2	ndcluster	RLMM Teacher Effect	Overall RLMM Teacher VAM	RQR Teacher VAM	Rteacher Median SGP	Rnumber students	rpim1	rpim2	rdcluster	Overall RMTET	Overall Cluster	
7639	10	4	0.4967	13.1583	3.6109	51.5	18	0.3079	-0.0492	3	0.0000	3.4042	-4.1343	29.5	18	-1.1642	-1.0298	4	-0.9281	4	
5132	10	4	4.6026	16.9293	6.0198	61.5	24	1.1003	0.1508	4	0.0000	10.6220	2.1259	51.5	24	1.2701	0.0315	3	1.1852	2	
1714	10	4	14.1408	28.3909	17.0350	76.0	22	3.0923	-0.0242	2	0.0000	10.8276	0.6914	48.0	22	0.7264	-0.1063	3	1.9044	2	
8674	15	4	-12.7435	-4.1422	-15.1236	31.0	21	-2.6685	0.2916	1	0.0000	6.8107	-2.4475	43.0	21	-0.6611	-0.0555	2	-1.6148	3	
7642	15	4	-9.9282	-1.5707	-12.3996	28.0	19	-2.3511	-0.0876	1	0.0000	14.1804	5.1728	54.0	19	2.4138	-0.2682	1	0.0313	3	
3007	15	4	-2.5479	8.7008	-3.1471	40.5	22	-0.6870	-0.2086	3	0.0000	8.9883	-0.1006	61.5	22	1.2367	1.4085	3	0.2749	1	
5518	15	4	-6.9544	3.2270	-9.6157	35.0	23	-1.6379	-0.0302	1	0.0000	4.2631	-5.5102	39.0	23	-1.6645	0.0495	4	-1.6512	3	
8057	20	4	-0.9140	11.0948	-1.6215	51.0	15	-0.1194	0.1579	3	0.0000	4.9277	-3.6708	51.0	15	-0.6729	1.0005	2	-0.3461	4	
7468	20	4	-6.0218	5.0619	-8.9916	36.0	21	-1.4394	-0.0751	1	0.0000	1.6614	-7.1928	39.0	21	-2.3517	0.3984	4	-1.8955	3	
4203	20	4	5.4203	19.8034	7.0542	71.5	16	1.6122	0.5215	4	0.0000	4.9966	-3.5385	42.0	16	-1.0756	0.0731	4	0.2683	1	
5064	25	4	15.9662	30.5511	17.7457	77.0	23	3.3457	-0.1088	2	0.0000	9.1029	-0.3729	50.0	23	0.5328	0.2768	3	1.9393	2	
8815	25	4	10.0300	23.9038	14.1974	63.0	21	2.1481	-0.3429	4	0.0000	6.2688	3.5508	36.0	21	0.1127	-1.5569	2	1.1304	2	
1433	25	4	-4.5141	5.4393	-4.2633	44.0	18	-0.9020	0.1399	3	0.0000	2.7168	-2.6217	46.5	18	-0.4659	0.3472	2	-0.6839	4	
4726	25	4	-7.0327	3.3654	-9.3174	29.0	24	-1.7912	-0.3350	1	0.0000	14.7273	6.1868	52.5	24	2.9554	-0.5699	1	0.3861	1	
1743	40	5	-12.4500	-12.3092	-11.5660	28.5	46	-4.0796	-0.1395	3	0.8938	5.1915	-0.3518	50.5	22	0.1325	-0.0874	2	-1.8735	1	
2406	40	5									-1.6561	-2.4985	-2.6413	39.0	19	-0.9900	-0.1999	1	-0.9900	1	
7314	40	5									-3.8120	-1.4316	-3.7961	38.0	19	-1.1763	-0.1187	1	-1.1763	1	
2481	40	5	-2.1257	-1.1250	0.0494	52.5	44	-0.3461	0.3520	2	-4.7025	-0.8579	-1.3182	46.0	25	-0.8070	0.3522	1	-0.5766	1	
6060	40	5									6.6404	8.7742	4.9463	46.0	14	0.9228	-0.7854	2	0.9228	2	
3000	40	5									-4.9803	-0.8310	-1.5714	40.5	22	-1.0108	0.1205	1	-1.0108	1	
4123	40	5	4.6392	6.3779	5.9786	60.0	46	1.5894	0.0666	1	-5.7471	-2.0762	-4.9614	35.5	24	-1.4984	-0.0714	1	0.0455	2	
7655	40	5									7.9057	12.6972	7.0954	70.0	21	2.1154	0.1203	3	2.1154	2	
8674	40	5	1.7292	2.7621	3.9916	55.0	39	0.6432	0.1096	1	-5.1698	-2.5116	-3.0475	44.0	21	-1.1035	0.3187	1	-0.2302	2	
8424	40	5	4.4194	6.8033	3.8514	57.0	46	1.2682	-0.2538	1	10.6475	16.9835	10.8908	66.0	24	2.6628	-0.3520	3	1.9655	3	
5370	40	5	-1.0261	-1.7472	0.0117	48.0	35	-0.5141	0.0552	2	1.5989	5.9366	4.5287	56.0	21	0.6929	0.1786	2	0.0894	2	
9080	40	5	4.8140	6.7376	5.3694	57.0	41	1.4390	-0.1900	1	-10.7526	-9.1231	-11.3459	29.5	22	-2.8322	0.1722	1	-2.8322	1	
7065	40	5									13.5694	21.0331	21.5584	87.0	20	4.4038	0.4443	3	2.9214	3	
6480	40	5									-5.1736	-2.6014	-5.5511	36.0	23	-1.5100	-0.0929	1	-1.5100	1	
3566	40	6	0.7075	4.5320	-0.7756	48.5	18	0.5129	0.8792	1	-0.9020	4.4050	-0.7633	48.5	66	-0.4564	0.4427	1	-0.4564	1	
7148	40	6									-1.8376	1.3378	-5.5482	32.0	26	-2.2945	-0.2230	1	-0.8908	1	
5306	40	6									-1.0775	2.5172	-1.5207	38.0	67	-1.2614	-0.1056	1	-1.2614	1	
2080	40	6	3.1228	3.5709	4.6340	60.5	80	2.3247	-0.8207	1	1.2735	8.1937	4.5100	57.5	58	1.5520	0.1792	2	2.3247	2	
9604	40	6																			
2535	40	6	-1.8471	-1.9796	-0.8787	49.0	101	-0.8660	-0.7297	1	2.5436	10.8772	6.2206	58.5	68	2.4602	-0.2931	2	2.4602	2	
8259	40	6	-1.7539	-2.2177	-4.1372	37.0	82	-1.9716	0.1712	1	-0.8078	2.4840	-4.4474	42.0	80	-0.3644	-0.3985	1	-0.3644	1	
6439	35	7									0.5283	-3.7792	-3.1876	30.0	10	-1.0445	1.0493	1	-1.0445	1	
8607	35	7									3.4678	8.3076	2.6144	58.0	105	2.6713	0.0317	2	2.6713	2	
4876	35	7																			
2485	35	7	-2.8860	-2.8789	-5.6153	40.0	69	-1.4257	0.1039	1											
8125	35	7	5.3149	9.2151	5.5016	66.0	79	2.2821	0.0189	1											
6895	35	7	-2.4290	-0.3826	-3.9052	45.0	109	-0.8563	-0.1228	1											
9595	35	7																			
9473	30	8	0.0000	2.0157	-0.8140	49.0	114	-1.6292	0.0686	2	-3.1418	-2.2713	-4.7476	42.0	58	-1.2624	-0.6876	1	-1.2624	1	
3360	30	8									0.3479	2.8279	2.6676	65.0	109	1.5481	0.6148	1	1.5481	1	
6057	30	8	0.0000	3.2937	-0.2685	50.0	107	0.8985	0.9650	1											
4073	30	8									-0.5727	-0.3871	-2.0676	44.0	62	-0.3811	-1.7008	2	-0.3811	2	
8607	30	8									-0.0040	0.1459	-2.1727	50.0	8	0.4262	-0.1213	1	0.4262	1	
8258	30	8									-0.0673	-11.9186	-16.1028	19.0	6	-3.3193	0.8351	3	-3.3193	3	
7021	30	8									0.3811	1.3656	0.5925	58.5	52	1.2730	0.8330	1	1.2730	1	
7241	30	8									-0.1977	-2.7402	-5.7682	32.0	17	-1.0890	-0.3322	2	-1.0890	2	
5512	30	8	0.0000	0.6532	-0.3154	50.0	49	0.7307	-1.0336	1	0.1127	3.5568	3.4835	54.5	18	1.5412	-0.1285	1	1.5412	1	
1470	30	8																			

Teacher	School	grade	MLMM Teacher Effect	Overall MLM Teacher VAM	Overall MQR Teacher VAM	Mteacher Median SGP	Mnumber students	mprin1	mprin2	mcluster	RLMM Teacher Effect	Overall RLM Teacher VAM	RQR Teacher Median SGP	Rnumber students	rprin1	rprin2	roluster	Overall RMTI	Overall Cluster			
7539	10	4	1.3185	-0.0511	2.5530	56.5	16	0.7081	0.3148	1	3.0444	2.7414	8.5573	66.0	16	2.0895	0.2618	3	1.9388	3		
5132	10	4	6.6054	8.1194	10.3425	56.0	19	2.2178	-0.8271	1	0.4579	-2.5916	1.7753	40.0	19	-0.1321	-0.6740	1	1.0428	3		
1714	10	4	11.2890	15.1108	16.6056	74.5	18	4.2248	0.0932	4	9.2737	13.2909	18.7579	82.0	18	5.1338	-0.0124	4	4.6793	4		
8574	15	4	-6.2063	-11.7158	-10.3029	29.0	21	-2.6457	-0.4508	3	-3.2460	-9.6807	-6.7773	39.0	21	-1.8866	0.2219	2	-2.2661	4		
7542	15	4	1.8052	0.1508	1.5895	47.5	20	0.8562	-0.3420	2	-0.3711	-5.3977	-2.4920	57.0	20	-0.1776	0.8245	1	0.0898	1		
3007	15	4	4.1263	4.1618	4.3918	59.0	21	1.4536	0.1150	1	2.3889	-0.3967	3.7002	52.0	21	0.8688	-0.1911	1	1.1712	3		
5518	15	4	-5.5743	-11.0064	-5.6527	41.0	17	-1.7678	0.2399	3	-5.8831	-16.0595	-8.5903	34.0	17	-3.0639	0.3594	2	-2.4183	2		
8057	20	4	-0.4800	-3.5561	1.5243	48.0	16	-0.0737	-0.0172	2	0.4145	-1.5484	2.9476	55.0	16	0.5666	0.1816	1	0.2464	1		
7458	20	4	-0.0818	-2.3602	-1.7856	34.0	19	-0.7233	-1.0168	2	0.9710	-1.0912	0.4310	59.0	18	0.6358	0.5562	1	-0.0438	1		
4203	20	4	1.3113	-0.7278	3.1059	56.0	16	0.6840	0.2958	1	-3.8376	-10.9829	-5.1867	31.5	16	-2.2187	-0.2888	2	-0.7673	1		
5064	25	4	1.0303	-0.5016	2.1288	59.0	22	0.7205	0.5362	1	0.1252	-3.5556	1.9464	46.5	22	-0.0079	-0.2183	1	0.3563	1		
8815	25	4	1.8986	0.2988	3.0334	57.0	21	0.8364	0.2912	1	0.2099	-2.8717	2.1352	46.0	21	0.0527	-0.2900	1	0.4445	1		
1433	25	4	-7.1228	-13.6317	-9.7744	43.0	17	-2.2769	0.6641	3	-3.0694	-8.2037	-3.4866	36.0	17	-1.6085	-0.2387	2	-1.9427	2		
4726	25	4	-8.7883	-15.9358	-11.4476	33.0	24	-3.0736	0.1825	3	-1.2480	-4.8038	-1.3182	46.0	24	-0.8650	0.0638	1	-1.8193	2		
5629	25	4	-1.1811	-4.6446	-0.9878	41.0	20	-0.6406	-0.3798	2	0.7700	-1.5127	3.6685	44.5	20	0.2930	-0.5560	1	-0.1713	1		
1743	40	5									-0.8234	-0.0993	1.8506	47.0	20	-0.9394	-0.9382	1	-0.9394	1		
7492	40	5									-0.8501	3.2332	-0.6291	45.0	78	-1.0443	0.1639	1	-1.0443	1		
2481	40	5	-4.7780	6.9934	-6.2511	38.5	82	-1.7320	-0.0652	1												
4123	40	5	6.2272	20.6299	6.8853	62.0	79	2.1862	-0.0316	1												
8424	40	5																				
6480	40	5																				
3126	40	5	-1.4492	12.1138	-1.6726	45.0	85	-0.4542	0.0968	1												
3566	40	6																				
5306	40	6																				
2080	40	6	1.6310	5.5447	1.7236	50.0	95	0.9049	-0.0535	1												
9604	40	6																				
2535	40	6	-3.1576	-1.0494	-3.6936	41.5	90	-2.2918	0.0070	1												
8259	40	6	2.2804	6.4883	2.8580	51.0	94	1.3869	0.0465	1												
3373	40	6																				
6439	35	7																				
4875	35	7																				
8125	35	7	4.9405	3.5552	4.6021	59.0	118	0.8372	-0.1374	1												
6835	35	7	-11.2038	-12.5985	-9.6574	35.0	115	-2.2782	0.0223	1												
7065	35	7	6.2632	9.0737	6.9627	64.0	67	1.4409	0.1151	1												
9252	35	7																				
6067	30	8	-0.5001	-0.2802	-1.3107	42.0	97	-0.8412	-0.5723	1												
4073	30	8																				
6380	30	8	3.4989	4.8387	4.0585	55.0	102	2.2046	0.0838	1												
7021	30	8																				
7241	30	8																				
5512	30	8	-2.9987	-2.8312	-1.8731	46.0	83	-1.3634	0.4885	1												
1470	30	8																				

Appendix 4: NCES District Results

2006:

Teacher	School	grade	MLMM Teacher Effect	Overall MLMM Teacher VAM	Overall MQR Teacher VAM	Overall RQR Teacher VAM	RLMM Teacher Effect	RLMM Teacher VAM	Overall RLM Teacher VAM	Overall Teacher VAM	Rteacher Median SGP	Number students	rprin1	rprin2	rdcluster	Overall RIMTEI	Overall Cluster
		3	-0.4262	-0.5441	-0.5705	20.0	-0.2076	-0.2529	-0.3000	41.0	26	-1.9219	0.3882	2	-3.1916	2	
		3	0.1872	0.3062	0.2509	83.0	-0.0384	-0.0574	-0.1288	51.0	11	-0.5095	0.3430	2	0.6206	1	
		3	-0.0734	-0.1076	-0.1767	43.0	-0.1243	-0.1694	-0.2235	50.0	15	-1.3855	0.6155	2	-1.1691	3	
		3	0.2398	0.3687	0.3763	72.0	0.1021	0.1448	0.0917	59.0	13	0.8772	0.1801	1	1.7098	4	
		3	-0.1868	-0.2555	-0.1934	39.0	0.1183	0.1522	0.1631	67.0	19	1.3155	0.5858	1	-0.3106	1	
		3	-0.0658	-0.0949	-0.0772	42.0	-0.3023	-0.4670	-0.4629	23.5	10	-3.4022	-0.0991	3	-2.1830	2	
		3	-0.1455	-0.2181	-0.1821	41.5	-0.0417	-0.0579	-0.1036	39.5	14	-0.8267	-0.3237	2	-1.2411	3	
		3	-0.1026	-0.1292	-0.0669	40.0	-0.1425	-0.1712	-0.2022	42.0	27	-1.3936	0.1750	2	-1.2940	3	
		3	-0.0651	-0.0645	-0.0537	46.0	0.1137	0.1288	0.0997	54.0	41	0.7367	-0.1129	1	0.0031	1	
		3	0.0196	0.0272	0.0122	58.5	0.1674	0.2180	0.1437	58.0	18	1.2406	-0.0973	1	0.7040	1	
		3	0.3257	0.4596	0.4556	79.0	0.1184	0.1563	0.0751	54.0	17	0.7516	-0.1247	1	2.0159	4	
		3	0.1251	0.2126	0.0521	39.5	0.1311	0.2025	0.0264	61.0	10	0.9912	0.2501	1	0.6294	1	
		3	0.1207	0.1629	0.2177	72.5	-0.0153	-0.0195	-0.0162	52.0	20	-0.1364	0.2256	2	0.7033	1	
		3	-0.1756	-0.2574	-0.2446	46.0	-0.2011	-0.2740	-0.3052	22.0	15	-2.5437	-0.6711	3	-2.1666	2	
		3	0.0277	0.0351	-0.0023	50.5	-0.1408	-0.1703	-0.2122	44.0	26	-1.3450	0.2935	2	-0.7125	3	
		3	-0.1063	-0.1682	-0.1484	34.0	0.0381	0.0442	-0.0074	56.0	19	1.4749	-0.8327	2	-0.2005	1	
		3	0.0194	0.0234	0.0912	47.5	0.0381	0.0442	-0.0074	56.0	34	-0.3675	-0.8327	2	-0.2005	1	
		3	-0.0193	-0.0257	-0.0669	49.0	0.2054	0.2687	0.2287	69.0	21	1.9125	0.3487	1	0.7092	1	
		3	0.1779	0.2483	0.2592	67.0	0.4382	0.5688	0.4854	70.0	19	3.5670	-0.4879	4	2.6593	4	
		3	0.3928	0.5374	0.4975	67.0	0.4386	0.5617	0.4886	78.0	19	3.8134	-0.0342	4	3.5786	4	
		3	-0.1766	-0.2574	-0.2446	46.0	-0.2011	-0.2740	-0.3052	22.0	15	-2.5437	-0.6711	3	-2.1666	2	
		3	0.0277	0.0351	-0.0023	50.5	-0.1408	-0.1703	-0.2122	44.0	26	-1.3450	0.2935	2	-0.7125	3	
		4	-0.1063	-0.1682	-0.1484	34.0	0.0381	0.0442	-0.0074	56.0	19	1.4749	-0.8327	2	-0.2005	1	
		4	-0.0957	-0.4887	-0.3085	32.0	-0.1936	-0.6478	-0.5262	7.0	5	-3.8640	-0.3391	4	-3.4773	3	
		4	-0.0731	-0.1565	-0.1366	40.0	-0.0982	-0.1622	-0.1828	29.5	18	-1.3527	0.0213	2	-1.4538	2	
		4	0.0021	0.0038	0.0683	42.5	0.0275	0.0410	0.0395	47.0	24	0.4155	-0.0701	1	0.1920	1	
		4	-0.0048	-0.0090	-0.0171	59.5	0.0469	0.0699	0.0923	51.0	20	0.7667	-0.0474	1	0.5036	1	
		4	0.0223	0.0453	0.0473	50.5	0.0375	0.0595	0.0300	46.5	24	0.4705	-0.1386	1	0.4422	1	
		4	-0.0663	-0.1230	-0.0840	32.0	-0.1150	-0.1712	-0.1786	41.5	24	-1.1047	0.5676	2	-1.3170	2	
		4	0.0369	0.0614	0.1019	60.0	0.0524	0.0722	0.0288	49.0	31	0.6208	-0.1271	1	0.8153	1	
		4	0.0194	0.0428	0.0976	57.0	0.0204	0.0345	0.0581	57.0	17	0.6925	0.3351	1	0.7137	1	
		4	0.1442	0.3910	0.4086	75.0	0.2063	0.4079	0.4867	72.5	12	3.3553	-0.2017	3	3.5614	4	
		5	-0.4147	-0.4738	-0.4825	24.5	-0.3709	-0.4145	-0.4510	20.5	42	-3.8552	0.3210	3	-3.9063	3	
		5	-0.0253	-0.0438	-0.0001	43.0	-0.2157	-0.3342	-0.1836	33.0	9	-2.3288	0.4155	3	-1.3372	2	
		5	0.1942	0.2454	0.2178	62.0	0.2039	0.2442	0.2307	67.0	25	1.8117	0.3593	1	1.8923	1	
		5	-0.0917	-0.1258	-0.1149	37.5	-0.0673	-0.0858	-0.1626	34.5	18	-1.8554	-0.1270	2	-1.3102	2	
		5	0.0623	0.0785	0.0683	57.0	0.2523	0.2952	0.2721	61.0	29	1.9564	-0.1278	1	1.3650	1	
		5	0.3295	0.3768	0.3089	67.0	0.2294	0.2541	0.1749	53.0	46	1.3689	-0.3784	1	2.1714	1	
		5	0.0629	0.0847	0.1019	39.0	0.0224	0.0282	-0.0012	33.0	19	-0.5824	-0.6178	2	-0.2727	2	
		5	-0.0361	-0.1249	-0.1414	41.5	0.0561	0.0687	0.0531	48.0	21	0.1388	-0.0003	2	-0.4616	2	
		5	-0.1509	-0.1963	-0.2232	33.0	0.0646	0.0798	0.0626	38.0	22	-0.0851	-0.5470	2	-0.9600	2	
		5	-0.0806	-0.1177	-0.1375	49.5	0.0364	0.0446	-0.0121	51.0	22	-0.0177	0.2609	2	-0.3310	2	
		5	0.0027	0.0034	-0.0159	42.0	0.2476	0.3008	0.2385	62.0	23	1.9109	-0.0560	1	0.8336	1	
		5	-0.1823	-0.2369	-0.2671	46.5	-0.2297	-0.2814	-0.3197	23.0	22	-2.8445	-0.0625	3	-2.2505	2	
		5	0.0098	0.0116	-0.0258	45.0	0.0325	0.0371	0.0015	44.0	35	-0.2151	-0.0876	2	-0.1700	2	
		5	0.2567	0.3374	0.3208	71.0	0.1263	0.1548	0.1163	57.0	21	0.8890	0.1801	1	1.8789	1	
		5	0.2479	0.3339	0.2714	68.0	0.3316	0.4179	0.3772	80.0	19	3.1886	0.4685	1	2.9070	1	

2007:

Teacher	School	grade	MLMM Teacher Effect	Overall MLM Teacher VAM	MQR Teacher VAM	Overall Mteacher Median SGP	Mnumber students	mpim1	mpim2	mcuster	RLMM Teacher Effect	Overall RLM Teacher VAM	RQR Teacher VAM	Rteacher Median SGP	Rnumber students	rpm1	rpm2	rcuster	Overall RMTI	Overall Cluster
3			0.7090	0.5992	0.7419	91.0	17	4.0587	0.1473	4	0.4672	0.6224	0.5413	87.0	17	4.6370	0.3657	4	4.3478	4
3			-0.0631	-0.3848	-0.0878	48.0	16	-0.4275	0.1254	2	-0.0854	-0.3712	-0.1810	37.0	16	-1.1076	0.2536	1	-0.7675	2
3			0.5106	0.3911	0.5999	73.5	12	2.9047	0.3635	4	0.2264	0.1820	0.2065	68.5	12	2.2768	0.4487	3	2.5902	3
3			-0.0492	-0.2909	-0.0487	62.0	15	0.1451	-0.2422	3	-0.0270	-0.2318	-0.0071	67.0	15	0.5276	-0.0974	2	0.3364	2
3			0.1325	-0.0924	0.1067	55.5	20	0.6898	0.2724	3	0.0372	-0.1267	0.0202	51.0	20	0.4050	0.2743	2	0.5474	2
3			-0.2868	-0.7017	-0.3091	37.0	18	-1.6380	-0.0731	1	-0.2409	-0.6868	-0.4481	26.0	13	-2.8414	0.1669	1	-2.2397	1
3			-0.4338	-0.8684	-0.4644	19.5	20	-2.7100	0.2707	1	-0.1512	-0.5173	-0.2207	43.0	20	-1.3912	-0.0863	1	-2.0506	1
3			-0.1340	-0.7132	0.0246	53.0	9	-0.5809	-0.7364	2	-0.0492	-0.5307	0.1097	51.0	9	-0.2907	-0.9195	2	-0.4358	2
3			-0.3865	-0.6410	-0.5376	25.0	14	-2.3210	0.5043	1	-0.2896	-0.6961	-0.5191	30.0	14	-2.8319	0.5040	1	-2.5764	1
3			0.3492	0.1235	0.4492	56.0	9	1.7287	0.4989	3	0.1438	0.0131	0.2216	62.0	9	1.6174	0.0670	3	1.6780	3
3			-0.1419	-0.7953	0.0003	46.0	31	-0.8896	-0.6406	2	0.1214	-0.3975	0.1977	54.0	31	0.6240	-0.2162	2	-0.1328	2
3			0.0732	-0.2469	0.0977	62.0	20	0.5748	-0.2084	3	0.0455	-0.1971	0.1108	61.0	20	0.7976	-0.2031	2	0.6662	2
3			-0.3084	-0.6768	-0.3584	27.5	14	-1.9425	0.2903	1	-0.1085	-0.4170	-0.1985	39.0	14	-1.2083	0.1570	1	-1.5754	1
3			0.1490	-0.3509	0.2528	66.0	19	0.9005	-0.5376	3	0.0040	-0.3983	0.0402	52.0	19	-0.0374	-0.4536	2	0.4316	2
3			-0.3904	-0.6921	-0.3803	30.0	20	-2.0513	0.1682	1	-0.2123	-0.5156	-0.3251	39.0	20	-1.8662	0.1877	1	-1.9688	1
3			0.2708	-0.0526	0.2614	71.0	20	1.5684	-0.2027	3	0.1188	-0.1169	-0.1542	48.0	20	0.6900	0.1515	2	1.1242	3
4			0.0625	-0.1068	0.0936	63.0	23	1.0134	-0.7715	3	-0.0718	-0.5824	-0.1542	31.0	23	-1.6610	-0.3276	2	-0.3238	1
4			-0.1850	-0.2767	-0.1892	31.0	24	-2.0150	0.3647	1	0.0707	-0.1869	0.0652	54.0	24	1.1895	-0.1400	1	-0.4127	1
4			-0.2291	-0.3860	-0.2413	29.0	24	-2.5663	0.1683	1	-0.2079	-0.6742	-0.3390	28.0	24	-3.0573	-0.0849	2	-2.8118	3
4			0.2496	0.2091	0.2906	65.0	17	2.6490	0.0612	2	-0.1553	-0.6869	-0.2874	42.0	17	-2.2062	0.6897	2	0.2214	1
4			-0.0727	-0.3335	-0.1538	39.5	22	-1.4127	-0.1246	1	0.0706	-0.3781	0.0973	60.5	22	1.1313	0.7128	1	-0.1407	1
4			-0.0833	-0.2395	-0.0802	39.5	22	-1.0602	0.0492	1	0.0221	-0.4015	-0.0391	37.0	22	-0.3651	-0.5526	2	-0.7076	1
4			-0.0481	-0.1730	0.0102	53.5	6	-0.1373	-0.3687	3	0.0643	-0.1742	0.1178	63.0	6	1.6602	0.3769	1	0.7615	1
4			0.0958	-0.0222	0.1176	52.0	20	0.8473	0.0295	3	0.1584	-0.1472	0.1892	60.0	20	2.2112	-0.0095	1	1.5293	2
4			0.2103	0.3961	0.1543	71.0	19	2.6817	0.5919	2	0.0490	-0.0339	-0.0084	51.0	19	1.0873	-0.6648	1	1.8845	2
5			-0.0468	-0.3587	-0.4261	44.0	15	-2.5764	0.6233	1	-0.1013	-0.6504	-0.2437	27.0	15	-2.4932	-0.3109	1	-2.5343	3
5			-0.0028	-0.0306	-0.0792	55.0	14	0.1418	0.1547	1	-0.0380	-0.4395	-0.2004	32.0	14	-1.4641	0.0024	1	-0.6611	2
5			0.0643	0.2143	0.2096	66.0	16	2.8056	-0.2374	2	0.1189	-0.1045	0.1579	70.5	16	2.6659	-0.1443	2	2.8358	4
5			-0.0144	-0.1000	-0.0664	46.5	20	-0.4293	-0.2344	1	0.0705	-0.2262	0.0200	62.0	20	1.4755	-0.2954	2	0.5231	1
5			-0.0143	-0.1310	-0.1259	30.5	16	-1.1209	-1.0914	1	-0.0258	-0.3696	-0.0670	40.0	17	-0.4611	-0.0378	1	-0.7910	2
5			0.0124	0.0053	0.0387	70.0	16	1.1781	0.7853	2	0.0131	-0.2153	-0.0761	36.0	16	0.0770	0.7861	1	0.6276	1

2008:

Teacher	School	grade	MLMM Teacher Effect	Overall MLMM Teacher VAM	MQR Teacher VAM	Mteacher Median SGP	Mnumber students	mipm1	mipm2	mcluster	RLMM Teacher Effect	Overall RLMM Teacher VAM	RQR Teacher VAM	Rteacher Median SGP	Rnumber students	rpin1	rpin2	roluster	Overall RMTETI	Overall Cluster
		3	0.3310	0.2649	0.3752	74.0	19	2.5999	0.0662	2	0.3064	0.4593	0.3312	72.0	18	3.1263	-0.4174	3	2.8631	1
		3	-0.2563	-0.5730	-0.3819	17.0	6	-2.1026	-0.3656	1	-0.2093	-0.2923	-0.2052	33.0	25	-1.7558	0.2838	2	-1.7558	2
		3	0.2382	0.1436	0.1140	52.0	31	1.3831	-0.3612	2									-2.1026	2
		3	0.1114	0.0113	0.0752	55.0	14	0.9514	-0.0179	2	0.0871	0.0836	0.0638	64.5	30	1.1024	0.2990	1	1.1024	1
		3	0.2883	0.2118	0.2996	70.5	22	2.2535	0.0798	2	0.0488	0.0736	0.0716	53.0	9	0.6991	-0.0256	1	0.8253	1
		3	-0.2434	-0.4544	-0.4177	33.0	13	-1.6109	0.2360	2	0.1132	0.1695	0.2050	71.0	22	1.7673	0.4070	3	2.0104	1
		3	-0.1948	-0.3953	-0.3636	34.0	15	-1.3303	0.1396	1	-0.2003	-0.3295	-0.3010	21.0	13	-2.2754	-0.1639	2	-1.9432	2
		3	0.0725	-0.0367	-0.0098	41.0	16	0.3759	-0.4169	3	0.0548	0.0520	-0.0653	50.0	15	0.3358	-0.1866	1	-0.4972	3
		3	0.0776	-0.0331	-0.0523	79.0	17	1.1638	1.1355	2	0.0450	0.0596	0.0726	46.0	11	0.4936	-0.2500	1	0.4947	3
		3	-0.4226	-0.9363	-0.9224	12.0	5	-3.8366	0.4198	1	0.0492	0.0263	0.0909	66.0	17	0.9554	0.5169	1	1.0596	1
		3	-0.0323	-0.1633	-0.0790	47.0	22	0.0189	0.0601	3	-0.2390	-0.7440	-0.5519	14.0	5	-3.7552	0.2847	4	-3.7959	4
		3	-0.0537	-0.2191	-0.1186	26.0	17	-0.6206	-0.6774	3	-0.1617	-0.2792	-0.3365	22.0	15	-2.1173	-0.3197	2	-1.3690	2
		3	0.1835	0.1887	0.0752	58.5	6	1.4222	-0.0450	2	0.1355	0.2597	0.2873	67.5	6	2.0639	0.1109	3	1.7431	1
		3	-0.3873	-0.6316	-0.4975	16.0	17	-2.6381	-0.1271	1	-0.1219	-0.1730	-0.2142	23.0	17	-1.4262	-0.3066	2	-2.0321	2
		3	0.2680	0.1898	0.2393	63.0	20	1.9704	-0.1349	2	0.0923	0.1312	-0.0132	54.0	20	0.7862	-0.2326	1	1.3783	1
		4	-0.0184	-0.0348	-0.0542	35.5	32	-1.0349	-0.0739	1	0.1148	0.2556	0.1650	51.5	16	1.6443	-0.2309	1	2.3468	3
		4	0.3129	0.3200	0.3475	76.0	16	3.0493	-0.1449	2	-0.0873	-0.0973	-0.0913	46.0	24	-0.8596	0.3943	2	-0.8804	1
		4	0.0023	-0.0851	-0.0405	42.5	24	-0.3013	-0.2943	1	-0.0031	-0.0052	0.0094	50.0	35	0.1076	0.2840	1	0.1076	1
		4	-0.0465	-0.0432	-0.1036	45.5	24	-1.0436	-0.4809	1	0.0307	0.1279	0.0315	61.5	24	0.9850	0.6971	1	-0.0278	1
		4	-0.1580	-0.2103	-0.2152	32.0	15	-2.4727	-0.5419	1	-0.1725	-0.2373	-0.1859	34.0	45	-2.1913	0.0892	2	-2.1913	2
		4	0.2083	0.1989	0.1701	69.5	18	1.7352	-0.2833	2	0.0397	0.0692	0.1067	57.0	12	0.9718	0.5374	1	1.3535	3
		4	0.0055	0.0104	-0.0005	47.5	22	-0.3930	-0.3958	1	-0.1105	-0.1612	-0.1436	54.0	15	-0.9669	1.0140	2	-0.6799	1
		4	0.2567	0.0399	-0.1523	14.0	5	-0.8735	2.0265	1	0.0795	0.0204	-0.2022	27.0	5	-0.9694	-1.6734	2	-0.9214	1
		4	0.1966	0.3830	0.1489	61.0	19	1.9344	0.1885	2	0.0295	0.1663	-0.0150	41.0	33	0.2076	-0.6508	1	0.2076	1
		5	-0.0627	-0.1769	-0.0767	24.0	17	-1.2268	-0.4390	1	-0.2428	-0.1858	-0.5524	34.0	12	-2.0116	0.9436	2	-1.6192	1
		5	-0.0792	-0.1778	-0.1559	36.0	31	-1.2803	0.3470	1	-0.0729	-0.0524	-0.1348	50.0	21	-0.1089	0.4698	1	-1.2803	1
		5	0.1661	0.3535	0.2869	75.0	21	3.5514	0.1588	2	0.1860	0.2994	0.1138	60.0	37	1.9349	-0.1957	1	1.9349	2
		5	0.0500	0.0240	0.0158	46.5	40	0.6570	-0.0283	1	0.1193	0.2515	0.0350	43.0	37	1.1076	-0.4605	1	1.1076	3
		5	-0.0790	-0.2774	-0.1469	46.0	29	-1.2439	0.8356	1	-0.1386	-0.5815	-0.5075	2.0	5	-2.9684	-0.8756	2	-1.2439	1
		5	0.0048	-0.1300	0.0196	24.0	5	-0.4573	-0.8741	1	0.1534	0.3367	0.0466	63.0	37	1.8341	0.1319	1	1.8341	2
		5									-0.0044	0.1688	-0.1195	39.0	33	0.2123	-0.0134	1	0.2123	3