Identifying effective and ineffective schools for accountability purposes: a comparison of four generic types of accountability models

Fen Yu
Louisiana State University and Agricultural and Mechanical College, fyu1@lsu.edu

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Part of the Education Commons

Recommended Citation
IDENTIFYING EFFECTIVE AND INEFFECTIVE SCHOOLS
FOR ACCOUNTABILITY PURPOSES:
A COMPARISON OF FOUR GENERIC TYPES OF ACCOUNTABILITY MODELS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Educational Theory, Policy, and Practice

by
Fen Yu
B. A., Wuhan University, 2000
M.Ed., Kent State University, 2003
August, 2007
ACKNOWLEDGEMENTS

There are many people who have contributed to this long but rewarding journey. First, I owe a tremendous amount of gratitude to Dr. Charles Teddlie, whose guidance, support, and dedication have played a crucial role in my doctoral journey. Thank you for being generous with your time and your commitment to my graduate work. Secondly, I also owe a tremendous amount of gratitude to Dr. Eugene Kennedy, who has selflessly shared his extensive knowledge of statistical modeling with me and has always been available to me whenever I need consultation.

Appreciation also goes to the other members of my dissertation committee: Dr. Diana Taylor, Dr. Kim S. MacGegor, and Dr. Timothy Slack. Thank you for validating this research and providing me with constructive comments. Thank you also for always being responsive to my emails requesting a date for general examinations, proposal defense, and final examine. I can not ask for a better committee than this one.

To Mindy Crain, Vicki Anderson and Li Liu, thank you for putting the data together for me and providing me with consultations on SAS programming and the Louisiana School and District Accountability Model. I am very glad that we have now become colleagues at the LDE.

Most importantly, I would like to thank my husband, Gefu Ji, who always puts my needs ahead of his own. We don’t make much as graduate assistants, but he has always been thrifty on himself so that I can be a little “extravagant” every now and then. And above all these, he is unusually patient, non-judgmental, and very supportive of my academic work.

Finally, I’d like to thank my parents and my parents-in-law. My parents always remind me to study hard, and my parents-in-law always remind me to live happily. Together they keep me balanced.
# TABLE OF CONTENTS

Acknowledgements ........................................................................................................... ii

List of Tables ......................................................................................................................... vi

List of Figures ......................................................................................................................... vii

Abstract ................................................................................................................................. viii

Chapter One: Introduction ..................................................................................................... 1
  Perspectives Regarding Accountability Models ................................................................. 1
  Statement of Purpose for the Study ....................................................................................... 5
  Significance of the Study ....................................................................................................... 6
  Research Questions .............................................................................................................. 7
  Definition of Terms .............................................................................................................. 9
  Scope and Limitations ......................................................................................................... 11
  Summary ............................................................................................................................. 12

Chapter Two: Literature Review ............................................................................................ 14
  Introduction ......................................................................................................................... 14
  The Evolution of School Accountability .............................................................................. 15
    School Performance Monitoring ....................................................................................... 16
    School Accountability ....................................................................................................... 18
    The Pros and Cons of High-Stakes Accountability .......................................................... 21
  No Child Left Behind ......................................................................................................... 28
    Key Accountability Provisions ......................................................................................... 30
    Criticisms of NCLB .......................................................................................................... 33
    NCLB Flexibilities ............................................................................................................ 37
  Louisiana’s School Accountability System ......................................................................... 39
    Before 1999 ...................................................................................................................... 39
    After 1999 ......................................................................................................................... 40
  Four Generic Types of Accountability Models ................................................................... 44
    School Effectiveness Research ......................................................................................... 45
    Hierarchical Linear Modeling (HLM) .............................................................................. 50
    Four Types of Accountability Models .............................................................................. 52
  Comparisons of Different School Accountability Models ............................................... 58

Chapter Three: Methodology ............................................................................................... 60
  Introduction ......................................................................................................................... 60
  Restatement of Research Questions .................................................................................. 61
  Phase I: Status Models ....................................................................................................... 63
    Phase I Model Specifications .......................................................................................... 63
    Phase I Sample ................................................................................................................ 64
    Phase I Measurement ...................................................................................................... 65
    Phase I Data Analysis ...................................................................................................... 66
<table>
<thead>
<tr>
<th>Chapter Four: Results</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>88</td>
</tr>
<tr>
<td>Results from the Status Model (Phase I)</td>
<td>88</td>
</tr>
<tr>
<td>Results for Research Question #1</td>
<td>89</td>
</tr>
<tr>
<td>Results for Research Question #2</td>
<td>91</td>
</tr>
<tr>
<td>Summary of the Findings from Research Question 1 and 2</td>
<td>95</td>
</tr>
<tr>
<td>Results from the Improvement Model (Phase II)</td>
<td>96</td>
</tr>
<tr>
<td>Results from Research Question #3</td>
<td>97</td>
</tr>
<tr>
<td>Results from Research Question #4</td>
<td>98</td>
</tr>
<tr>
<td>Results from Research Question #5</td>
<td>101</td>
</tr>
<tr>
<td>Results from the Growth Model (Phase III)</td>
<td>103</td>
</tr>
<tr>
<td>Results from Research Question #6</td>
<td>103</td>
</tr>
<tr>
<td>Results from Research Question #7</td>
<td>105</td>
</tr>
<tr>
<td>Results from the Value-Added Model (Phase IV)</td>
<td>107</td>
</tr>
<tr>
<td>Results from Research Question #8</td>
<td>108</td>
</tr>
<tr>
<td>Results from Research Question 9, 10, 11, and 12</td>
<td>113</td>
</tr>
<tr>
<td>An Overall Comparison of the Four Models</td>
<td>114</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Five: Conclusion</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Findings from the Four Models</td>
<td>117</td>
</tr>
<tr>
<td>Major Findings from the Status Model</td>
<td>117</td>
</tr>
<tr>
<td>Major Findings from the Improvement Model</td>
<td>119</td>
</tr>
<tr>
<td>Major Findings from the Growth Model</td>
<td>122</td>
</tr>
<tr>
<td>Major Findings from the Value-Added Model</td>
<td>123</td>
</tr>
<tr>
<td>Methodological and Policy Implications</td>
<td>125</td>
</tr>
<tr>
<td>Methodological Implications</td>
<td>125</td>
</tr>
<tr>
<td>Policy Implications</td>
<td>126</td>
</tr>
</tbody>
</table>

References | 128 |
LIST OF TABLES

2.1 Average Four-Year Increase in NAEP Scores .....................................................25
2.2 Potential Effects of High Stakes Testing ...........................................................29
3.1 The AMOs for Math .........................................................................................63
3.2 Student Demographic Information for Phase I Study ........................................65
3.3 Louisiana School Accountability Indicators and Weighting ...............................69
3.4 Performance Labels .......................................................................................70
3.5 Growth Labels ...............................................................................................71
3.6 Student Demographic Information for Phase II Study .......................................73
3.7 The Amount of Improvement in Terms of Decrease in the Distance between Baseline Performance and Proficiency in the Target Grade .................................76
3.8 Student Who Have More Than One Data Points in 2005 ....................................79
3.9 Student Demographic Information for Phase III Analysis .................................81
3.10 Identifying Students for Value-Added Analysis ................................................83
3.11 Student Demographic Information for Phase IV data analysis ..........................84
4.1 Performance of the Five Subgroups .................................................................95
4.2 Performance Labels Based on SPS ..................................................................98
4.3 Growth Labels of the Schools .......................................................................102
4.4 The Effects of the Growth Model on Subgroup Performance ............................107
4.5 Three-Level Unconditional Model ..................................................................110
4.6 Three-Level Conditional Model .....................................................................112
4.7 Comparing the Four Models .........................................................................114
4.8 Using “Fairness” as the Additional Criterion ....................................................115
LIST OF FIGURES

2.1 Trends in State Accountability Systems .................................................................20
2.2 Trends in Percentile Rank of State Means ...............................................................22
2.3 Analysis of ACT for Alabama State ........................................................................24
2.4 Estimated Gains for a Typical Students in States ....................................................27
2.5 Focusing on middle-level students .......................................................................35
2.6 Examples of Back-Loaded Trajectories Based on Michigan’s Plan for English ....42
2.7 Example of Straight-with-Plateaus Trajectory Based on Florida’s Plan for Mathematics .................................................................43
2.8 An Illustration of a Status Model ...........................................................................53
2.9 An Illustration of an Improvement Model ..............................................................54
2.10 An Illustration of a Growth Model ........................................................................55
2.11 An Illustration of a Value-Added Model ...............................................................56
3.1 A Flowchart of the Status Model ...........................................................................68
3.2 A Flowchart of the Improvement Model ..................................................................72
3.3 An Example of North Carolina’s Model .................................................................77
3.4 A Flowchart of the Growth Model .........................................................................78
4.1 Before and After Applying Confidence Interval: Status Model ...............................92
4.2 Schools That Have Reached Louisiana’s Academic Goal of 2014 ............................97
4.3 Schools with One or More Subgroups Failing to Grow Two Points .......................101
4.4 Students Who Are Projected to Be Proficient .......................................................105
4.5 The “Value” Added by Schools ..............................................................................113
ABSTRACT

The stakes associated with student performance have been raised to new highs under the No Child Left Behind Act (NCLB, 2001). Many people are concerned with the adequacy and appropriateness of the statistical models used in identifying low-performing schools for educational accountability. The purpose of this study was to compare four generic types of accountability models (i.e., status models, improvement models, growth models, and value-added models) and see if they reach consistent/inconsistent conclusions regarding the effectiveness of the same set of schools. Further, the four models were also compared in terms of “fairness”. A fair model is defined as one that produces school effectiveness indices that have low correlations with various student background variables (Webster, Mendro, Orsak, & Weerasinghe, 1998). The sample included in this study consisted of all 297 K-5 schools in Louisiana. The results indicate that (1) the school effectiveness indices produced by the status model, the improvement model, and the growth model diverged significantly from those produced by the value-added model but converged highly among themselves; and (2) the school effectiveness indices produced by the value-added model had the lowest correlation with various student background variables. The methodological and policy implications of these findings were discussed.
CHAPTER ONE: INTRODUCTION

Perspectives Regarding Accountability Models

Accountability has become a central issue since the No Child Left Behind Act (NCLB, 2001) was signed into law by President George W. Bush in January 2002. NCLB requires all students, regardless of their race/ethnicity, gender, disability status, migrant status, economically disadvantaged status, or limited English proficiency status, to reach proficiency in reading/English language arts and mathematics by the end of the school year of 2013 – 2014 (CTB/McGraw-Hill, 2002; National Association of School Boards of Education, 2002). In order to ensure that the ultimate goal will be realized, NCLB requires educational institutions (i.e., schools and districts) to make Adequate Yearly Progress (AYP). The quantified form of AYP is the Annual Measurable Objectives (AMO), which is usually defined as percentage of students (by subgroup and subject) scoring at or above the proficient level of academic achievement. An educational institution is considered as making AYP only if the student population as whole and each of the six subgroups\(^1\) reach AMO of the year. Schools and districts that do not make AYP for two or more consecutive years will be held accountable and therefore subjected to a series of escalating consequences (CTB/McGraw-Hill, 2002; National Association of School Boards of Education, 2002).

In addition to AMO, NCLB also provides a secondary indicator of AYP, Safe Harbor, for schools and districts that may have been negatively influenced by certain subgroups (e.g., schools serving a large number of students with limited English proficiency). Using Safe Harbor as the secondary indicator, a subgroup is considered as making AYP if the students scoring

\(^1\) The six subgroups cover race/ethnicity, gender, disability status, migrant status, economically disadvantaged status, and limited English proficiency status (The National Association of State Boards of Education, 2002).
below the proficiency level are reduced by 10% from the previous year, regardless of the initial yearly objectives (National Association of School Boards of Education, 2002).

NCLB has been applauded for its emphasis on equity, its relentless insistence of 100% proficiency (i.e., all students, regardless of their ethnic, racial, or special-needs status, must reach proficiency by the end of the school year of 2013 – 2014), and its “no excuses” approach to district and school accountability for student performance (Lockwood, 2005; Public Affairs Research Council of Louisiana, 2004). However, it has also incurred many criticisms, most of which are related to how AYP is determined (Doran & Izumi, 2004).

Before growth models were allowed, AYP was determined jointly by a status model (i.e., AYP) and an improvement model (i.e., Safe Harbor), both of which use very simplistic measures. Status models measure only the current status of student achievement. When used in educational accountability, status models reflect students’ family backgrounds (i.e., SES) more than school effectiveness (Teddlie, Reynolds & Sammons, 2000). Improvement models, which are also called status-change models, are not much more advanced than status models (Hanushek & Raymond, 2002). Improvement models compare the same grades in two different years (e.g., 3rd graders in 2005 vs. 3rd graders in 2006, 4th graders in 2005 vs. 4th graders in 2006). This technique is like comparing apples to oranges because the 3rd graders of one year can be quite different from the 3rd graders of another year in almost all important aspects (e.g., racial composition, percentage of students with limited English proficiency). This is especially true for schools and districts that have high student mobility rates (Goldschmidt, Roschewski, Choi, Auty, Hebbler, & Blank et al., 2005).

As dissatisfaction with simplistic accountability models have increased substantially under the current NCLB law, growth models have been proposed as a useful addition to AYP
and Safe Harbor models. Growth models are characterized by using linked data (i.e., panel data), vertically scaled measurements (i.e., a continuous measurement scale to allow comparison from one grade to the next), and more sophisticated statistical modeling techniques (e.g., hierarchical linear modeling). All of these characteristics help eliminate many extraneous variables that lead to biased conclusions regarding a school’s effectiveness and therefore make growth models more desirable for high-stakes accountability (Lissitz, Doran, Schafer, & Willhoft, 2006).

Although perceived as more fair, growth models had been criticized for implicitly setting lower standards for disadvantaged students. This is true for value-added models, a special class of growth models. Value-added models decompose the variance of the test scores into portions that are explained by student inputs (i.e., adjusting for student backgrounds), and into other portions that are presumed to be directly related to schools. Schools are held accountable only for the portions of variance that they have control of (Lissitz et al, 2006). This idea is somewhat similar to regression models used in earlier school effects research (Teddlie & Reynolds, 2000), except that value-added models normally use panel data and hierarchical linear modeling techniques.

Since one of the most important features of NCLB is to hold all students to the same standards, growth models were not considered by the U. S. Department of Education for NCLB accountability purposes until November 2005, when the U. S. Secretary of Education Margaret Spellings announced the growth model pilot project (Olson & Hoff, 2005). In response to the Secretary’s call for growth models, many states have proposed new growth models (excluding value-added models) that incorporate various NCLB requirements, the most important of which are the ultimate goal of 100% proficiency by the school year of 2013-14 and closing the achievement gaps among subgroups. The newly proposed growth models hold all students to the
same standards, but allow a reasonable amount of time (i.e., by the school year of 2013-14) for disadvantaged students to catch up. Schools and districts that are on track and making accelerated growth towards the ultimate goal of 100% proficiency are exempted from negative classification even if they fall short of the intermediate goals (Tennessee Department of Education, 2006; North Carolina Department of Public Affairs, 2006). Up until November 2006, the U.S. Department of Education (USDE) had approved growth models proposed by Tennessee, North Carolina, Delaware, Arkansas, and Florida (USDE, 2006a, 2006b).

The Louisiana State Department of Education (LDE) is also interested in growth models. According to former Assistant Superintendent Robin Jarvis, “the state’s accountability commission is interested in value-added analyses, but wants to wait until the state has several years of data from new state tests before making a decision” (Olson, 2005). There are at least two reasons behind Louisiana’s seemingly slow reaction to growth models. First, Louisiana adopted a statewide accountability system in 1999 (roughly two years before NCLB), and it appears to be working well. For example, Louisiana’s fourth graders demonstrated the most improvement in the nation in math on the 2000 National Assessment of Educational Progress (NAEP), and Louisiana’s eight graders were the third most improved in math on the 2000 NAEP (Goldschmidt, Roschewski, Choi, Auty, Hebbler, Blank, & Williams, 2003). The advent of NCLB has disturbed Louisiana’s statewide accountability system that is already in place and functioning well. Therefore, Louisiana’s present plan is to keep the current accountability system while trying to adjust for NCLB requirements (Linn, Baker, & Herman, 2005).

Second, Louisiana has managed, at least for the present time, to keep most of its schools out of corrective actions through various methods, such as (1) calculating confidence intervals around AMO, (2) counting the “basic” level of achievement as “proficient”, (3) using a back-
loaded trajectory (i.e., requiring only modest increases of proficiency rate in early years and much larger increase of proficiency rate later on), and (4) specifying the minimum subgroup reporting size (n >10) (Goldschmidt et al., 2003; Linn, Baker, & Herman, 2005).

These methods will only work for a few years because a higher percentage of students are required to score at or above proficiency every two years. Using the subject mathematics as an example, 36.9% of the students (by school and subgroup) are required to reach proficiency in 2002, 47.4% in 2004, 57.9% in 2007, 68.4% in 2010, 78.9% in 2011, 89.4% in 2012, and finally 100% proficiency in the school year of 2013-14 (i.e., the ultimate goal of NCLB). It is obvious that more and more schools will be identified as failing AYP as the time gets closer to 2014 (Bulletin 111, 2006; Linn, Baker, & Herman, 2005).

Statement of Purpose for the Study

The purpose of this study was to compare four generic types of accountability models (i.e., status models, improvement models, growth models, and value-added models) and see if they reach consistent/inconsistent conclusions regarding the effectiveness of K-5 schools in Louisiana. Since each of the four generic types of accountability models encompasses a number of specific models, a representative model must be chosen for each generic model, and the selected models will be run on the data collected on Louisiana students and schools. Specifically, the AYP model under NCLB was chosen as the representative of the status models; Louisiana’s current school and district accountability model was chosen as the representative of the improvement models; North Carolina’s Modified ABCs Model was chosen as the representative of growth models; and a value-added model specified by the author of this study based on the model that is used to evaluate Massachusetts charter schools (Massachusetts Department of Education, 2006) was chosen as the representative of value-added models.
The AYP model was chosen because it is the primary indicator of school performance under NCLB. Louisiana’s current school and district accountability model was chosen because it is a rather sophisticated accountability model that is already in place and functioning well. North Carolina’s Modified ABCs Model represents growth models that accommodate NCLB requirements and were recently approved by the USDE to be used for NCLB accountability purposes. Value-added models, although not yet approved by the USDE for NCLB accountability purposes, are the most advanced and fair.

Significance of the Study

The findings of this study should have important methodological implications. School accountability is fundamentally an issue of identifying ineffective schools and holding them responsible. Different accountability models define school effectiveness in different ways and, therefore, are likely to hold different sets of schools accountable. Determining the best procedure for defining effective/ineffective schools and specifying the models accordingly will always be of methodological importance.

The findings of this study should also have important policy implications. NCLB sets the ambitious (some say unrealistic) goal of 100% proficiency for all students by the school year of 2013-14, and holds schools and districts accountable based on their proficiency rates (NCLB, 2001). However, since the 1960s, school effectiveness researchers in the USA and elsewhere have consistently demonstrated that the majority of the variance in student achievement between schools is attributable to the socioeconomic status (SES) of the student body (e.g., Teddlie, Reynolds, & Sammons, 2000). Louisiana is one of the poorest states in the USA and serves a very high percentage of economically disadvantaged students. Therefore, Louisiana schools and districts are put at a disadvantageous position under NCLB because their students are less likely
to do well on exams even if they receive the same quality of instruction as students who come from more affluent families (e.g., Doran & Izumi, 2004; Public Affairs Research Council in Louisiana, 2004).

When stakes associated with school performance are high, which is the case under NCLB, there should be a credible model that can be used to fairly and accurately identify effective and ineffective schools. In this study, four generic types of accountability models are compared and the findings should have important implications regarding whether it is feasible to include alternative models in the Louisiana school accountability system.

Research Questions

This study consists of four phases, and each phase concerns one generic type of accountability model. In the first three phases, parallel research questions are addressed. Specifically, the model is first examined based on all students and then its effects on subgroups are examined. In the last phase, findings from the first three phases are compared. The 12 research questions are presented below.

Phase I (Status Model) Research Questions:

1) According to the Status Model, how many K-5 schools in Louisiana can be categorized as making AYP for the school year of 2004 – 05?

2) If a K-5 school is categorized as making AYP based on the student population as whole, will each of the following subgroups also make AYP?
   a. African American Students
   b. Economically Disadvantaged Students
   c. Students with Limited English Proficiency
   d. Students with Disabilities
Phase II (Improvement Model) Research Questions:

3) According to the Improvement Model, how many K-5 schools in Louisiana can be categorized as making AYP for the school year of 2004 – 05?

4) If a K-5 school is categorized as making AYP based on the student population as a whole, will each of the following subgroups also make AYP?
   e. African American Students
   f. Economically Disadvantaged Students
   g. Students with Limited English Proficiency
   h. Students with Disabilities

5) According to the Improvement Model, how many K-5 schools in Louisiana meet or exceed the expected improvement?

Phase III (Growth Model) Research Questions:

6) According to the Growth Model, how many K-5 schools in Louisiana can be categorized as making AYP?

7) If a K-5 school in Louisiana is categorized as making AYP based on the student population as a whole, will each of the following subgroups also make AYP?
   i. African American Students
   j. Economically Disadvantaged Students
   k. Students with Limited English Proficiency
   l. Students with Disabilities

Phase IV (Value-Added Model) Research Questions:

8) According to the value-added model, which schools add the most “value” to student learning?
9) Do schools that are categorized as making AYP by the status model add more “value” to student learning than those categorized as not making AYP?

10) Do schools that are categorized as making AYP by the improvement model add more value to student learning than those categorized as not making AYP?

11) Do schools that are categorized as making AYP by the growth model add more value to student learning than those categorized as not making AYP?

12) Do the four models generate consistent or inconsistent conclusions regarding the effectiveness of K-5 schools in Louisiana?

Definition of Terms

- School Accountability Models

  There are four generic types of accountability models: status models, improve models, growth models, and value-added models (Goldschmidt et al., 2005).

- Status Models

  A status model takes a snapshot of the performance of a subgroup/school/local educational authority (LEA) at one point in time, and compares that proficiency level with an established target (Goldschmidt et al., 2005). For example, the AYP model under NCLB is a status model. AYP takes a snapshot of the percentage of students meeting or exceeding the proficiency level of achievement in a specific school year, and then compares it to the target proficiency level of that year.

- Improvement Models

  An improvement model compares the change of status at two points in time (Goldschmidt et al., 2005; Riddle, 2005). For example, the percentage of 4th graders making proficiency in 2000 (i.e., status in 2000) can be compared to the percentage of 4th graders making
proficiency in 2001 (i.e., status in 2001). The resulting difference provides a measure of improvement. Safe Harbor, the secondary indicator of school performance under NCLB, is based on an improvement model. Safe Harbor applies if a school fails to make AYP of the year but reduces the number of below proficient scores of a student group by 10% from the previous year’s comparable group (Goldschmidt et al., 2005; Riddle, 2005).

- Growth Models

The fundamental difference between growth models and the two models introduced previously (i.e., status models and improvement models) is that in growth modeling cohorts of students are tracked (usually through unique identification numbers) over time as they move along their K-12 educational careers (Riddle, 2005). The intention is to see whether students have made any progress. For example, the performance of 4th graders in 2000 can be compared to the performance of the same students in 1999 when they were in the third grade (Goldschmidt et al., 2005). The change in score from one year to the other is the actual growth, which is usually compared to the expected growth.

- Value-Added Models

Value-added models are a very special class of growth models (Lissitz, Doran, Schafer, & Willhoft, 2006). Similar to growth models, value-added models also require (1) linked data, and (2) vertically scaled measurements, or some other techniques that can be used to create a developmental scale. The key distinction between the two is whether student demographic characteristics or previous achievement have been adjusted/accounted for so that the variances in student achievement can be better attributed to individual teachers, schools, or LEAs (Lissitz et al., 2006; Riddle, 2006). Value-added models typically use multilevel modeling techniques (Choi, Goldschmidt, & Yamashiro, 2006).
Hierarchical Linear Models (HLM)

HLM is a relatively new statistical technique. It provides a solution to the nesting problems (e.g., students nested in classrooms which in turn are nested in schools) that usually occur in organizational research. Hierarchical models come under a variety of names. They are labeled as multilevel linear models in sociological research, mixed-effects models and random-effects models in biometric applications, random-coefficient regression models in econometrics literature, and covariance components models in statistical literature (Raudenbush & Bryk, 2002).

Scope and Limitations

Much of the research on accountability models focuses only on primary schools/grades (e.g., Choi, Seltzer, Herman, & Yamashiro, 2004; Zhang & Zhang, 2006). The present study shares this particular limitation in that it only examined schools that had a K-5\(^2\) configuration in the state of Louisiana. Since four types of accountability models were examined, it would be overly complicated to include all schools with all sorts of grade configurations.

Another limitation of this study is that student achievement was indicated by Math performance only. The original plan was to use both Math and English Language Arts (ELA). However, the procedures used to examine Math would be identical to the procedures used to examine ELA. In order to avoid repetition, ELA was dropped from analysis.

A third limitation of this study is that the samples used for the four models were not identical. This happens for a couple of reasons. First, the four models vary greatly in terms of complexity. For example, the simplest model, the status model, uses cross-sectional data (i.e., one year) while the most sophisticated model, value-added model, uses longitudinal data. It is

\(^2\) The K-5 schools in this study were those that have grades (a) 1-5, (b) 3-5, (c) K-5, and (d) PK, K-5. Additionally, magnet schools, charter schools, and lab schools were eliminated from analysis. The resultant sample consisted of 297 schools.
unlikely that they are going to share the same sample. Secondly, in Louisiana different tests were
used for different grades. A norm-referenced test (NRT) was implemented at Grades 3 and 5,
while a criterion-referenced test (CRT) was implemented at Grade 4. NRT compares students in
relation to each other (as opposed to a set of content standards), so it doesn’t provide a simple
measure of proficiency. Only CRT compares student performance to the state content standards
and provides ratings like “advanced”, “proficient”, “basic”, “approaching basic”, and
“unsatisfactory”. Therefore, the status model, which only performs a simple count of proficiency
rate, used only fourth grade data (i.e., CRT scores). Thirdly, value-added models require data
with good quality. For example, the value-added model used in this study adjusts for student
backgrounds and previous test scores, so the students need to not only have longitudinal test
scores, but also no missing values on the background variables.

Last but not the least, all four accountability models used in this study are extant
accountability models that are being used in various states to evaluate school performance. Great
efforts have been made to keep what is done in this study consistent with how the models have
actually been used in reality. However, adjustments have been made due to the data and time
available. The adjustments have been discussed in this study all along.

Summary

This chapter introduces four generic types of accountability models: status models,
 improvement models, growth models, and value-added models. The purpose of this study was to
test these models using longitudinal data collected on Louisiana students and schools to see if
they reach consistent/inconsistent conclusions regarding the effectiveness of K-5 schools in
Louisiana. In Chapter Two, a detailed review of the relevant literature is presented. Topics
covered include the evolution of school accountability, No Child Left Behind, Louisiana’s school
and district accountability system, four generic types of accountability models, and research studies that compare the appropriateness and adequacy of various accountability models. Chapter Three has four phases, each of which deals with one type of accountability models. The design components for each phase are discussed separately for each model in the order of (1) model specifications, (2) sample, (3) measurements, and (4) data analysis. In Chapter Four, findings from each of the four generic types of accountability models are presented. In the last chapter, the methodological and policy implications of these findings are discussed.
CHAPTER TWO: LITERATURE REVIEW

Introduction

This chapter is divided into five major sections: (1) the evolution of school accountability, (2) No Child Left Behind, (3) Louisiana’s school accountability system, (3) four generic types of accountability models, and (5) studies that compare the adequacy and appropriateness of various accountability models.

Section One examines school accountability from a historical point of view. School accountability started as school performance monitoring, and then developed into school accountability as high-stakes were attached to it. Major events in the evolution of school accountability include: the publication of the Coleman Report (1966), the establishment of the program The National Assessment of Educational Progress (NAEP), the publication of A Nation at Risk (1983), and the enactment and subsequent renewal of Elementary and Secondary Educational Act (1996). The pros and cons of high-stakes accountability are also discussed.

Section Two is a brief introduction to various aspects of NCLB, including key accountability provisions, criticisms of NCLB, and new flexibilities to educational accountability under NCLB.

Section Three concerns the evolution of a statewide accountability system in the state of Louisiana. The efforts to boost student performance before 1999 was largely futile, but the establishment of statewide school and district accountability system in 1999 seems to be working well and has produced some observable results.

In Section Four, development in school effectiveness research and hierarchical linear modeling are introduced to set the stage for discussions on the four generic types of
accountability models. Following this is a detailed introduction to status models, improvement models, growth models, and value-added models.

Section Five is a review of research studies that have been conducted to examine the appropriateness and adequacy of various accountability models. A general consensus among the findings of these studies is that simple measures like mean scores provide misleading information about the effectiveness of schools/teachers. Value-added models provide a useful alternative to educational accountability.

The Evolution of School Accountability

It is hard to trace the exact origin of school accountability. According to Cuban (2004, p. 19), “at no time in the history of U.S. public schools have those responsible for schools been unaccountable”. Primitive forms of school accountability appeared in the early 19th century when tax-supported public schools came into existence. At that time, trustees were elected by the local community members to serve on the state-chartered district school board, and they were legally obligated to ensure that public school children were adequately housed, taught, and supplied with learning materials (Cuban, 2004). The focus of school accountability was on spending school funds efficiently, and a good school was defined as “one that efficiently provided students access to adequate buildings, staff and materials” (p. 28).

The purposes that school accountability is intended to serve and the means through which the purposes are served have changed dramatically over time. The 1960s is generally regarded as the beginning of contemporary school accountability (Benveniste, 1985; Ravitch, 2005). It started as school performance monitoring, and then emerged as school accountability when high-stakes (i.e., rewards and sanctions) were attached to school performance indicators (Kochan, 1998).
School Performance Monitoring

Three major events shaped the development of school performance monitoring (Kochan, 1998). These events were: the publication of the *Equality of Educational Opportunity Study* (EEOS) (more commonly known as the *Coleman Study/Report* because of its head researcher James S. Coleman), the creation of the program of *National Assessment of Educational Progress* (NAEP), and the publication of *A Nation at Risk*.

*The Equality of Educational Opportunity Study*

In 1964, James S. Coleman and his colleges, commissioned by the United States Department of Health, Education, and Welfare, conducted the *Equality of Educational Opportunity Study* (EEOS), the main purpose of which was to “assess the availability of equal educational opportunities to children of different race, color, region, and national origin” (Inner-University Consortium for Political and Social Research, 1995). After analyzing test scores and questionnaire responses from a national sample of over 600,000 students and teachers, Coleman and his colleges concluded that family background variables influenced academic achievement the most in children’s early years, and that going to schools enlarged the achievement gap between the blacks and whites (ICPSR, 1995).

The Coleman Study was important for many reasons. One of the reasons was that for the first time in educational research the focus of research shifted from inputs to outcomes (i.e., academic achievement). As Ravitch (2002) remarked,

Before the Coleman report, educational reform had focused solely on the issue of resources, on the assumption that more generous provisions for teachers’ salaries, facilities, textbooks, and supplies would fix whatever ailed the nation’s schools. After the Coleman report, reformers advanced a broader array of proposals, many of which sought changes in performance rather than (or in addition to) increases in resource (Ravitch, 2002, p. 14).
Additionally, the Coleman Study also demonstrated inadequacies in the collection and analysis of educational data. In order to have better information on outcome (i.e., student achievement), Congress authorized the creation of the first educational performance indicator in the USA: *The National Assessment of Educational Progress* (Kochan, 1998).

*The National Assessment of Educational Progress*

NAEP is also known as “the nation’s report card”. It is the only nationally representative and continuing assessment that monitors trends in the knowledge and skills that America’s students acquire at school. It reports results for student achievement at grades 4, 8, and 12 in various subject areas, such as reading, mathematics, science, writing, US history, civics, geography, etc. When first implemented in 1969, NAEP only reported results at grade levels (e.g., fourth-graders) and subgroups within grade levels (e.g., female fourth-graders), and no information on the performance of individual students, schools, school districts, or states was made public. In 1990, NAEP started reporting test results at the state-level, which enabled inter-state comparison of academic achievement (National Assessment Governing Board, 2006). Since test results are reported at the state-level, schools and districts have little motivation to game the system (e.g., teaching to the test) in order to perform better on the tests. Therefore, NAEP is generally regarded as a good/unbiased indicator of what students know and can do in the USA.

*A Nation at Risk*

In 1983, the National Commission on Excellence in Education published a very influential report: *A Nation at Risk*. Essentially, the authors of this report attributed America’s mediocre economic performance in the global market to the mediocre student performance on national and international tests (Sirotnik, 2004). They claimed that weakness in American education had made the country vulnerable to “unfriendly foreign power”, and therefore had put
the nation at risk (National Commission on Excellence in Education, 1983). This report attracted a great deal of attention from the general public, and it triggered a nationwide educational reform movement. The reform included increasing high school graduation requirements, lengthening the school year, and notably adding more tests for students to take (Cuban, 2004). As Kochan (1998) remarked, “the report [A Nation At Risk] was largely responsible for triggering a decade-long wave of educational reform activity and creating a policy climate which decision makers could finally justify expensive initiatives to design and collect more detailed performance information”.

School Accountability

School accountability is school performance monitoring with sanctions attached to it (Kochan, 1998). The development of school accountability has been greatly influenced by the federal legislation known as the Elementary and Secondary Education Act (ESEA), which was first signed into law in 1965 and has been reauthorized approximately every five years. Major events in the development of school accountability system include: (1) ESEA of 1965, (2) Improving America’s Schools Act (IASA) of 1994, and (3) NCLB of 2001. Under the influence of ESEA of 1965 and its various reauthorizations, the majority of the states have gone through the processes of adopting a statewide testing program, imposing rewards and sanctions on schools based on test performance, developing curriculum and achievement standards, and finally building a statewide accountability system that aligns standards and assessment.

ESEA of 1965

In 1965, the U.S. Congress passed the first and most extensive federal education law in history: the Elementary and Secondary Education Act (ESEA). The ESEA is part of President Lyndon Johnson’s legislative plan entitled War on Poverty, and it allocates, through Title I
compensatory education program, substantial federal funds to improve the educational
opportunities of school children from low-income families (Archuleta, 2002). Title I requires
annual evaluation to prevent districts from diverting or wasting federal funds. Educational
achievement is made the yardstick in judging the success of ESEA (Cuban, 2004). As Hamilton
and Korets (2002) remarked, “standardized achievement tests became a central means of
evaluating Title I programs. Some observers viewed this as a key step toward the use of tests as
monitoring and accountability devices.” (p. 16) During the next two decades after ESEA of
1965, many states started to implement a statewide testing program. By the 1970s, approximately
60% of the states had adopted statewide testing programs. By the 1990s, most states had put a
statewide testing program in place (Cuban, 1994).

IASA of 1994

IASA of 1994 was a reauthorization of the ESEA of 1965. It was characterized by being
increasingly outcome-oriented. Compared with all previous ESEA reauthorizations, IASA had
narrower and more specific provisions regarding standards and assessment, many of which have
served as foundations for NCLB of 2001 (Riddle, 2006).

For example, IASA required states to develop and implement standards of Adequate
Yearly Progress (AYP), which later on became the cornerstone of NCLB. However, the AYP
standards under IASA were “transitional” because IASA gave states several years (1994-2002)
to develop and implement curriculum content standards, pupil performance standards, and
assessments linked to these standards. Before states accomplished these tasks, they would use
“transitional” assessments and “transitional” student performance standards. Therefore, the AYP
standards based on these “transitional” assessment and performance standards were also
“transitional” in nature (Riddle, 2006). Additionally, the IASA stated broadly that all students
must meet the proficient and advanced levels of achievement, which later on became one of the most important provisions under NCLB. The IASA didn’t specify a timeline for meeting the ultimate goal of 100% proficiency like NCLB does. Therefore, most states didn’t include the ultimate goal in their AYP plans (Riddle, 2006).

In addition to the enactment of the IASA of 1994, the 1990s also saw the emergence of state-wide accountability systems imposed by individual states. By the school year of 2001-2002, roughly before the enactment of NCLB, about 90% of all states issued report cards; more than 35% of all states included explicit financial rewards, sanctions, or a combination of both (Kane & Staiger, 2002). (See Figure 2.1)

![Figure 2.1](image)

Trends in State Accountability Systems (Kane & Staiger, 2002, p. 93)

NCLB of 2001

Few congressional actions have influenced American public education as NCLB has. NCLB requires, for the first time in history, every state to have an approved accountability plan to ensure academic proficiency for all children (Zwick, 2006). Rather than being vague, NCLB
has specific and stringent provisions regarding (1) standards and assessment, (2) verification of standards and assessment through NAEP, (3) state and district report cards, (4) AYP, and (5) school improvement, corrective actions and restructuring. The specifics of these provisions will be discussed at length later in a separate section of this chapter.

The Pros and Cons of High-Stakes Accountability

High-stakes accountability has always been a controversial issue (Benveniste, 1985). First, researchers can not agree on whether high-stakes accountability increases student learning. Secondly, one of the major justifications of high-stakes accountability is to change the behavior of teachers and students in desirable ways (Amrein & Berliner, 2002; Raymond & Hanushek, 2003). Nevertheless, there is disagreement on whether the educational practices have been changed for the better (Center for Public Education, 2006).

High-Stakes Accountability and Student Learning

Assessing the effects of high-stakes accountability on student learning is not an easy task. Statewide accountability systems differ significantly in terms of the test chosen, grades monitored, subject tested, and performance criterion used (Raymond & Hanushek, 2003), which makes it difficult to compare the effects of accountability on student achievement across different states. Additionally, test inflation makes within state comparisons less meaningful. Possible reasons for test inflation include: (a) using the same test over and over again; (b) excluding low-performing students from taking the test; and (c) narrowing the curriculum or simply teaching to the test. Consequently, the current average performance is usually higher than the mean of old norms, which was based on a more representative/realistic sample of students under a less distorted learning environment (Linn, 2000). In the 1990s, test scores almost always went up when a state adopted a new accountability testing program. The upward trend continued
until the old test was replaced by a new test and all of sudden scores dropped to a level that was as low as when the old test was first implemented (Linn, 2000). Figure 2.2 shows what happened to the reading and mathematics test scores when a state replaced the old test with a new test.

![Figure 2.2](image)

Trends in percentile rank of state means (Linn, 2000, p. 7).

The rollercoaster pattern of test scores is enough evidence that accountability test results lack generalizability or robustness. Therefore, many researchers resorted to tests other than those associated with high-stakes accountability (i.e., independent tests), such as NAEP, for cross state comparisons (Amrein and Berliner, 2002). In fact, NAEP has been used as indicator of student learning in several studies examining the effects of school accountability.

For example, Amrein and Berliner (2002) examined the effects of high-stakes accountability on student achievement by comparing the growth rate of student achievement before and after the introduction of high-stakes accountability. Student achievement was indicated by test scores on NAEP, American College Test (ACT), Scholastic Aptitude Test (SAT), and Advanced Placement (AP), all of which were independent tests. The study was based on the assumption that if high-stakes accountability could improve student learning, which
seemed to be the case when state testing programs were used, then the increase in student learning should be reflected in NAEP, ACT, SAT, and AP scores, which measure similar sets of domains.

One facet of validity is that the scores on a test are indicators of performance in the domain from which the test items are drawn. Thus, the score a student gets on a ten-item test of algebra, or on their driving test, ought to provide information about how that student would score on any of the millions of problems we could have chosen from the domain of algebra, or on how that student might drive in innumerable traffic situations (Amrein & Berliner, 2002, p. 21).

The sample included in Amrein and Berliner’s (2002) study were 18 states in which most severe consequences were attached to test scores. For each state, Amrein and Berliner depicted students’ average score across years for each test, and then compared the state’s trend line to that of the nation. It was assumed that a faster growth rate in student achievement would be observed after the introduction of high-stakes accountability, if high-stakes accountability indeed had a positive effect on student achievement.

Figure 2.3 presents a comparison of Alabama’s ACT performance against that of the nation’s. Alabama started imposing high-stakes on test scores in 1983, when high school students were required, for the first time, to pass graduation exam. Amrein and Berliner (2002) used ACT in 1983 as “baseline” and hypothesized that growth in average ACT scores for Alabama would be higher than that of the national average from 1984 on because of the introduction of high-stakes accountability system. It was found that Alabama gained 0.1 point over the nation from 1984 to 1985, 0.3 point from 1984 to 1992, 0.1 point from 1992 to 1993, and 0.1 point from 1992 to 2001. These increases were regarded as negligible.

Using analysis similar to what they employed with Alabama’s ACT scores, Amrein and Berliner (2002) plotted all other states in the study against the nation for all four tests (NAEP, ACT, SAT and AP), and came to the conclusion that student learning remained at the same level
as it was before the implementation of high-stakes accountability for all but one analysis, in which student learning was indeterminable. “Analyses of these data revealed that if the intended goal of high-stakes policy is to increase student learning, then that policy is not working” (Amrein & Berliner, 2002, p.1).

Amrein and Berliner’s (2002) study attracted a lot of attention. A number of researchers examined their study and pointed out that their methodology and conclusions were problematic. The most frequent criticism was that there was no comparison/control group. Only states with high-stakes accountability systems were included in the study while states without such systems were eliminated. “AB [Amrein and Berliner] violate the first principle of social science research – the need to control for the condition of interest…. The natural comparison group, however, is the states that had not adopted accountability systems.” (Raymond & Hanushek, 2003, p. 124)

Rosenshine (2003) reanalyzed the NAEP data, and added 14 to 18 states\(^3\) that didn’t have high-stakes accountability as a control group. For both groups of states (with vs. without high-

---

\(^3\) The number of states without high-stakes accountability systems varied from 14 to 18 because sometimes a state didn’t participate in NAEP for a certain subject area (e.g., reading) in a certain year (e.g., 1994).
stake accountability), Rosenshine calculated the difference between (a) 4th grade NAEP mathematics in 1996 and 4th grade NAEP mathematics in 2000, (b) 8th grade NAEP mathematics in 1996 and 8th grade NAEP mathematics in 2000, and (c) 4th grade NAEP reading in 1994 and 4th grade NAEP reading in 1998. The difference was viewed as growth over the four-year period, and the two types of states were compared in terms of the magnitude of growth. (See Table 2.1) Rosenshine’s study indicated that states with high-stakes accountability systems performed much better on NAEP reading and mathematics than those states without high-stakes accountability systems did.

Table 2.1
Average Four-Year Increase in NAEP Scores (Rosenshine, 2003, p.24)

<table>
<thead>
<tr>
<th>NAEP Test</th>
<th>Average Four-Year Increase in Clear-High-Stakes States</th>
<th>Average Four-Year Increase in States Without High-Stakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>4th grade mathematics</td>
<td>3.45 (n=11 states)</td>
<td>2.40 (n=15 states)</td>
</tr>
<tr>
<td>(1996-2000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8th grade mathematics</td>
<td>3.42 (n=7 states)</td>
<td>1.63 (n=13 states)</td>
</tr>
<tr>
<td>(1996-2000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th grade reading</td>
<td>3.44 (n=9 states)</td>
<td>1.21 (n=14 states)</td>
</tr>
<tr>
<td>(1994-1998)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Having a control group eliminated many threats to internal validity, and therefore significantly increased the trustworthiness of Rosenshine’s study. However, Rosenshine’s study also had its limitations. It used a cross-sectional design, in which different groups of students were compared. Specifically, 4th graders in 1996 were compared to 4th graders in 2000, and 8th graders in 1996 were compared to 8th graders in 2000. Differences in NAEP reading and math scores observed in these two groups may have nothing to do with schooling, but only attributable to differences in changed student composition. For example, a state may simply have fewer economically disadvantaged students in 1996 and more such students in 2000. Under this circumstance, an observed decrease in NAEP scores is not evidence that the quality of education
is decreasing. Another limitation of Rosenshine’s study was that it used very simple statistical measures as indicators of student learning (i.e., NAEP mean scores). Factors such as students’ previous achievement and student backgrounds were not statistically controlled.

Limitations of Rosenshine’s study were addressed later on in other studies in which sounder design and more sophisticated statistical models were employed. For example, Raymond and Hanushek (2003) also performed a reanalysis of the NAEP data. In addition to using states that didn’t have high-stakes accountability systems as a control group, Raymond and Hanushek also followed two cohorts of students as they moved from 4th grade to 8th grade (i.e., they tracked one cohort of students who entered 4th grade in 1992 and became 8th graders in 1996 and then another cohort of students who entered 4th grade in 1996 and became 8th graders in 2000). The expected growth in NAEP mathematics, as students moved from 4th grade to 8th grade, was calculated with regression analysis, in which certain student background variables were statistically controlled.

Results of this study indicated that the expected gain in NAEP mathematics of a typical student, after controlling for parental education level and school spending, was 0.7 for states without an accountability system, 1.2 for states with a report card, and 1.6 for states with consequential accountability. (See Figure 2.4) Therefore, Raymond and Hanushek (2003) concluded that “testing and accountability as practiced have led to significant gains in student performance over that [which was] expected without formal systems” (p. 122).

In addition to the studies by Rosenshine (2003) and Raymond and Hanushek (2003), many other researchers (e.g., Braun, 2004; Carnoy and Leob, 2003) performed additional reanalysis of the NAEP data as a reaction to Amrein and Berliner’s (2002) study. They unanimously concluded that states with high-stakes accountability had higher NAEP scores.
High-Stakes Testing and Behavior Change

High-stakes accountability has been intended as a mechanism to change educational practices in desirable ways. However, many researchers view high-stakes accountability as a distraction to normal teaching and learning, which generated many detrimental side effects. The frequently cited negative effects include:

a) narrowing the curriculum by excluding from it subject matter not tested;

b) excluding topics either not tested or not likely to appear on the test even within tested subjects;

c) reducing learning to the memorization of facts easily recalled for multiple-choice testing; and

d) devoting too much classroom time to test preparation rather than learning (Center for Public Education, 2006, p. 4)

Hamilton and Stecher (2004) reviewed both the positive and negative effects of high-stakes testing and generated a more balanced view. High-stakes testing has influenced students,
teachers, administers, and policymakers in both positive and negative ways. Positive effects include: motivating students and teachers to work harder and smarter, causing administrators to adjust school policies to meet the needs of curriculum and instruction, and helping policymakers to judge the effectiveness of educational policies and relocate resources. Negative effects include: putting high pressure on students and therefore discouraging them from trying, causing teachers to “teach to the test”, distracting administrators from other school duties. Table 2.2 provides a more detailed summary of the effects of high-stakes testing at various individual levels.

It should be noted, however, that these intended and unintended effects were not based on rigorous research. They appeared mostly in essays, anecdotal reports, testimonials, and protests (Center for Public Education, 2006). As pointed out by Cimbricz (2002), despite the great volumes of literature on the effects of high-stakes testing, only “a small body of work” was empirical studies; and within that small body of work, only “a handful” of them were studies that met professional standards of qualitative and quantitative research.

No Child Left Behind

In 2002, NCLB was signed into law as the most recent reauthorization of the ESEA of 1965. NCLB is both comprehensive and complex. It covers three broad areas: (1) accountability provisions, (2) teacher quality provisions, and (3) provisions on state flexibility in the use of federal funds (CTB/McGraw-Hill’s Office of Public and Governmental Affairs, 2002). The area that is most pertinent to this study is the accountability provisions. In this section, the specifics of the accountability provisions (e.g., standards and assessment, verification of standards and assessment) are discussed. Additionally, the potential effects of these accountability provisions are also discussed.
Table 2.2
Potential Effects of High Stakes Testing (Hamilton and Stetcher, 2004, p. 86)

<table>
<thead>
<tr>
<th>Types of Effects</th>
<th>Effects on Students</th>
<th>Effects on Teachers</th>
<th>Effects on Administrators</th>
<th>Effects on Policymakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Effects</td>
<td>Provide students with better information about their own knowledge and skills</td>
<td>Support better diagnosis of individual student needs</td>
<td>Cause administrators to examine school policies related to curriculum and instruction</td>
<td>Help policymakers judge the effectiveness of educational policies</td>
</tr>
<tr>
<td></td>
<td>Motivate students to work harder in school</td>
<td>Help teachers identify areas of strength and weakness in their curriculum</td>
<td>Help administrators judge the quality of their programs</td>
<td>Improve policymakers’ ability to monitor schools system performance</td>
</tr>
<tr>
<td></td>
<td>Send clearer signals to students about what to study</td>
<td>Help teachers identify content not mastered by students and redirect instruction</td>
<td>Lead administrators to change school policies to improve curriculum or instruction</td>
<td>Foster better allocation of state educational resources</td>
</tr>
<tr>
<td></td>
<td>Help students associate personal efforts with rewards</td>
<td>Motivate teachers to work harder and smarter</td>
<td>Help administrators make better resource allocation decisions, e.g., provide professional development</td>
<td></td>
</tr>
<tr>
<td>Negative Effects</td>
<td>Frustrating students and discourage them from trying</td>
<td>Lead teachers to align instruction with standards</td>
<td>Lead administrators to enact policies to increase test scores but not necessarily increase learning</td>
<td>Provide misleading information that leads policymakers to suboptimum decisions</td>
</tr>
<tr>
<td></td>
<td>Making student more competitive</td>
<td>Encourage teachers to participate in professional development to improve instruction</td>
<td>Cause administrators to reallocate resources to tested subjects at the expense of other subjects</td>
<td>Foster a “blaming the victims” spirit among policymakers</td>
</tr>
<tr>
<td></td>
<td>Cause students to devaluate grades and school assessment</td>
<td>Lead administrators to waste resources on test preparation</td>
<td>Lead administrators to waste resources on test preparation</td>
<td>Encourage a simplistic view of education and its goals.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Entice teachers to cheat when preparing or administering tests</td>
<td>Distract administrators from other schools needs and problems</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Devalue teachers’ sense of professional worth</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Key Accountability Provisions

The accountability provisions under NCLB are very specific, and therefore leave little room for ambiguity. Major components of the accountability provisions include: (1) standards and assessments, (2) verification of state standards and assessments through NAEP, (3) report cards, (4) Adequate Yearly Progress, and (5) school improvement, corrective actions and restructuring (CTB/McGraw-Hill’s Office of Public and Governmental Affairs, 2002; National Association of State Boards of Education, 2006).

Standards and Assessments

Standards and assessments serve as the foundation of any accountability system. Since IASA of 1994, states have been required to have challenging academic content standards and student academic achievement standards, and valid and reliable assessments to go with these standards. However, before the passage of NCLB, only 17 states (including Louisiana) fully complied with the IASA requirements (Public Affairs Research Council of Louisiana, 2004).

The requirements on standards and assessments have been greatly reinforced under NCLB. According to NCLB,

- States must adopt challenging academic content standards and student academic achievement standards in reading and math;
- Starting in 2005-06, states must implement annual assessment on reading and mathematics in grades 3-8;
- Starting in 2007-08, states must implement assessment on science at least once in grades 3-5, 6-9, and 10-12;
- Assessment must be administered to at least 95% of the student population as a whole and 95% of each of the six subgroups;
Assessments must be valid, reliable, and aligned with state academic standards (CTB/McGraw-Hill’s Office of Public and Governmental Affairs, 2002; National Association of State Boards of Education, 2006).

Verification of State Standards and Assessments through NAEP

In order to prevent states from lowering standards and therefore escaping corrective actions, NCLB requires all states to participate in NAEP so that they will have a secondary indicator of student learning. It is expected that a state with high scores on state assessment should also have high scores on NAEP. The main NCLB provisions regarding verification of standards and assessments are:

- State test scores must be verified by a secondary indicator of student progress, the National Assessment of Educational Progress (NAEP);
- States must participate in NAEP every other year at the fourth and eighth grade levels;
- No federal rewards or sanctions are attached to NAEP results (CTB/McGraw-Hill’s Office of Public and Governmental Affairs, 2002; National Association of State Boards of Education, 2006).

Report Cards

Publishing report cards is not a new initiative, since about 45 of the states had already been publishing report cards before NCLB. Under NCLB, however, the content of the report cards have been greatly enriched. Specifically,

- Starting in 2002-03, state and district report cards must include disaggregated data for each assessment by the six required subgroups;
- State and district report cards must include information on test participation rate for the student population as a whole and by subgroups;
• States and district report cards must include the latest two-year trend data; and
• State and district report cards must include results on other public school indicators
  (CTB/McGraw-Hill’s Office of Public and Governmental Affairs, 2002; National

Adequate Yearly Progress

AYP is the essence of NCLB. Through its various provisions, AYP makes it almost
impossible for states to slip through the NCLB system. Although controversial, AYP holds all
students, regardless of their backgrounds and previous achievement levels, to the same standards.
AYP is also intended to fix the problem of achievement gaps. A state/district/school is
considered as making AYP only if the student population as a whole as well as the students in
each subgroup meets AYP. The main AYP provisions are as follows:

• Adequate Yearly Progress is defined in relation to the ultimate goal of 100% proficiency
  by the school year of 2013-14;
• A state/district/school is considered as making AYP only if the student population as a
  whole and students in each subgroup make AYP.
• States must set a baseline and equally-spaced intermediate goals (i.e., annual measurable
  objectives) to ensure that the ultimate goal of 100% will be realized.
• “Safe Harbor” applies if a school fails to make AYP but reduces students below
  proficiency by 10% from the previous year\(^4\) (CTB/McGraw-Hill’s Office of Public and

\(^4\) Safe Harbor is usually used when a school fails to make AYP of the year but has made significant improvement. If
a school reduces students scoring below proficiency by 10%, it is considered as making AYP even if it doesn’t reach
the AYP cut score (CTB/McGraw-Hill’s Office of Public and Governmental Affairs, 2002; National Association of
School Improvement, Corrective Actions, and Restructuring

NCLB has more teeth than any other federal legislation on education because of the escalating consequences associated with not making AYP. The escalating consequences are realized through (1) public school choice, (2) supplemental service, (3) corrective action, and (4) restructuring.

- **Public School Choice:** If a school fails to make AYP for two continuous years, the school district must provide all students in the failing school public school choice, that is, allowing students in the failing school to attend another school in the same district and compensating their transportation fees;

- **Supplemental Service:** If a school fails to make AYP for three continuous years, the school district must provide supplemental educational services to the low achieving, disadvantaged students.

- **Corrective Action:** Corrective action applies to schools that fail to make AYP for four consecutive years. The school district must implement corrective actions, such as adopting a new curriculum or replacing staff;

- **Restructuring:** Restructuring applies to schools that fail to make AYP for five consecutive years. Restructuring actions include (1) conversion to a charter school, (2) a state takeover, (3) contracting a private management firm, or (4) staff restructuring (CTB/McGraw-Hill’s Office of Public and Governmental Affairs, 2002; National Association of State Boards of Education, 2006).

**Criticisms of NCLB**

NCLB provides schools with incentives to improve their performance. As Mathers and Brat (2005) remarked:
When schools are held accountable, they have incentives to change because they know that there will be repercussions for their poor performance. In a sense, this is a way to force change in schools that have been performing poorly and haven’t made the changes necessary on their own (Mathers & Brat, 2005, p. 13).

Additionally, NCLB brings in competition, which is viewed by many as an important factor of school success. For example, through the school choice program schools will have an extra incentive to do better because they don’t want to lose all of their students (Mathers and Brat, 2005). Despite of these benefits, NCLB has incurred many criticisms, the majority of which are related to AYP.

Criticisms of AYP

Under NCLB, high-stakes are attached to AYP. However, many researchers pointed out that AYP was a weak indicator of school performance (Bryk, Thum, Easton, & Luppescu, 1998; Choi, Goldscmidt, & Yamashiro, 2006; Choi, Seltzer, Herman, Yamashiro, 2004; Doran & Izumi, 2004; Meyer, 1995; Webster & Mendro, 1998). It is based on simple statistical procedures of counting the percentage of students reaching proficiency level on a test. Doran and Izumi (2004) enumerated a couple of disadvantages of using simple statistical measures.

First, AYP based on simple measures provides misleading and invalid results. Since the 1960s, school effectiveness researchers in U.S. have consistently demonstrated that the majority of the variance in student achievement between schools is attributable to the socioeconomic status (SES) of the student body (Teddlie, Reynolds, & Sammons, 2000). Therefore, measurement without taking student SES into consideration is seriously biased. It puts schools serving poor students at a disadvantageous position, since their students are less likely to do well on exams even if they receive the same quality of instruction as students who come from more affluent families (Doran & Izumi, 2004).
Second, AYP is a crude measurement. It is responsive only to middle-level students who are closest to scratching the proficiency line. Moving this subgroup of students up to proficiency level will significantly change the outside appearance of the school when it comes to AYP. Schools don’t get credit for making their high-performing students even better or moving their low-performing students closer to proficiency. Consequentially, teachers may focus their teaching only on middle-level students. The instructional needs of high-performing students and low-performing students may be ignored, since the former need very little help to reach proficiency and the latter need a lot of help but still may not make it (Doran and Izumi, 2004) (See Figure 2.5).

Third, AYP may lead to student exclusion. The school environment may become “unfriendly” to racial or ethnic groups, special education students, or students with limited English proficiency because these students are less likely to reach proficiency than their more affluent peers.

All types of schools, whether elementary, middle, or high school, must make AYP. Students who perform poorly on state tests obviously hurt schools looking to make AYP.
This is why schools, to the extent they can, will work to avoid enrolling those students who are at risk of failing the exams. The same pressure could lead schools to push low-performing students out, either to another school (if one can be found that will accept them) or out of the school system entirely. Given the connection between performance on test, socioeconomic status, and race, the students most likely to be targeted for exclusion will be poor and/or racial minorities (Ryan, 2004, p.969).

Last but not the least, stringent AYP provisions may also lead to unequal distribution of good teachers. “Attaching consequences to test results creates obvious incentives for teachers to avoid schools that are likely to provide bad results (Ryan, 2004, p. 974)”. Good teachers flock to good schools where there is less pressure to make AYP, and low-performing schools have to accept whoever that are left (Ryan, 2004).

Other Criticisms

Although AYP is the major source of criticism, other NCLB provisions have also been attacked. Mathers and Brat (2005) identified some additional problems with NCLB. First, the NCLB goals are unrealistic. The ultimate goal of NCLB is 100% proficiency in 12 years, which is viewed by many as unrealistic and unattainable. According to Edward Haertel of CRESST (National Center for Research on Evaluation, Research and Student Testing) and Stanford University, it would take 110 years for all states to reach the ultimate goal of 100% proficiency if progress were based on NAEP experience (Lewis, 2002). Setting unrealistic goals has its consequences. With the odd combination of overly ambitious goals and severe sanctions, states may be encouraged to lower standards, make tests easier, or lower the scores needed to be deemed proficient (Mathers & Brat, 2005). As a matter of fact, Louisiana, Colorado, Connecticut, and Texas have already been accused of tinkering with their scoring systems in order to have a higher percentage of students reaching proficiency. Still, there are other states that have been accused of not only altering the scoring system but also making the tests easier to pass (Ryan, 2004).
Second, NCLB lacks real school choice. Under NCLB, if a school fails to make AYP for two consecutive years, the Local Educational Authority is required to provide students in this school with school choice (i.e., allowing them to go to a different school in the same district and compensating them for the transportation fee) (National Association of State Boards of Education, 2002). NCLB only allows within district transfer, but often times there are none or very few schools in the district that are well-functioning. Therefore, students in poor-performing schools are usually discouraged to change to another school. In the State of Louisiana, 61 schools in five parishes were required to offer school choice. Of the 40,485 students eligible for school choice, only 1% actually transferred to another school (Public Affairs Research Council of Louisiana, 2004).

NCLB Flexibilities

Stormed by criticisms, the U. S. Department of Education has decided to grant some flexibilities to NCLB. The flexibilities that are most relevant to this study are those on growth models.

The fact that AYP is a crude and invalid measurement of school performance has been an issue of great concern. Many researchers have proposed growth models as a supplement to AYP. Growth models account for the fact that different schools and students have very different starting points, and recognize progress being made at all levels (Tennessee Department of Education, 2006). Supporters of growth models view them as more fair and accurate than the AYP models (i.e., percent proficient and Safe Harbor). However, growth models contradict some of the statutory NCLB provisions. For example, NCLB holds all students to the same standards, regardless of their subgroup status and starting points of academic achievement. Growth models,
on the other hand, have the disadvantage of implicitly setting lower standards for the
disadvantaged children (Riddle, 2006).

The U. S. Department of Education had been reluctant to accept growth models until
November 2005, when the U.S. Secretary of Education Margaret Spellings announced a growth
model pilot program. The pilot program allowed up to 10 states to use growth models to make
AYP determination for the school year of 2005-2006 (Riddle, 2006). The first phase of this pilot
program was finished in March 2006, and eight states were permitted to go forward and have
their proposals evaluated by a national peer review panel (Thompson Publishing Group, 2006).
By May 2006, the peer review panel has given permission to two states, North Carolina and
Tennessee, to use their proposed growth models to determine if their schools are making AYP in
2005-2006. Unit now, 20 states have submitted proposals to participate in the pilot in the school
all states are considering using growth models (Shaul, 2006).

Growth models do not loosen up NCLB provisions much. They are governed by a long
list of criteria, some of which are:

- The ultimate goal of 100% proficiency by the school year of 2013-14 stays unchanged;
- Achievement gaps among subgroups should be narrowed in order for schools or LEAs to
  make AYP;
- Annual achievement goals should be based on the ultimate goal of 100% proficiency, not
  on pupil backgrounds, school characteristics, past performance, or “typical” performance
growth rate;
- Assessment results must be consistent from grade to grade and year to year;
- Progress of individual students must be tracked within a state data system (Riddle, 2006).
The technical aspects of growth models will be introduced in great detail later in this chapter after a review of Louisiana’s school accountability system.

**Louisiana’s School Accountability System**

Louisiana adopted a statewide comprehensive accountability system in 1999. Before the accountability system was put into place, Louisiana had implemented many waves of educational reform (Public Affairs Research Council of Louisiana, 1999).

**Before 1999**

Data collection for school accountability is a relatively recent phenomenon in Louisiana (Kochan, 1998). The 1970s were characterized by an overall focus on education inputs. In 1975, a major reform initiative named “school approval process” (SAP) was enacted to ensure that “all public schools meet certain minimum standards relating to school safety, facilities, and instruction”. Also in this decade, standards for instructional personnel were raised, and higher standards for student performance were established (Kochan, 1998).

In 1980s, after the publication of *A Nation at Risk* (1983), the Louisiana legislature passed the omnibus education bill “Children First Act”, which simultaneously established three programs for educational accountability. These programs are (1) the School Incentive Program (SIP), the Louisiana Teacher Evaluation Program (LaTEP), and the Progress Profiles (School Report Card) Program. SIP was intended to recognize and reward schools that were making exceptional progress, but it only lasted for a year because of no further funding. LaTEP was a high-stakes professional accountability program assessing teacher performance, which encountered bitter opposition from teacher unions because it provided a mechanism for revoking the credentials of incompetent teachers. LaTEP only lasted for a few years. The “Progress Profiles Program”, Louisiana’s education performance indicator system, is the only original
component of the “Children First Act” that is still in operation today. The purposes of the program are “to establish a database for education planning, to increase accountability at all levels, and to inform the parents of school children and the general public on the condition of public education” (Children First Act, 1988; Kochan, 1998, p. 64). At the beginning of this program, it produced a mix of school input and student outcome indicators including class size, faculty education and certification, plus student attendance, suspension, expulsion, and dropout rates. The program was converted to Student Information System (SIS) during mid-1990s. SIS is a sophisticated system that allows staff at the Louisiana Department of Education (LDE) to trace individual students anywhere in the state education system for information like demographics, enrollment/attendance patterns, and disciplinary records (Kochan, 1998).

The reforms (except for Progress Profiles Program) during the early years brought little impact because most of them were “either ignored, not implemented properly, challenged in court, repeated, gradually eroded or watered down” (Public Affairs Research Council of Louisiana, 1999, p. 2) As a result, Louisiana ranked near the bottom on almost all measures of academic performance, such as test scores, dropout rates, college remediation rates, and so forth (Tolbert, 2003).

After 1999

In order to boost test performance, the Board of Elementary and Secondary Education (BESE) in Louisiana approved a new school accountability program, the purpose of which is to hold schools accountable for student achievement results (Public Affairs Research Council of Louisiana, 1999). With a comprehensive statewide accountability system in place, test scores in Louisiana began to improve (Reeves, 2003).

The first three years of the new accountability system were characterized by increasing scores on the Louisiana Educational Assessment Program for the 21st Century, the state’s
criterion-referenced test, and the norm-referenced Iowa Tests. On annual School Performance Scores, which include student achievement data, attendance, and dropout rates, the state average increased from 69.4 in 1998-1999 to 81.3 in 2000-2001 (Reeves, 2003, p. 18).

Test scores also showed signs of improvement when compared with other states across the nation. For example, 4th graders in Louisiana exhibited the most improvement in math on NAEP of 2000, and 8th graders exhibited the third largest improvement. Louisiana’s success with its accountability system has received national recognition. In 1999, Louisiana, together with seven other states, received an “A” from the Thomas B. Fordham Foundation for its accountability system. Moreover, Louisiana’s accountability system was rated “B” in 2001 and “A” in 2002 in Education Week (Reeves, 2003).

The advent of NCLB disturbed an accountability system that was functioning well in Louisiana. Louisiana’s plan is to change its current accountability system as little as possible while trying to accommodate NCLB requirements (Coleman, 2003). Therefore, schools in Louisiana are now subjected to a dual accountability system, the major impact of which will be the likely identification of up to three-fourths of all public schools as failing (Public Affairs Research Council of Louisiana, 2004). At present, Louisiana has managed to put only a very small portion (less than 10% each year) of its schools into school improvement because NCLB granted states some flexibilities.

First, states have full autonomy in defining their own standards of academic achievement (Public Affairs Research Council of Louisiana, 2004). Before NCLB, Louisiana defined “basic” level of achievement as exhibiting “only the fundamental knowledge and skills needed for the next level of schooling” and “proficient” level of achievement as exhibiting “competency over challenging subject matter”. The “basic” and “proficient” levels defined by Louisiana’s accountability system were approximately equivalent to the “basic” and “proficient” levels on
NAEP. After NCLB was signed into legislation, however, Louisiana redefined its “proficient” level as “at or above basic”. In other words, student at the “basic” level will also be counted as “proficient”. Louisiana has been accused of easing academic standards due to these changes (Public Affairs Research Council of Louisiana, 2004; Porter, Linn, and Trimble, 2005).

Second, states can select their own trajectories to reach 100% proficiency by 2013-14. The trajectories can be “backloaded” (i.e., requiring smaller increases at the beginning and much bigger increases in later years) (See Figure 2.6), “straight with plateaus” (i.e., a stair-step approach that requires increases at least every third year) (See Figure 2.7), or simply “straight” (i.e., equal increase each year). Louisiana, like other 24 states, chose the “backloaded” approach. Therefore, Louisiana is able to keep a low percentage of schools in school improvement for the time being. However, this method will only work for a few years, since an increasing higher proficiency rate is required in the years to come and by 2013-14 a 100% proficiency rate is required (Porter, Linn, and Trimble, 2005; Public Affairs Research Council of Louisiana, 2004).

![Figure 2.6](image)

Figure 2.6
Example of Back-loaded Trajectories Based on Michigan’s Plan for English (Porter, et al., 2005, p. 35)
Third, states can decide the minimum group size that is required before disaggregated results are used to determine AYP for a school. This is usually referred to as “minimum N”. According to NCLB, a school/LEA is not considered as making AYP unless the student population as a whole and each subgroup makes AYP. However, if a subgroup is not large enough to be reported, then it won’t negatively affect the AYP status of a school/LEA. The minimum subgroup reporting size in Louisiana is 10 (Porter, Linn, and Trimble, 2005; Public Affairs Research Council of Louisiana, 2004).

Fourth, the majority of states have chosen to calculate a confidence interval around AMO. A school/LEA is considered as making AYP so long as the percentage of students scoring at or above proficiency reaches the lower band of the confidence interval. Louisiana uses a confidence of 99%, while the rest of the states either use lower than 95%, 95%, 98%, 99%,
As discussed earlier, the strategies that Louisiana is using to keep its schools out of school improvement will only work for a few years. Eventually, all students are expected to reach the proficient level of academic achievement. Louisiana, together with many other states, is hoping that NCLB requirements will be eased in its later years. Alternatively, Louisiana can seek to include growth models to help decide AYP, like what many other states are doing right now. Louisiana is one of the poorest states in the U. S., and it serves a very large percentage of disadvantaged students. Therefore, growth models and value-added models should be able to better identify failing schools.

Four Generic Types of Accountability Models

There are four general types of accountability models: status models, improvement models, growth models, and value-added models. Status models and improvement models are usually based on simple statistical measures such as percent proficient or means. They are widely used in educational accountability due to their simplicity. Growth models and value-added models, on the other hand, are much more complicated. The development of growth models and value-added models are built upon both findings from school effectiveness research and advance in statistical modeling (e.g., hierarchical linear modeling). In this section, I will first discuss major findings from school effectiveness research and the application of HLM in educational research to set the stage; then I will introduce the four general types of accountability models noted above; and finally, research studies examining the appropriateness of various models used in educational accountability will be presented.
School Effectiveness Research

The purpose of school accountability is, essentially, to assess the effectiveness of schools and hold the less effective schools accountable. However, what is an effective school? Should we judge the effectiveness of a school based on absolute measures, such as the percentage of students scoring at or above proficiency or average test scores? If so, schools serving children from median- or high-income families will almost always be more effective than schools serving children from low income families. As consistently indicated by numerous research studies, students’ SES explains the majority of the variances in student achievement (Teddlie & Reynolds, 2000). Consequentially, school effectiveness defined in absolute terms (i.e., percent proficient or mean scores) reflects the average SES of the student body in the school more than the school’s ability to educate its students. In other words, absolute measures are indicators of average SES of student body, not the “true” effectiveness of schools.

In order to find the “true” effectiveness of schools, researchers started to try to separate the variance that is attributable to student backgrounds (e.g., SES, limited English proficiency status) from the total variance in student achievement. The earliest statistical technique used in this endeavor was multiple regression. In multiple regression analysis, the importance of a predictor variable to an outcome variable is indicated by the percent of variance that the predictor variable can explain in the total variance that exists for the outcome variable. For example, if the total variance in student achievement is 2,500 and by adding the variable SES to the regression model the total variance is reduced to 500, then the variable SES is said to be able to explain about 80% (i.e., 2,000 out of 2,500) of the total variance in student achievement.

By calculating all the variance that is explained by non-school factors and separating them from the total variance in student achievement, researchers expected that the leftover
variance could be attributed to school effects. According to Teddlie and Reynolds (2000), on average schools are capable of explaining about 10-15% of the total variance in student achievement. The more variance school factors can explain in the total variance of student achievement, the more effective the school is. For example, a school that explains about 30% of the variance in student achievement is more effective than a school that explains only about 10% of the variance in student achievement.

However, can we really partition out the effects of all non-school factors? In other words, can we really measure all the non-school variables that could possibly affect student achievement and then separate the variance explained by them from the total variance? Obviously we can not, because the non-school variables we can measure are limited by the availability of instruments and practical considerations. Therefore, no matter what percentage of variance is attributed to schools effects, this variance is always a combination of the effects of schools and other unknown/unmeasured factors. The fact that we can not separate all non-school factors from school effects is an issue that has always perplexed school effectiveness researchers. Understanding this problem is important to the later discussion on growth models and value-added models in this chapter.

Studying the magnitude of school effects was only the beginning of school effectiveness research. Today, this line of research has grown into three separate but interrelated strands (Teddlie & Reynolds, 2000).

Strand 1: School Effects Research

This strand of school effectiveness research involves studying the existence of school effects, the magnitude of school effects, the context effects that influence the generalizability of school effects across different settings, the consistency of school effects at one point in time, the
stability of school effects across time, differential school effects on student subgroups within a school, and the continuity of school effects. (Teddlie et al. 2000) School effects researchers are usually quantitatively oriented, and they tend to use various statistical modeling techniques to answer their research questions.

Strand 2: Effective Schools Research

This strand of school effectiveness research involves studying the processes (e.g., effective leadership, effective teaching, high expectations, etc.) that are associated with effective/ineffective schools. (Reynolds, Teddlie, Creemers, Scheerens & Townsend, 2000) Effective schools researchers are usually more qualitatively oriented. Often they separate schools into four categories: high SES and high student achievement, high SES and low student achievement, low SES and high student achievement, and low SES and low student achievement. Effective schools are those serving students from low income families but still producing relatively high student achievement. Effective schools researchers are usually interested in what processes make these schools effective, although they serve students from low income families.

Strand 3: School Improvement Research

This strand of school effectiveness research involves examining the processes through which schools can be made more effective (Reynolds et al., 2000). School improvement researchers are usually interested in the effects of “interventions”, which are changes to school routines that are intended to make the school perform better on standardized tests.

Findings from the three strands of school effectiveness research have important implications regarding how to build a credible school accountability model. First, through examining a wide variety of predictor variables, school effectiveness researchers have identified variables that are most relevant to student achievement, which include prior achievement, SES,
ethnicity, gender, pupil/teacher ratio, school location (urban, suburban, and rural), and so forth (Rumberger & Palardy, 2003). These variables form the preexisting conditions under which a school operates, and a school may be positively or negatively influenced by these variables. For example, a school serving students from low-income families is less likely to outperform a school serving students from medium- or high-income families, even if the quality of schooling is exactly the same at both schools. In a fair accountability system, the preexisting conditions should be statistically controlled so that they don’t confound the measurement of quality of schooling, which is what schools should be held accountable for.

Second, school effectiveness researchers have called to attention the inadequacy of cross-sectional designs and demonstrated the importance of conducting longitudinal studies. A cross-sectional design is a design in which researchers collect data at one point of time (e.g., Creswell, 2002). In a cross-sectional design, a school can be identified as effective in one year and ineffective in another only because the student population has changed. To give a hypothetical example, let’s suppose that School A was identified as effective in 2000; however, in 2001 there was an influx of immigrant children who didn’t speak English and the average test scores in School A dropped suddenly. School A was ranked as ineffective in 2001, although the administrators, faculty and staff worked as hard as (or perhaps even harder than) they did in the previous year. Accountability systems based on cross-sectional designs only provide “snapshots” reporting of school performance. A credible accountability system should employ a longitudinal design, in which cohorts of students are followed (by assigning each of them a unique identification number) as they move through the school system.

Third, school effectiveness researchers have argued convincingly that single-level models (e.g., ordinary least squares, ANOVA) have limited use in educational research, and have called
for studies based on multilevel models (Kennedy & Mandeville, 2000; Reynolds & Teddlie, 2000). Empirical investigations of school effects usually involve searching for statistical associations between variables measured at different levels: student-level, classroom-level, and/or school-level. These variables are hierarchical in nature because students are nested within classrooms (two-level), classrooms are nested within schools (two-level), or students are nested within classrooms which in turn are nested within schools (three-level) (Lee, 2000).

Furthermore, studies of student growth often involve repeated observations within individuals, which create another layer of nested structure in addition to the hierarchical structure generated by organizational settings (e.g., Raudenbush & Bryk, 2002). For example, in a three-level hierarchical model, repeated measures can be viewed as being nested within individual students who in turn are nested within schools.

Before the advent of hierarchical linear modeling, researchers used to aggregate lower level variables to a higher level or disaggregate higher level variables to a lower level. For example, suppose a researcher was interested in studying the effects of principal leadership, school social climate and school SES composition (i.e., percentage of students receiving free or reduced price lunch) on student achievement. The predictor variables are all measured at the school level, while the criterion variable (i.e., student achievement) is measured at the student level. To aggregate the data to the school level, researchers can compute an average achievement score for each school and then use schools as the units of analysis; to disaggregate the data to the student level, the researchers can assign the same value to all students in the same school for each predictor variable. Neither method is appropriate. The former leads to loss of student level information and the latter assumes that all predictor variables influence students to the same degree.
Hierarchical Linear Modeling (HLM)

HLM is a relatively new statistical technique. It provides a solution to the nesting problems that usually occur in organizational research. Hierarchical models come under a variety of names. They are labeled as multilevel linear models in sociological research, mixed-effects models and random-effects models in biometric applications, random-coefficient regression models in econometrics literature, and covariance components models in statistical literature (Raudenbush and Bryk, 2002). HLM provides possible solutions to the following three major problems caused by using single-level models to analyze multilevel data (Lee, 2000).

Aggregation Bias

Aggregation bias occurs when a variable measured at the individual level (e.g., student SES) not only influences the individuals but also has an effect at the group level after it is aggregated. As pointed out by Lee (2000), “the average SES of a school may influence a student’s academic achievement above and beyond his or her own SES” (p. 127). In other words, a student’s academic achievement is influenced not only by his/her own SES status (i.e., receiving free or reduced price lunch), but also by the average SES of the school (i.e., 70% of the students in this school receiving free or reduced price lunch). In hierarchical models the same variable, SES in this case, can be modeled at different levels, so aggregation bias is no longer a problem.

Misestimated Standard Errors

Standard errors can be misestimated when dependent cases are treated as if they are independent. For example, students in the same school are not independent of each other because (a) they are not randomly assigned to schools in the first place, and (b) they share many educational experiences and become more similar over time. With Ordinary Least Squares
(OLS) regression, students in the same school are mistakenly treated as independent cases, which lead to overestimated or underestimated standard errors (Lee, 2000). This problem is fixed in hierarchical models because each individual has a separate error term and each error term is assigned a pair of subscripts that help identify the individual and its group status.

Heterogeneity of Regression slopes.

In single-level models, only one set of regression slopes are estimated. However, “relations between characteristics of students (such as race, ethnicity, SES) and academic achievement may vary across schools and may be functions of group-level variables” (Lee, 2000, p. 128). In other words, multiple sets of regression slopes should be estimated, one for each school. For example, variables such as race, ethnicity and SES may have weaker effects on student achievement in more equitable schools than in less equitable schools. This problem is fixed in hierarchical models because each school has a separate model and therefore has a separate set of slopes.

Findings from school effectiveness research and advancement in statistical modeling serve as foundational work for developing sounder and more sophisticated accountability models. For a long time, accountability has been regarded largely as a political issue and handled by policy-makers who might not be fully aware of the findings from school effectiveness research and advancement in statistical modeling. Therefore, simplistic measures, such as those employed in status models and improvement models, are widely used and accepted. With the enactment of NCLB, under which stakes associated with testing results have been raised to a new high, it becomes increasingly difficult to tolerate accountability models that are invalid and inaccurate. Therefore, there has been an increasingly loud call from a variety of stakeholders for better accountability models.
Four Types of Accountability Models

There exist four general types of accountability models: status models, improvement models, growth models, and value-added models. Status models (e.g., AYP) and improvement models (e.g., Safe Harbor) are approved accountability models under NCLB. Growth models are being piloted by the U. S. Department of Education as a possible supplement to the current AYP model. Two states, Tennessee and North Carolina, have been permitted to use their proposed growth models for NCLB purpose, while about 20 growth models proposed by other states are still under review. Value-added models are regarded by many as a special type of growth models, but value-added models have not yet been approved by the U. S. Department of Education for NCLB purpose. The characteristics of each of the four general types of the accountability models are discussed in detail in this section.

Status Models

A status model takes a snapshot of the performance of a subgroup/school/LEA at one point in time, and compares that proficiency level with an established target (Goldschmidt et al., 2005). For example, the AYP model under NCLB is a status model. AYP takes a snapshot of the percent of students meeting or exceeding the proficiency level of achievement in a specific school year, and then compares it to the target proficiency level of that year. The target proficiency level is called Annual Measurable Objective (AMO), and it is established based on three criteria: (1) the ultimate goal of 100% proficiency for all students, (2) the baseline\(^5\), and (3) equally-spaced intermediate goals. AMO is the same for all subgroups, schools and LEAs statewide for a given subject area and grade level (Riddle, 2005). As shown in Figure 2.8, the fundamental questions under status models are: How are students performing this year? Do their

\(^5\) A state’s AYP baseline will be “the higher of either the percentage of students at the proficient level (or higher) who attend the school that is at the 20\(^{th}\) percentile in the state, or the percentage of students in the state’s lowest achieving subgroups who score at the proficient level (or higher)”. (NASBE, 2005, p. 3)
performances meet the established target? The performance of this year is usually not compared to the performance of the previous year.

**Figure 2.8**
An Illustration of a Status Model (Goldschmidt et al., 2005, p. 3)

**Improvement Models**

An improvement model compares the change of status at two points in time. For example, the percentage of $4^{th}$ graders making proficiency in 2000 (i.e., status in 2000) can be compared to the percentage of $4^{th}$ graders making proficiency in 2001 (i.e., status in 2001). The resulting difference provides a measure of improvement. Safe Harbor, the secondary indicator of school performance under NCLB, is based on an improvement model. Safe Harbor applies if a school fails to make AYP in a particular year but reduces the number of below proficient scores of a student group by 10% from the previous year’s comparable group (CCSSO, 2005; Riddle, 2005). As shown in Figure 2.9, the fundamental question answered by improvement models is: Compared with students in the same grade/school/LEA last year, are students doing better this year? It is worth noting that the two groups under comparison are two different groups (e.g., this year’s $4^{th}$ graders vs. last year’s $4^{th}$ graders). Improvement / change from year to year is
subjected to the influence of many factors other than schooling. Therefore, rewarding/punishing
schools or LEAs based on the result of improvement models is questionable.

![Figure 2.9](image)

**Improvement Model**

The basic question under this model is, “On average, are students doing better this year as compared to students in the same grade last year?”

- **Year**
- **Year**

Growth Models

The fundamental difference between growth models and the two models introduced previously (i.e., status models and improvement models) is that in growth modeling cohorts of students are tracked (usually through unique identification numbers) over time as they move along their K-12 educational careers (Riddle, 2005). The intention is to see whether students have made any progress. For example, the performance of 4th graders in 2000 can be compared to the performance of the same students in 1999 when they were in the third grade (Goldschmidt et al., 2005). The change in score from one year to the other is the actual growth, which is usually compared to the expected growth. A school/LEA is considered to be making adequate progress if its actual growth exceeds the expected growth. As shown in Figure 2.10, the fundamental questions under growth models are: How much did the cohorts grow? Did the growth exceed the expected growth?
Modeling growth can be a complicated issue. First, it usually requires at least three years of student level data, with measurement occasions matched to individual student record. Second, it may require vertical scaling, which involves developing a continuous scale that has the same meaning across grades. According to Lissitz, Doran, Schafer, and Willhoft (2006):

Tests are scaled so that the process of subtracting the prior-year’s score from the current year’s score earned by the same students in the same subject is interpretable. For example, say third-grade scores are scaled from 300 to 399 and fourth grade scores from 400 – 499, and so on. Assuming an interval scale, a difference of 100 points would represent a year’s growth anywhere on the scale. Further, a difference of 100 points would represent the same year’s growth between any two adjacent grades. Comparisons of growth could then be made meaningfully between schools and perhaps even teachers and students whose prior achievement (starting points) were different. (p.5)

Value-added Models

Value-added models are a very special class of growth models (Lissitz et al., 2006). Similar to growth models, value-added models also requires (1) longitudinal, student-level data, and (2) vertical scaling, or some other techniques that can be used to create a developmental

---

55

---

6 “The simplest form of a growth model is a model that uses gains from one year to the next as the outcome. Growth models with three or more time points typically model the individual outcome as a function of time where the model generates a growth trajectory for each individual” (CCSSO, 2005, p. 15)
scale. The key distinction between the two is whether student demographic characteristics or previous achievement have been adjusted / accounted for so that the variances in student achievement can be better attributed to individual teachers, schools, or LEAs (e.g., Riddle, 2006).

The distinction between simple growth models and VAMs [value-added models] primary lies in the motivation and the types of inferences that one seeks. The key motivation for implementing a value-added model is to obtain empirically an estimate of a teacher’s effectiveness or school’s effectiveness by decomposing the variance of the test scores into portions that are explained by student inputs (e.g., prior achievement), and into other portions that are believed to be directly related to the (presumably) causal inputs of the current classroom teacher or the school. In other words, VAMs attempt to answer the question, how much value has a teacher added to a student’s learning? (Lissitz et al., 2006, p. 8).

As shown in Figure 2.11, subtracting the score of Year \( \times \) from the score of Year \( \times +1 \) gives us the actual growth. The actual growth will then be compared to the expected growth. The difference between the actual growth and the expected growth is called value added. The total value-added can be further dissected, through statistical techniques, into value-added by teachers, value-added by schools, etc.

![Figure 2.11](image)

Figure 2.11
An Illustration of a Value-added Model (Goldschmidt et al., 2005, p. 5)
The earliest value-added model to be used statewide in the USA was the Tennessee Value-added Assessment System (TVAAS), which was developed by Dr. William Sanders (Sanders & Horn, 1995). TVAAS sets different goals for different students/designated student groups/schools based on the student’s/student group’s/school’s previous scores. Growth is calculated by subtracting the expected growth from the actual growth. Therefore, the model can be used to determine whether students/designated student groups/schools are below, at, or above their level of expected performance. The model is also capable of estimating the unique contribution of the teacher and the school to a child’s growth in scores over time. Since 1992, schools in Tennessee had been classified into Grade A, B, C, D, or F based on their unique contribution (as calculated by the TVAAS model) and other additional information (U. S. Government Accountability Office, 2006).

In addition to TVAAS, there are many other value-added models, such as McCaffrey et al.’s RAND Model (McCaffrey, Lockwood, Koretz, Louis, and Hamilton, 2004), Bryk et al.’s Chicago Public School Productivity Model (Bryk, Thum, Easton, & Luppescu, 1998), Choi et al.’s CRESST Student Growth Distribution Model (Choi, Seltzer, Herman, & Yamashiro, 2004), Doran et al.’s REACH Model (Doran & Izumi, 2004), etc. Those models are unique in their own way, but “the differences in inferences based on different VAMs will be much less than the differences in inferences between a VAM and a status model such as AYP” (Goldschmidt, 2005, p. 16)

For a comparison among status models, growth models, and value-added models, please refer to Appendix 1; for a comparison among the various types of value-added models, please refer to Appendix 2.
Comparisons of Different School Accountability Models

A number of research studies have been conducted to examine the appropriateness of the models used in educational accountability systems (e.g., Choi, Goldschmidt, & Yamashiro, 2006; Choi, Seltzer, Herman, & Yamashiro, 2004; Linn & Haug, 2002; Meyer, 1995; Raudenbush, 2004; Webster, Mendro, Orsak, & Weerasinghe, 1998). For example, Webster, et al (1998) compared value-added models against a number of statistical models for estimating school and teacher effects on student learning and other educational outcomes. Webster, et al believed that “fairness” should be the criterion used to judge the appropriateness of statistical models designed to rank schools and teachers. Therefore, school effectiveness indices produced by (1) unadjusted student test scores, (2) gain scores, and (3) various ordinary least squares (OLS) models were compared against those produced by various hierarchical linear models (HLM), which was the methodology of choice. The first three types of models were judged as appropriate if the school or teacher effectiveness indices produced by them (1) had a high correlation with the school or teacher effectiveness indices produced by HLM, and (2) had a low correlation with individual student background variables and aggregate school variables. Webster et al concluded that the two-stage, two-level (students nested within schools) model is the model of choice for estimating school effects and that the two-stage, two-level (students nested in teachers) model is the model of choice for estimating teacher effects.

Another example of such comparison studies comes from Choi, Goldschmidt, and Yamashiro (2006), in which performance classifications based on AYP results are compared to those based on results from an array of value-added models. The longitudinal dataset used in the study came from an urban school district in the Pacific Northwest. The outcomes of interest in this study are ITBS reading scores for third graders in 2001 and these same students’ test scores
in 2003 when they were in fifth grade. Choi, et al first classified the schools into AYP schools (i.e., schools that have a high enough proficiency rate) and non-AYP schools. It was found that among the 51 AYP schools only 12 had an estimated gain that was statistically greater than the district mean gain. In contrast, almost half of the non-AYP schools have gains greater than the district average. Choi, et al also compared an array of value-added models that differ in the number of background variables adjusted. It was concluded that value-added models provide both the most informative and the most valid picture of school performance.

Another contributor to this line of research was Meyer (1995). Meyer did a simulation study comparing results produced by average test scores and those produced by value-added models. It was concluded that simplistic indicators used to assess school performance are highly flawed and, therefore, are of limited value. Value-added models provide a promising alternative, but the validity of the results are dependent on a number of factors, such as the quality and appropriateness of the test, the adequacy of the control variables included in the appropriate statistical models, and the technical validity of the statistical models used to construct the indicators.

The proposed study will add to the literature on the comparison of models used in educational accountability systems in a unique way. First, it will compare the same schools across four different models looking for consistencies and inconsistencies in the classification of those schools’ performance. Second, it will use data derived from Louisiana school accountability system, thereby allowing for comparisons that should have implications for education accountability in the state.
CHAPTER THREE: METHODOLOGY

Introduction

The preceding chapters introduced four generic types of accountability models: (1) status models, (2) improvement models, (3) growth models, and (4) value-added models. As each of the four generic types of accountability models encompasses a number of specific models, a representative model was chosen for each generic model, and the selected models were run on the data collected on Louisiana students and schools. Specifically, the AYP model under NCLB was chosen as the representative of the status models; Louisiana’s current school and district accountability model was chosen as the representative of the improvement models; North Carolina’s Modified ABCs Model was chosen as the representative of growth models; and a value-added model specified by the author of this study based on the model that has been used to evaluate Massachusetts charter schools (Massachusetts Department of Education, 2006) was chosen as the representative of value-added models. Therefore, this study consists of four phases:

- Phase I: The AYP Model (Status Model)
- Phase II: Louisiana’s Current School Accountability Model (Improvement Model)
- Phase III: North Carolina’s Modified ABCs Model (Growth Model)
- Phase IV: A Value-Added Model

In this chapter, the research design components are discussed separately for each of the four models in the order of (1) model specification, (2) sample, (3) measurement, and (4) data analysis. These design components are discussed separately for each model because different models are required for each set of slightly different research questions. Consequently, the associated design components are also likely to be different.
Restatement of Research Questions

The following is a restatement of the 12 research questions, which were originally proposed in Chapter 1. These research questions are mostly parallel questions. Specifically, for each model the student population as a whole was examined first and then the model’s effect on the four subgroups were examined.

Phase I (Status Model) Research Questions:

1) According to the Status Model, how many K-5 schools in Louisiana can be categorized as making AYP for the school year of 2004 – 05?

2) If a K-5 school is categorized as making AYP based on the student population as whole, will each of the following subgroups also make AYP?
   a. African American Students
   b. Economically Disadvantaged Students
   c. Students with Limited English Proficiency
   d. Students with Disabilities

Phase II (Improvement Model) Research Questions:

3) According to the Improvement Model, how many K-5 schools in Louisiana can be categorized as making AYP for the school year of 2004 – 05?

4) If a K-5 school is categorized as making AYP based on the student population as a whole, will each of the following subgroups also make AYP?
   e. African American Students
   f. Economically Disadvantaged Students
   g. Students with Limited English Proficiency
   h. Students with Disabilities
According to the Improvement Model, how many K-5 schools in Louisiana meet or exceed the expected improvement?

Phase III (Growth Model) Research Questions:

6) According to the Growth Model, how many K-5 schools in Louisiana can be categorized as making AYP for the school year of 2004-05

7) If a K-5 school in Louisiana is categorized as making AYP based on the student population as a whole, will each of the following subgroups also make AYP?
   i. African American Students
   j. Economically Disadvantaged Students
   k. Students with Limited English Proficiency
   l. Students with Disabilities

Phase IV (Value-Added Model) Research Questions:

8) According to the value-added model, which schools add the most “value” to student learning?

9) Do schools that are categorized as making AYP by the status model add more “value” to student learning than those categorized as not making AYP?

10) Do schools that are categorized as making AYP by the improvement model add more “value” to student learning than those categorized as not making AYP?

11) Do schools that are categorized as making AYP by the growth model add more “value” to student learning than those categorized as not making AYP?

12) Do the four models generate consistent or inconsistent conclusions regarding the effectiveness of the same set of K-5 schools in Louisiana?
Phase I Model Specifications

The NCLB AYP model was chosen as the representative of status models. The AYP model is the primary indicator of school performance under NCLB. It computes a simple frequency count of the students scoring at or above the proficient level of academic achievement, and compares the observed proficiency rate against the expected proficiency rate. A subgroup/school/LEA is considered as making AYP if the observed proficiency rate reaches or exceeds the expected proficiency rate.

The expected proficiency rate, which is also called AMO (Annual Measurable Objective), is set based on three criteria: (1) the ultimate goal of 100% proficiency for all students by the school year of 2013-14, (2) the baseline, and (3) equally-spaced intermediate goals. AMO is the same for all subgroups, schools and LEAs statewide for a given subject area (Riddle, 2005). As shown in Table 3.1, the AMO for Math for the school year of 2004-05 in the state of Louisiana is for all subgroups/schools/LEAs to reach a proficiency rate of 41.8% (Bulletin 111, 2006).

Table 3.1
The AMOs for Math (Bulletin 111, 2006)

<table>
<thead>
<tr>
<th>School Year</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-2002</td>
<td></td>
</tr>
<tr>
<td>2002-2003</td>
<td>30.1%</td>
</tr>
<tr>
<td>2003-2004</td>
<td>30.1%</td>
</tr>
<tr>
<td>2004-2005</td>
<td>41.8%</td>
</tr>
<tr>
<td>2005-2006</td>
<td>41.8%</td>
</tr>
<tr>
<td>2006-2007</td>
<td>41.8%</td>
</tr>
<tr>
<td>2007-2008</td>
<td>53.5%</td>
</tr>
<tr>
<td>2008-2009</td>
<td>53.5%</td>
</tr>
<tr>
<td>2009-2010</td>
<td>53.5%</td>
</tr>
<tr>
<td>2010-2011</td>
<td>65.2%</td>
</tr>
<tr>
<td>2011-2012</td>
<td>76.9%</td>
</tr>
<tr>
<td>2012-2013</td>
<td>88.6%</td>
</tr>
<tr>
<td>2013-2014</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
It is worth noting that Louisiana uses a backloaded trajectory, which requires only modest increases of proficiency rate in early years and much larger gains in later years. For example, the required proficiency rate for Math increases from 30.1% to 53.5% (i.e., a 23.4% increase) during the eight years between 2002 and 2009. However, an increase from 65.2% to 100% (i.e., a 34.8% increase) is required for the last four years between 2010 and 2013. At present, Louisiana has managed to keep a very low percentage (less than 10%) of her schools in school improvement (Louisiana Department of Education, 2006). However, the backload trajectory will only work for a few years, since the ultimate goal is 100% proficiency by the school year of 2013-14 (Porter, Linn, & Trimble, 2005).

Phase I Sample

The Phase I sample consisted of all 24,159 students in Grade 4 for the school year of 2004-05 in all 297 K-5 schools in the state of Louisiana. Students with missing information on race/ethnicity, gender, SES (socioeconomic status), LEP (limited English proficiency) status, or SPED (special education) status were eliminated because each subgroup’s proficiency rate must be examined. Only the fourth graders were included in the Phase I study because Grade 4 was the only grade that has a statewide criterion-referenced test (CRT). Grades 3 and 5 used a norm-referenced test (NRT), in which students were compared to each other (e.g., percentile rank) as opposed to a set of content standards. Test scores on NRT could not be judged using levels like advanced, proficient, basic, approaching basic, and unsatisfactory. Therefore, it was not possible to do a frequency count of the students scoring at or above proficiency for students in Grades 3 and 5.

Of the 24,159 students included in the Phase I study, about 47.61% were females while 52.39% were males. African American students accounted for about 53.67% of the total student
population statewide, with White, Hispanic, Asian, and Native American students accounting for 42.55%, 2.17%, 1.26%, and 0.35% respectively. Students were not evenly distributed in terms of race/ethnicity, since some schools (40 schools) had lower than 10% African American students while others (78 schools) had higher than 90% African American students. Students were not evenly distributed in terms of SES, either. On average, 70.42% of the total student population received free or reduced-price lunch. However, some schools (15 schools) had less than 30% of the students receiving free or reduced-price lunch while others (82 schools) had more than 90% of the students enrolled in the federal lunch program. The K-5 schools included in Phase I study usually had a very small percentage of students with limited English proficiency, with LEP students accounting for 1.34% of the total student population. Finally, a high percentage of students were enrolled in special education programs. For example, some schools (15 schools) had more than 35% SPED students. The average percentage of SPED students was 21.42%. (See Table 3.2)

Table 3.2
Student Demographic Information for Phase I Study

<table>
<thead>
<tr>
<th></th>
<th>Statewide/Average</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(24,159 Students, 297 Schools)</td>
<td>81</td>
<td>203</td>
<td>19</td>
</tr>
<tr>
<td>Native American</td>
<td>0.35%</td>
<td>11.6%</td>
<td>0%</td>
</tr>
<tr>
<td>Asian</td>
<td>1.26%</td>
<td>19.7%</td>
<td>0%</td>
</tr>
<tr>
<td>African American</td>
<td>53.67%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>2.17%</td>
<td>40.7%</td>
<td>0%</td>
</tr>
<tr>
<td>White</td>
<td>42.55%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>1.34%</td>
<td>38.5%</td>
<td>0%</td>
</tr>
<tr>
<td>Students with Disabilities</td>
<td>21.42%</td>
<td>50.0%</td>
<td>4.4%</td>
</tr>
<tr>
<td>Economically Disadvantaged</td>
<td>70.42%</td>
<td>100%</td>
<td>9.7%</td>
</tr>
</tbody>
</table>

Phase I Measurement

LEAP 21 Math Test was used as the measurement instrument. The Louisiana Educational Assessment Program for the 21st Century (LEAP 21) was a criterion-referenced testing (CRT)
program that was designed to measure how well a student had mastered state content standards in subject areas of English Language Arts (ELA), Mathematics, Science, and Social Studies. The LEAP 21 was administered at grades 4 and 8, and it had five achievement ratings: Advanced, Mastery, Basic, Approaching Basic, and Unsatisfactory (Louisiana Department of Education, 2006b). For test scores used in this study (fourth grade math), a score between 419 and 500 was within the Advanced level; a score between 370 and 418 was within the Proficiency (or Mastery) level; a score between 315 and 369 was within the Basic level; a score between 282 and 314 was within the Approaching Basic level; and a score between 100 and 281 was within the Unsatisfactory level (Louisiana Department of Education, 2006a).

NCLB requires each state to report the percentage of students scoring at or above the proficiency level of achievement but grants the state full authority to define what the proficiency level of achievement is. Louisiana, like 22 other states, started counting the “basic” level and above as “proficient” (Porter, Linn, & Trimble, 2005). In order to be consistent with those changes, a score at or above 315 (instead of 370) was regarded as within the “proficient” level of achievement for the Phase I data analyses.

Phase I Data Analysis

Phase I data analysis consisted of three steps. In Step One, students with math score equal to or greater than 315 were assigned a code of 1 while students with a math score less than 315 were assigned a score of 0. The percentage of students coded as 1 was calculated, and the observed proficiency rate for each school was compared to the target proficiency rate of 41.8%. Schools that had 41.8% or more of its students reaching proficiency were labeled as making AYP and vice versa. In Step Two, a 99% confidence interval around AMO was calculated for schools that failed to make AYP in Step One. The algorithm used to calculate confidence
intervals is presented in Equation 3.1. Confidence interval made it easier for schools to pass, especially for those that had a small student population. Schools were re-categorized as making AYP if their proficiency rate was equal or higher than the lower band of the AMO confidence interval. Finally in Step Three, safe harbor was applied if a school still failed to make AYP in Step Two. To enter Safe Harbor, a school had to reduce its non-proficient students by 10% from last year’s comparable group. The algorithm for Safe Harbor calculation was presented in Equation 3.2.

\[
\text{AMOCI}^7 = 41.8 - 2.326 \times \text{SQRT}\left(\frac{41.8 \times (100 - 41.8)}{\text{#Testers}}\right) \tag{3.1}
\]

\[
\text{Safe_Harbor} = \text{ProficiencyRate2005} + \\
+ 1.15 \times \text{SQRT}\left(\frac{\text{ProficiencyRate2005} \times (1 - \text{ProficiencyRate2005})}{\text{#Testers2005}} - (1 - (1 - \text{ProficiencyRate2004})) \times 0.9\right) \tag{3.2}
\]

If a school failed to pass in all three steps it was regarded as failing AYP of the year. This answers Research Question # 1. For those schools that were categorized as making AYP, the proficiency rate of the four subgroups was also examined through the same three steps presented above. A school was considered as making AYP only if the student population as a whole and each of the subgroups passed AYP. This answers Research Question # 2. Graph 3.1 is a flowchart of the data analysis process for the Status Model.

Phase II: Improvement Model

Phase II Model Specifications

Louisiana’s current school and district accountability model was chosen as the representative of Improvement Models. It is a rather sophisticated model in which a single School Performance Score (SPS) is calculated annually for each school. The SPS score ranges from 0.0 to 120.0 and beyond, and a score of 120.0 indicates that a school has reached

\footnote{AMOCI refers to the annual measurable objective after confidence interval is applied.}
Figure 3.1
A Flowchart of the Status Model

The calculation of SPS was a three-step process. In Step One, grade-level indices were calculated for each student. The algorithms for grade-level index calculation were presented in Equation 3.3-3.5\(^8\). Step Two involves combining the grade-level indices into a CRT (Criterion-Referenced Test) index and a NRT (Norm-Referenced Test) index. In this study, for example, Grades 3 and 5 were implemented the NRT test (i.e., the ITBS Test) while Grade 4 was implemented the CRT test (i.e., the LEAP Test). Therefore, the indices for Grade 3 and 5 were

\[^8\) Please refer to the document entitled “Recommendations Regarding the Louisiana School Accountability System” (1998) by the Accountability Technical Advisory Committee regarding why the index for each indicator is calculated in such a way.
combined to form the NRT index while the index for Grade 4 was used as the CRT Index. In Step Three, the SPS score was determined by weighting the CRT index and the NRT index, as well as student attendance index. The CRT index counts for 60%; the NRT index counts for 30%; and attendance index counts for the remaining 10% (The Accountability Technical Advisory Committee, 1998; Public Affairs Research Council of Louisiana, 1999; Bulletin 111, 2006). (See Table 3.3)

Table 3.3
Louisiana School Accountability Indicators and Weighting (Bulletin 111, 2006, p. 1)

<table>
<thead>
<tr>
<th>Through 2005, K-12 Indictors and Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRT (60% K-12)</td>
</tr>
<tr>
<td>Grades 4, 8, 10, 11</td>
</tr>
<tr>
<td>NRT (30% K-12)</td>
</tr>
<tr>
<td>Grades 3, 5, 6, 7, 9</td>
</tr>
<tr>
<td>Attendance (10% K-6; 5% 7-12)</td>
</tr>
<tr>
<td>Grades K-12</td>
</tr>
<tr>
<td>Dropout Rate (5% 7-12)</td>
</tr>
<tr>
<td>Grades 7-12</td>
</tr>
</tbody>
</table>

In this study, only the CRT and NRT indices were used to calculate the SPS index because the other three models (i.e., the status model, the growth models, and the value-added models) only used test scores as indicators of school performance. Attendance/dropout was not part of the model. Therefore, the weights of CRT and NRT scores had been increased to 66.7% and 33.3% respectively.

Grade 3 Index = (4.181 * SS) – 693.6 \hspace{1cm} (3.3)

Grade 5 Index = (3.101 * SS) – 599.3 \hspace{1cm} (3.4)

Grade 4 Index = \( \frac{N_{\text{advanced}} \times 200 + N_{\text{proficient}} \times 150 + N_{\text{basic}} \times 100 + N_{\text{approaching basic}} \times 50 + N_{\text{unsatisfactory}} \times 0}{N_{\text{total}}} \) \hspace{1cm} (3.5)

Where

SS = scaled score

\( N_{\text{advanced}} \) = the number of students scoring at the advanced level

\( N_{\text{proficient}} \) = the number of students scoring at the proficient level
\[ N_{\text{basic}} = \text{the number of students scoring at the basic level} \]

\[ N_{\text{approaching basic}} = \text{the number of students scoring at the approaching basic level} \]

\[ N_{\text{unsatisfactory}} = \text{the number of students scoring at the unsatisfactory level} \] (Bulletin 111, 2006, p. 8)

Once a single School Performance Score was determined for each school, a performance label was assigned based on a set of absolute standards. For the school year of 2004-05\(^9\), schools with SPS lower than 60.0 were labeled as “academically unacceptable”; schools with SPS of 60.0 to 79.9 were labeled as “One Star”; schools with SPS of 80.0 to 99.9 were labeled as “Two Stars”; schools with SPS between 100.0 to 119.9 were labeled as “Three Stars”; schools with SPS between 120.0 and 139.9 were labeled as “Four Stars”; and finally schools with SPS 140 or above were labeled as “Five Stars” (Bulletin 111, 2006). (See Table 3.4)

Table 3.4
Performance Labels (Bulletin 111, 2006, p. 16)

<table>
<thead>
<tr>
<th>Performance Label</th>
<th>School Performance Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academically Unacceptable</td>
<td>60.0 or below</td>
</tr>
<tr>
<td>One Star</td>
<td>60.0-79.9</td>
</tr>
<tr>
<td>Two Stars</td>
<td>80.0-99.9</td>
</tr>
<tr>
<td>Three Stars</td>
<td>100.0 – 119.9</td>
</tr>
<tr>
<td>Four Stars</td>
<td>120.0 – 139.9</td>
</tr>
<tr>
<td>Five Stars</td>
<td>140.0 or above</td>
</tr>
</tbody>
</table>

School Performance Score was also used to assign a “growth” label for each school. For the school year of 2004-05 the weighted average SPS 2002-03 and 2003-04 was used as baseline, which was compared against the SPS of 2004-05 to get the actual growth. The actual growth was then compared to the target growth to see whether schools had met or exceeded the growth target. The algorithm for growth target calculation is detailed in Equation 3.6 (Bulletin 111, 2006).

---

\(^9\) From this point forward, the school year of 2004-05 will be referred as 2005; the school year of 2003-04 will be referred to as 2004, and so forth.
Schools exceeding growth target by 5 points or more were labeled as “exemplary academic growth”, schools meeting or exceeding growth target by fewer than 5 points were labeled as “recognized academic growth”, schools that were improving but not meeting growth target were labeled as “minimal academic growth”, schools with SPS declining up to 5 points were labeled as “no growth”, and schools with SPS declining more than 5 points were labeled as “school in decline. (See Table 3.5)

\[
\text{Growth Target} = \left[ \text{PropRE} \times (120 - \text{SPS}) / N \right] + \\
\left[ \text{PropSE} \times (120 - \text{SPS}) / 2N \right] + \left[ \text{PropLEP} \times (120 - \text{SPS}) / 2N \right] 
\]

Where

- PropRE = the proportion of regular education students in the school
- SPS = Baseline School Performance Score
- N = the number of remaining years to 2014
- PropSE = the proportion of special education students
- PropLEP = the proportion of students with limited English proficiency (Bulletin 111, 2006, p. 11)

Table 3.5
Growth Labels (Bulletin 111, 2006, p. 16)

<table>
<thead>
<tr>
<th>Label</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exemplary Academic Growth</td>
<td>A school that makes its growth target, all subgroups grow at least two points, and the school is not in SI(^{10}).</td>
</tr>
<tr>
<td>Recognized Academic Growth</td>
<td>A school that makes its growth target but any subgroups does not grow at least two points and/or the school is in SI.</td>
</tr>
<tr>
<td>Minimal Academic Growth</td>
<td>A school improving (at least 0.1 points) but not meeting its growth target.</td>
</tr>
<tr>
<td>No Growth</td>
<td>A school with a change in SPS of 0 to -2.5 points.</td>
</tr>
<tr>
<td>School in Decline</td>
<td>A school with a declining SPS (more than -2.5 points).</td>
</tr>
</tbody>
</table>

\(^{10}\) SI means School Improvement. For details please refer to Bulletin 111 (2006).
A flow chart of how Louisiana’s School and District Accountability model works is presented in Figure 3.2. It is worth noting that “growth” here is actually what has been defined as “improvement” in this study. The performance of 2005\textsuperscript{11} (i.e., the status of 2005) was compared against the weighted average performance of 2003 and 2004 (i.e., the combined status of 2003 and 2004). The result would be an improvement (i.e., status change). Since the data used were not panel data (i.e., individual students have not been tracked over time), this model was not categorized as a growth model as defined in this study.

Figure 3.2: A Flowchart of the Improvement Model

Phase II Sample

The Phase II sample consisted of all 63,279 students enrolled in Grades 3, 4, and 5 for 2004-05 in all 297 K-5 schools in the state of Louisiana. Students with missing information on

\textsuperscript{11} From this point forward, 2005 means the school year of 2004-05, 2004 means the school year of 2003-04, and so on and so forth.
race/ethnicity, gender, SES (socioeconomic status), LEP (limited English proficiency) status, or SPED (special education) status were eliminated because each subgroup’s proficiency rates needs to be examined. Of the 63,279 students included in the Phase II study, about 48.30% were females while 51.70% were males. African American students accounted for about 50.26% of the total student population statewide, with White, Hispanic, Asian, and Native American students accounting for 45.51%, 2.48%, 1.34%, and 0.41%, respectively. About 67.91% of the students received either free or reduced-price lunch, and about 19.65% of the students were learning disabled. There were a very small percentage of students with limited English proficiency, with LEP students accounting for only 1.47% of the total student population. (See Table 3.6) Since Phase II analysis used the average SPS of 2003 and 2004 as baseline, the demographic information of the students in 2003 and 2004 are also included.

Table 3.6
Student Demographic Information for Phase II Study

<table>
<thead>
<tr>
<th></th>
<th>Year 2005</th>
<th>Year 2004</th>
<th>Year 2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schools (N)</td>
<td>297</td>
<td>297</td>
<td>297</td>
</tr>
<tr>
<td>Students (n)</td>
<td>63,279</td>
<td>65,867</td>
<td>66,277</td>
</tr>
<tr>
<td>Male (1)</td>
<td>51.70%</td>
<td>51.52%</td>
<td>51.47%</td>
</tr>
<tr>
<td>Native American (1)</td>
<td>0.41%</td>
<td>0.40%</td>
<td>0.41%</td>
</tr>
<tr>
<td>Asian (2)</td>
<td>1.34%</td>
<td>1.33%</td>
<td>1.30%</td>
</tr>
<tr>
<td>African American (3)</td>
<td>50.26%</td>
<td>49.58%</td>
<td>49.34%</td>
</tr>
<tr>
<td>Hispanic (4)</td>
<td>2.48%</td>
<td>2.32%</td>
<td>2.23%</td>
</tr>
<tr>
<td>White (5)</td>
<td>45.51%</td>
<td>46.38%</td>
<td>46.72%</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>1.47%</td>
<td>1.21%</td>
<td>0.99%</td>
</tr>
<tr>
<td>Students with Disabilities</td>
<td>19.65%</td>
<td>19.40%</td>
<td>18.64%</td>
</tr>
<tr>
<td>Economically Disadvantaged</td>
<td>67.91%</td>
<td>66.75%</td>
<td>65.42%</td>
</tr>
</tbody>
</table>

Phase II Measurements

LEAP 21 Math (CRT) and ITBS Math (NRT) were used jointly in Phase II analysis. LEAP 21 Math was used for Grade 4 while ITBS was used for Grades 3 and 5. The Iowa Test of Basic Skills (ITBS) is a nationally standardized norm-referenced Testing (NRT) program. In
Louisiana, ITBS was administered at grades 3, 5, 6, 7, 9 in subject areas of Reading, Language, Mathematics, Science, Social Studies, and Sources of Information. The ITBS had been used in Louisiana since 1998 (Louisiana Department of Education, 2006a).

Phase II Data Analysis

The first step in Phase II analysis was calculating the CRT index for Grade 4, and the NRT indices for Grades 3 and 5. Then these indices were weighted and combined to make a single School Performance Score for each school. Schools with a SPS lower than 60 were categorized as “academically unacceptable” and therefore were reviewed as ineffective, and vice versa (Research Question # 3). For schools that had a SPS higher than 60, the SPS for the four subgroups were also calculated. The purpose was to see whether schools were equally effective in educating disadvantaged students (Research Question # 4). Improvement in SPS was also examined, and the actual improvement was compared against the expected improvement (Research Question # 5).

Phase III: Growth Models

Phase III Model Specifications

North Carolina’s Modified ABCs Model was chosen as the representative of growth models. North Carolina calculates a four-year growth trajectory for all non-proficient students. This growth trajectory would classify a student as “proficient” within four years in the tested grades so long as the student meets the trajectory’s intermediate targets. In other words, students who are proficient and students who are on their four-year trajectory towards proficiency are counted as proficient (North Carolina Department of Public Instruction, 2006a). Two concepts are important in understanding how the Modified ABCs Model works, which are (1) change score and (2) four-year growth trajectory.
Change Scale

One of the main concepts of the North Carolina’s Modified ABCs model is “change score”, which is also called “c-score”. The algorithm for c-score calculation is presented in Equation 3.7. C-score looks like a z-score except that all the values on the right hand side of the equation were based on the normative distribution of student performance in the standard setting year\textsuperscript{12}. In other words, student performance is always compared back to the standard setting year. “What is different about the c-scale from the normative scales is that there is no reason why all students in the state could not score above “0” in any year after the standard setting year” (North Carolina Department of Public Instruction, 2006b, p. 2).

\[
\text{C-score} = \frac{\text{scale score} - \text{state mean}}{\text{standard deviation}} \tag{3.7}
\]

Four-Year Growth Trajectory

The other important concept in North Carolina’s Modified ABCs Model is the “four year growth trajectory”. The first step in calculating a four-year growth trajectory is to determine the number of years the student has been in the state public schools using historic files from the state’s accountability system. If the student has been in the state public schools for four years or more and is still not proficient, the student is counted as non-proficient. If the student has been in the state public schools for less than four years, a baseline score must be determined. The baseline score refers to the score a student obtained during the first year he/she appeared in state public schools and stayed for a full academic year. Once the baseline score is located, a c-score for the baseline will be calculated. The next step is to determine which grade the student will be in four years into the future and the c-score needed in order to be counted as “proficient” in that future grade. A student is regarded as on track towards proficiency if he/she is able to reduce the

\textsuperscript{12} The hypothetical standard setting year in this study was 2003, since 2003 is the earliest year in the years between 2003 and 2005.
difference between his baseline c-score and the c-score needed to be proficient in the future grade by 25% per year (North Carolina Department of Public Instruction, 2006a). (See Table 3.7)

Table 3.7
The Amount of Improvement in Terms of Decrease in the Distance between Baseline Performance and Proficiency in the Target Grade (North Carolina Department of Public Instruction, 2006, p. 12)

<table>
<thead>
<tr>
<th>Year in State-Tested Grade</th>
<th>Decrease From Baseline Assessment In Performance Discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25% of original gap</td>
</tr>
<tr>
<td>2</td>
<td>50% of original gap</td>
</tr>
<tr>
<td>3</td>
<td>75% of original gap</td>
</tr>
<tr>
<td>4 or more</td>
<td>Students must be proficient</td>
</tr>
</tbody>
</table>

Using an example provided by the North Carolina Department of Public Instruction (2006a), a student enters North Carolina in the 3rd grade and remains in North Carolina public schools for 4th and 5th grades. The student’s first full year in the state will be the fourth grade year. Therefore, the student will need to be on trajectory to be proficient by the end of 7th grade. Assume a score of 252 is needed in order to be proficient for Reading in the seventh grade. Using Formula (3.7), we can obtain a c-score of –1.00 for the score of 252. Since the third grade test is the student’s first test in the state, it will be used as baseline. If the student obtained a score of 220 in the third grade (equivalent to -3.08 on the c-scale), the difference between the baseline c-score and 7th grade proficiency c-score is 2.08 (3.08 minus 1.00). During the four years between 4th grade and 7th grade, students will need to reduce the distance between baseline performance and proficiency in target grade by 25% each year. In this case, the student started with a c-score of –3.08 in third grade he will need a c-score of –2.56 in the fourth grade, a c-

---

13 Year 2003 is the standard setting year for Reading in North Carolina. The mean performance for Reading in the standard setting year was 261.1 with a standard deviation of 9.06. Therefore, the calculation of c-score in the example became \((252-261.1)/9.06 = -1.00\).
score of -2.04 in fifth grade, a c-score of –1.52 in the sixth grade, and a c-score of –1.00 in the 7th grade. A c-score of –1.00 is the proficiency level in 7th grade.

Figure 3.3 is a visual presentation of the above example. The dots in the top red line are the c-scores needed to be proficient in the standard setting year; the dots in the bottom blue line are the actual c-scores the non-proficient student got. The distances between the four evenly distributed horizontal lines represent the distance that the student needs to minimize each year in order to be proficient when he is in 7th grade. For the current year (2004-05), when the student is in Grade 5 and has been in the system for two full years, he needs to reach the second lowest horizontal line (i.e., reduce the difference by 50%). The student did reach the second lowest horizontal line, so he/she is on track towards proficiency. Therefore, the student will be counted as proficient, although his/her score is lower than the proficiency line.

![Figure 3.3 An Example of North Carolina’s Model](image)

(Note: 1. the dots in the top (red) line are the c-scores needed in order to be proficient in each grade in the standard setting year; 2. the dots in the lower (blue) line are the actual c-scores a non-proficient student got; 3. the distances between the horizontal (green) lines are the differences that the non-proficient student needs to minimize in order to be counted as proficient)

In this study, a five-year rather than a four-year growth trajectory was used due the differences between Louisiana’s testing program and North Carolina’s testing program. North
Carolina has the same End-of-Grade test across grades 3-8, but Louisiana implements two different tests (ITBS in grades 3, 5, 6, and 7, and LEAP/GEE in 4 and 8) in grades 3-8. Therefore, a five-year growth trajectory was chosen to avoid complications caused by the two different tests. Figure 3.4 is a flowchart of how North Carolina’s Modified Model functions.

Figure 3.4
A Flowchart of the Growth Model

Phase III Sample

The sample in Phase III analysis consisted of all students who were in Grades 3, 4, and 5 in all 297 K-5 schools in the school year of 2004-05. These students formed two cohorts. Cohort 1 consisted of students who were in Grade 5 in 2005, Grade 4 in 2004, and Grade 3 in 2003. The focus was 2005, but students’ performance in two previous years when they were in lower grades
was retrieved to examine their growth patterns. Similarly, Cohort 2 consisted of students who were in Grade 4 in 2005 and Grade 3 in 2004. The focus was still the 2005 performance, but only one year’s previous performance was retrieved because this cohort would be in Grade 2 in 2003 and no test was implemented in Grade 2. (See Table 3.8)

Table 3.8
Students Who Have More than One Data Points in 2005

<table>
<thead>
<tr>
<th>Year 2003</th>
<th>Year 2004</th>
<th>Year 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 5</td>
<td>Grade 4</td>
<td>Grade 5</td>
</tr>
<tr>
<td>Grade 3</td>
<td>Grade 4</td>
<td>Grade 3</td>
</tr>
</tbody>
</table>

(Note: Cohort 1 refers to the students who had three years of data; Cohort 2 refers to students who had two years of data)

Cohort 1 of Phase III Sample consisted of 13,411 students. Students who had missing information on Gender, Ethnicity, SES, LEP status, or SPED status, who had missing information on test scores, or who changed schools during the two years, were eliminated from analysis. Of the 13,411 students included in the Phase III Cohort 1, about 48.89% were females while 51.11% were males. African American students accounted for about 43.58% of the total student population, with White, Hispanic, Asian, and Native American students accounting for 56.42 together. About 61.75% of the students received either free or reduced-price lunch, and about 16.22% of the students were learning disabled. There were a very small percentage of students with limited English proficiency, since LEP students accounted for only 1.14% of the total student population. (See Table 3.9)

The sample of Cohort 2 consisted of 17,612 students. Students who had missing information on Gender, Ethnicity, SES, LEP status, and SPED status, who had missing information on test scores, who changed schools during the two years, were eliminated from analysis. Of the 17,612 students included in the Phase III Cohort 2, about 48.76% were females
while 51.24% were males. African American students accounted for about 48.00% of the total student population, with White, Hispanic, Asian, and Native American students accounting for 52.00 together. About 65.98% of the students received either free or reduced-price lunch, and about 19.41% of the students were learning disabled. There were a very small percentage of students with limited English proficiency, since LEP students accounted for only 1.39% of the total student population.

Since Phase III analysis used linked data (i.e., tracking the same students over time), a total number of 74,557 records were used. Specifically, Cohort 1 has 13,411 students, each of whom had 3 records (i.e., students’ test scores in 2003, 2004, and 2005 when they were in Grades 3, 4, and 5); Cohort 2 has 17,142 students, each of whom had 2 records (i.e., students’ test scores in 2004 and 2005 when they were in Grades 3 and 4). (See Table 3.9)

Phase III Data Analysis

The first step in Phase III data analysis was identifying students who were tested but not proficient. If the student had been in Louisiana state public schools for five years or more and was not proficient for the current year, the student remained non-proficient. If the student had been in the state public schools for less than five years, the student’s test score during the first year (i.e., the year he first appeared in Louisiana public schools) was used as the baseline. Then the c-score for the baseline was calculated. The second step in Phase III data analysis was determining the c-score needed for proficiency in the target grade (i.e., the grade a student will be in five years after the baseline year). The third step involved determining whether the difference between the baseline c-score and the c-score needed to be proficient in the target year had been minimized by 25% each year. If the student was able to reduce the difference by at
least 25% each year during the five years, the student was counted as on trajectory towards proficiency.

Table 3.9
Student Demographic Information for Phase III Analysis

<table>
<thead>
<tr>
<th></th>
<th>Cohort 1 (Grade 5 in 2005, Grade 4 in 2004, and Grade 3 in 2003)</th>
<th>Cohort 2 (Grade 4 in 2005, and Grade 3 in 2004)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schools (N)</td>
<td>297</td>
<td>297</td>
</tr>
<tr>
<td>Students (n)</td>
<td>13,411</td>
<td>17,162</td>
</tr>
<tr>
<td>Male</td>
<td>51.11%</td>
<td>51.24%</td>
</tr>
<tr>
<td>African American</td>
<td>43.58%</td>
<td>48.00%</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>1.14%</td>
<td>1.39%</td>
</tr>
<tr>
<td>Students with Disabilities</td>
<td>16.22%</td>
<td>19.41%</td>
</tr>
<tr>
<td>Economically Disadvantaged</td>
<td>61.75%</td>
<td>65.98%</td>
</tr>
</tbody>
</table>

For Research Question # 6, the number of students who were already proficient and who were on track towards proficiency was counted for the whole group. Then a new proficiency rate was calculated. For Research Question # 7, the number of students who were able to make proficiency and who were on track towards proficiency was counted for each of the four subgroups. Then a new proficiency rate was calculated for each subgroup.

Phase IV: Value-Added Models

Phase IV Model Specifications

A value-added model calculates how much “value” a school adds to student learning. Here “value” refers to the difference between a school’s actual growth and the expected growth, with the expected growth being based on students’ previous test scores and various background variables, such as SES, gender, and various other background variables. The value-added model in Phase IV analysis was adapted from a three-level Hierarchical Linear Model that had been used to evaluate charter schools in Massachusetts (Raudenbush & Bryk, 2002; Massachusetts
Department of Education, 2006). In this model, measurement occasions were viewed as nested within students, and students were viewed as nested within schools. Specifically, the model was specified as follows:

**Level-1 Model:**

\[ Y_{tij} = \pi_{0ij} + \pi_{1ij}a_{tij} + e_{tij} \]  

(3.8)

Where

\[ Y_{tij} = \text{the achievement outcome measure of student } i \text{ in school } j \text{ at time } t; \]

\[ \pi_{0ij} = \text{the initial status for student } i \text{ in school } j; \]

\[ \pi_{1ij} = \text{the rate of change for student } i \text{ in school } j \]

\[ a_{tij} = \text{a measure of time (2003 = -1, 2004 = 0, and 2005 = 1)} \]

\[ e_{tij} = \text{the level 1 random error term for student } i \text{ in school } j \]

**Level-2 Model:**

\[ \pi_{0ij} = \beta_{00j} + \beta_{01j} (\text{SES}) + \beta_{02j} (\text{Gender}) + \beta_{03j} (\text{Ethnicity}) \]

\[ + \beta_{04j} (\text{LEP}) + r_{0ij} \]  

(3.9)

\[ \pi_{1ij} = \beta_{10j} + \beta_{11j} (\text{Ethnicity}) + r_{1ij} \]  

(3.10)

**Level-3 Model:**

\[ \beta_{00j} = \gamma_{000} + \mu_{00j} \]  

(3.11)

\[ \beta_{01j} = \gamma_{010} \]

\[ \beta_{02j} = \gamma_{020} \]

\[ \beta_{03j} = \gamma_{030} \]

\[ \beta_{04j} = \gamma_{040} \]

\[ \beta_{10j} = \gamma_{100} + \mu_{10j}^{14} \]  

(3.12)

\[ \beta_{11j} = \gamma_{110} \]

where

\[ \gamma_{000} = \text{the overall mean (grand mean of schools) for initial status;} \]

\[ ^{14} \text{This portion is the “value” added.} \]
\( \gamma_{010} = \) the effect of SES on initial achievement status;
\( \gamma_{020} = \) the effects of gender on initial achievement status;
\( \gamma_{030} = \) the effects of ethnicity on initial achievement status;
\( \gamma_{040} = \) the effects of LEP on initial achievement status;

Similarly,
\( \gamma_{100} = \) the overall mean (grand mean of schools) for academic year achievement growth rate;
\( \gamma_{110} = \) the effect of SES on achievement growth rate;

Phase IV Sample

The sample of Phase IV study consisted of all students that were enrolled in Grade 5 in 2005 in all 297 K-5 schools in the state of Louisiana. There were a total number of 20,231 students who met these conditions. During the next stage, the previous test scores of these 20,231 students when they were in Grade 4 in 2004 and Grade 3 in 2003 were retrieved. Students who did not have test records in all three grades were deleted, which reduced the total number of students from 20,231 to 13,411. Further, students who had all three years of data but had changed schools during the three years were eliminated. Therefore, the final sample of Phase IV analysis consisted of 11,997 students. (See Table 3.10)

Table 3.10
Identifying Students for Value-Added Analysis

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Started With</td>
<td>20,231</td>
</tr>
<tr>
<td>With Three Years of Data</td>
<td>13,411</td>
</tr>
<tr>
<td>With Three Years of Data and Without Changing Schools</td>
<td>11,997</td>
</tr>
</tbody>
</table>

Of the 11,997 students included in the Phase IV study, about 51.42% were females while 48.58% were males. African American students accounted for about 41.77% of the total student population statewide, with White, Hispanic, Asian, and Native American students accounting for
54.00%, 2.43%, 1.33%, and 0.48%, respectively. About 58.87% of the students received either free or reduced-price lunch, and about 16.51% of the students were learning disabled. There were a very small percentage of students with limited English proficiency, since LEP students accounted for only 1.13% of the total student population. (See Table 3.11) Since Phase IV analysis used linked data (i.e., tracking the same students over time), a total number of 35,991 records were used. (See Table 3.11)

Table 3.11
Student Demographic Information for Phase IV Analysis

<table>
<thead>
<tr>
<th></th>
<th>3 Years of Data; Same School</th>
<th>3 Years of Data; Changed School</th>
<th>Less than 3 Years of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schools (N)</td>
<td>297</td>
<td>297</td>
<td>297</td>
</tr>
<tr>
<td>Students (n)</td>
<td>11,997</td>
<td>13,411</td>
<td>20,231</td>
</tr>
<tr>
<td>Male (1)</td>
<td>48.58%</td>
<td>48.69%</td>
<td>50.91%</td>
</tr>
<tr>
<td>Native American (1)</td>
<td>0.48%</td>
<td>0.44%</td>
<td>0.44%</td>
</tr>
<tr>
<td>Asian (2)</td>
<td>1.33%</td>
<td>1.35%</td>
<td>1.23%</td>
</tr>
<tr>
<td>African American (3)</td>
<td>41.77%</td>
<td>44.30%</td>
<td>50.36%</td>
</tr>
<tr>
<td>Hispanic (4)</td>
<td>2.43%</td>
<td>2.47%</td>
<td>2.36%</td>
</tr>
<tr>
<td>White (5)</td>
<td>54.00%</td>
<td>51.44%</td>
<td>45.62%</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>1.13%</td>
<td>1.13%</td>
<td>1.19%</td>
</tr>
<tr>
<td>Students with Disabilities</td>
<td>16.51%</td>
<td>16.30%</td>
<td>20.01%</td>
</tr>
<tr>
<td>Economically Disadvantaged</td>
<td>58.87%</td>
<td>60.10%</td>
<td>66.62%</td>
</tr>
</tbody>
</table>

Phase IV Variables

Various student background variables were included in Phase IV analysis. The following are the definitions of these variables.

- Race/Ethnicity – Students in Louisiana schools are generally categorized into five ethnic groups: Native American, Asian, African American, Hispanic, and White.
- Gender – males and females
• SES – Socio-economic Status (SES) is usually determined based on students’ lunch status: receiving free lunch, receiving reduced-price lunch, or paying for lunch. Students receiving free or reduced-price lunch are generally regarded as having low SES.

• LEP – Limited English Proficiency (LEP) students are those living in a household where a language other than English is spoken.

Phase IV Data Analysis

Phase IV data analysis began with a level-1 model for individual growth, in which within-student measurements of achievement were modeled as a function of time (Raudenbush & Bryk, 2002; Rumberger & Palardy, 2003). The level-1 model was reproduced below.

Specifically, the test score of student i in school j at time t (Y_{tij}) was modeled as a combination of initial status (\pi_{0ij}), rate of change (\pi_{1ij}), and a random error (e_{tij}). Once the measure of time, a_{tij}, was coded as -1, 0, and 1 for 2003, 2004, and 2005 respectively, the slope represented achievement gains during the three-year (i.e., grades 3-5). The level-2 model started with a fully unconditional model, which helped set the stage for further analysis. The results of the unconditional model contained information on the average math score for students entering the third grade, the average growth rate of the students during the three years, and the variance in both initial status and learning between students and schools (Rumberger & Palardy, 2003). Student level variables, such as gender, ethnicity, SES, LEP status, and SPED status were added later on. The level-3 model also started with a fully unconditional model, and then school-level variables were added later on. School level variables were aggregated student level variables. For example, a student’s SES was indicated by “0” (i.e., paid lunch) or “1” (i.e., receiving free or reduced-price lunch), while a school’s SES referred to the percentage of students receiving free or reduced-price lunch at the school. SES was also studied at school level because SES could
have an effect above and beyond what was observed at the student level (Teddlie & Reynolds, 2000).

Null Level-1 Model: \( Y_{tij} = \pi_{0ij} + \pi_{1ij}a_{ij} + e_{tij} \)  

Null Level-2 Model: \( \pi_{0ij} = \beta_{00j} + r_{0ij} \)  
\[ \pi_{1ij} = \beta_{10j} + r_{1ij} \]  

Null Level-3 Model: \( \beta_{00j} = \gamma_{000} + \mu_{00j} \)  
\[ \beta_{10j} = \gamma_{100} + \mu_{10j} \]

Once the predictors at each level were specified, a value-added index was calculated for each school. This answers Research Question # 8. Next, schools that were identified as making AYP in Phase I Study were compared to the schools that were identified as adding the most value in Phase IV Study to see whether the two sets of schools overlapped. This answers Research Questions # 9. Similarly, the schools that were identified as effective in Phase II Study were compared to the schools that were identified as adding the most value in Phase IV Study to see whether they overlapped (Research Question # 10). Finally, the schools that were identified as making AYP in Phase III Study were compared to the schools that were identified as adding the most value in Phase IV Study to see whether they overlapped (Research Question # 11). Research Question # 12 concerned an overall comparison of the four models, which is addressed in the next section.

Comparisons of the Four Models

The final section of this chapter concerns the comparison of the four models. First, it will be interesting to see if the schools that were identified as effective/ineffective by each of the four models largely overlap. If a simple model can identify the same set of effective/ineffective schools as a very complicated model, then there is no need to use a complicated model. If,
however, the schools that are identified as effective/ineffective by each model differ significantly, criteria are needed to judge the validity of the findings of each model.

Similar to what Webster, Mendro, Orsak, & Weerasinghe (1998) did, “fairness” was used as the criterion to judge the appropriateness of the four accountability models employed in this study. Each of the four accountability models produced one school effectiveness index for each school. Specifically, the school effectiveness index produced by the status model was the percentage of students scoring at or above proficiency; the school effectiveness index produced by the improvement model was the composite School Performance Score (SPS); the school effectiveness index produced by the growth model was the percentage of students that are already proficient plus those that are on track towards proficiency; and the school effectiveness index produced by the value added model was the difference between expected growth and the actual growth. In this study, it referred to the $\beta_{10j}$ value (i.e., $\gamma_{100} + \mu_{10j}$) in Equation 3.12. It was assumed that school effectiveness indices produced by more fair models should have lower correlation with individual student background variables and student background variables that are aggregated at the school level.

In order to examine the magnitude of this relationship, a Pearson-Product-Moment correlation analysis was conducted between the school effectiveness indices produced by the four models and various student background variables. The models would be judged as appropriate if the school effectiveness indices produced by them had a low correlation with individual student background variables and student background variables aggregate at the school level.
CHAPTER FOUR: RESULTS

Introduction

The purpose of this study was to compare the four generic types of accountability models (i.e., status models, improvement models, growth models, and value-added models) and see if they reach consistent or inconsistent conclusions regarding the effectiveness of K-5 schools in Louisiana. The results from the four models are presented in this chapter, which consists of five sections. Each of the first four sections concerns findings from one model. Specifically, in each of the first four sections, a school effectiveness index was calculated for every school based on the student population as a whole, and then the model’s effect on 4 other subgroups’ performances was examined. The last section (i.e., the fifth section) is an overall comparison of the results from the four models. The four models produced four school effectiveness indices for each school, and it would be interesting to see if all four indices were pointing to the same direction regarding the effectiveness of the school. Additionally, the four models were compared in terms of “fairness”. A fair accountability model should produce school effectiveness indices that have a low correlation with various student background variables (e.g., SES) (Webster, Orsak, & Weerasinghe, 1998).

Results from the Status Model (Phase I)

As discussed in Chapter Three, Phase I data analysis mainly involves calculating the percentage of students scoring at or above the “proficient”\(^{15}\) level of academic achievement for each school and compare the observed proficiency rate against the target proficiency rate of the year. The target proficiency rate for fourth grade Math for the school year of 2004-05 is 41.8%. Therefore, schools having 41.8% or more of its students scoring at or above this level are

\(^{15}\) The “proficient” level of academic achievement actually refers to “basic and above” in Louisiana.
regarded as making AYP (Bulletin 111, 2006). In other words, these schools are regarded as more effective schools.

The following is a restatement of the two research questions of Phase I study, which were presented earlier in Chapters One and Three.

1) According to the status model, how many K-5 schools in Louisiana can be categorized as making AYP for the school year of 2004-05?

2) If a K-5 school is categorized as making AYP based on the student population as a whole, will each of the following subgroups also make AYP?
   a) African American students
   b) Economically disadvantaged students
   c) Students with limited English proficiency
   d) Students with disabilities

Results for Research Question # 1

The proficiency rate of the 297 K-5 schools in Louisiana was calculated, and the observed proficiency rates were compared against the target proficiency rate of 41.8%. It was found that 247 out of 297 (83.16%) schools were able to hit the target. However, there were still a significant number of schools (n = 50) that failed to hit the target. These schools would face various types of sanctions, depending on how many times the schools failed to hit the target previously.

Louisiana has used several “common practices” (i.e., methods that have also been used by other states) to avoid identifying too many schools for corrective actions. These methods include: (1) counting the “basic” level of academic achievement as “proficient”, (2) using a back-loaded growth trajectory (i.e., requiring only modest increase of proficiency rate in the
early years of NCLB and much larger increase of proficiency rate later on), (3) calculating a 99% confidence interval around AMO, and (4) specifying the minimum subgroup reporting size (n> =10) (Goldschmidt et al., 2003; Linn, Baker, & Herman, 2005).

Methods (1) and (2) had already been applied to the data analysis of Research Question #1, the results of which were presented earlier. Method (4) didn’t apply to Research Question #1 because the sample size was always equal or greater than 10 when the student population as a whole was examined. Therefore, the only method that was applicable to Research Question #1 but had not yet been applied is Method (3), calculating a confidence interval around AMO. In order to keep consistent with Louisiana’s practices, a 99% confidence interval was calculated.

Once the 99% confidence interval was applied, it was found that 284 out of 297 (95.62%) of the schools were able to hit the target proficiency rate. Confidence interval makes it easier for schools to hit the target, especially when schools have a small student population. In Figure 4.1, for example, School “009014” had an observed proficiency rate of 27.66%, but it was regarded as making the annual measurable objective of 41.8% because the lower bound of the annual measurable objective dropped from 41.8% to 26.62% once the confidence interval was applied. In other words, only 26.62% of the students were required to reach proficiency after the confidence interval was applied. School “009014”’s proficiency rate (i.e., 27.66%) was higher than the lower bound of AMO (26.62%), so it was regarded as making the target of year.

Included in Figure 4.1 is a sample of the output for Research Question #1. Column 1, “site05”, is the 6-digit school identification number; Column 2, “total”, is the total number of 4th graders in each school. Column 3, “totalprof”, is the number of students who scored at or above the proficient level on the fourth grade LEAP Math test. Column 4, “prate”, is the proficiency rate (i.e., the quotient of dividing “totalprof” by “total”). Column 5, “AYP”, is an indicator of
whether the proficiency rate of a school was high enough to be counted as making AYP before confidence interval was applied. “AYP” takes a value of 1 if the proficiency rate of the school is equal or greater than 41.8%. Otherwise, “AYP” takes a value of 0. Column 6, “amoci”, is lower bound of the AMO confidence interval, which is the new threshold after confidence interval was applied. If a school’s proficiency rate was higher than “amoci”, it could be re-categorized as making AYP even if its actual proficiency rate was lower than 41.8%. For example, all schools in Figure 4.1 had an observed proficiency rate (i.e., “prate”) lower than the target (i.e., all values in the “prate” column were smaller than 41.8%), but they were still categorized as making AYP because after the 99% confidence interval was applied, the observed proficiency rate is greater than the lower bound of AMO.

After the confidence interval was applied, there were still 13 schools failing to make AYP. The final opportunity for these 13 schools to make AYP was through Safe Harbor. According to the provision of Safe Harbor, a subgroup can be considered as making AYP if the number of students who didn’t make AYP is reduced by 10% from last year. A 75% confidence interval is allowed during the calculation of Safe Harbor status. Once Safe Harbor was applied to the 13 schools, only two schools were categorized as failing AYP for the subgroup of all students.

Results for Research Question # 2

Under NCLB, every subgroup’s performance is emphasized. A school is regarded as making AYP only if the student population as a whole, as well as all other subgroups, reaches the AMO of the year (CTB/McGraw-Hill, 2002; National Association of School Boards of Education, 2002). Louisiana has nine subgroups, which are (1) African Americans, (2) American Indians/Alaskan Natives, (3) Asians, (4) Hispanics, (5) Whites, (6) economically
disadvantaged students, (7) students with limited English proficiency, (8) students with disabilities, and (9) student population as a whole (i.e., all students) (Bulletin 111, 2006). In Research Question # 1, the performance of the student population as a whole was examined. In Research Question # 2, the performance of (1) African Americans, (2) Economically Disadvantaged Students, (3) Students with Limited English Proficiency, and (4) Students with Disabilities were examined. These subgroups were selected to be examined because they were the ones that were most likely to fail AYP.

<table>
<thead>
<tr>
<th>site05</th>
<th>Total</th>
<th>totalprof</th>
<th>prate</th>
<th>AYP</th>
<th>amoci</th>
<th>AYP99CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>006074</td>
<td>47</td>
<td>13</td>
<td>0.2765957447</td>
<td>0</td>
<td>0.3662341902</td>
</tr>
<tr>
<td>2</td>
<td>017073</td>
<td>43</td>
<td>14</td>
<td>0.2827142567</td>
<td>0</td>
<td>0.3678806155</td>
</tr>
<tr>
<td>3</td>
<td>025025</td>
<td>62</td>
<td>18</td>
<td>0.2032353906</td>
<td>0</td>
<td>0.3639358688</td>
</tr>
<tr>
<td>4</td>
<td>040043</td>
<td>43</td>
<td>13</td>
<td>0.302359614</td>
<td>0</td>
<td>0.350891559</td>
</tr>
<tr>
<td>5</td>
<td>050013</td>
<td>46</td>
<td>15</td>
<td>0.306122444</td>
<td>0</td>
<td>0.254557756</td>
</tr>
<tr>
<td>6</td>
<td>017017</td>
<td>113</td>
<td>38</td>
<td>0.3382311858</td>
<td>0</td>
<td>0.314623483</td>
</tr>
<tr>
<td>7</td>
<td>017007</td>
<td>58</td>
<td>20</td>
<td>0.3398636363</td>
<td>0</td>
<td>0.2746561733</td>
</tr>
<tr>
<td>8</td>
<td>017044</td>
<td>94</td>
<td>32</td>
<td>0.340425319</td>
<td>0</td>
<td>0.30431677</td>
</tr>
<tr>
<td>9</td>
<td>050103</td>
<td>38</td>
<td>13</td>
<td>0.3421052632</td>
<td>0</td>
<td>0.238900344</td>
</tr>
<tr>
<td>10</td>
<td>009072</td>
<td>93</td>
<td>32</td>
<td>0.3400660215</td>
<td>0</td>
<td>0.303415679</td>
</tr>
<tr>
<td>11</td>
<td>017051</td>
<td>55</td>
<td>19</td>
<td>0.3454545454</td>
<td>0</td>
<td>0.268860141</td>
</tr>
<tr>
<td>12</td>
<td>026078</td>
<td>103</td>
<td>36</td>
<td>0.3495146531</td>
<td>0</td>
<td>0.3057195373</td>
</tr>
<tr>
<td>13</td>
<td>017093</td>
<td>85</td>
<td>30</td>
<td>0.3529411765</td>
<td>0</td>
<td>0.297943427</td>
</tr>
<tr>
<td>14</td>
<td>010061</td>
<td>82</td>
<td>29</td>
<td>0.3535353535</td>
<td>0</td>
<td>0.285132234</td>
</tr>
<tr>
<td>15</td>
<td>016019</td>
<td>203</td>
<td>72</td>
<td>0.354699003</td>
<td>0</td>
<td>0.339870299</td>
</tr>
<tr>
<td>16</td>
<td>017091</td>
<td>76</td>
<td>27</td>
<td>0.3552317593</td>
<td>0</td>
<td>0.290308968</td>
</tr>
<tr>
<td>17</td>
<td>017075</td>
<td>75</td>
<td>27</td>
<td>0.3628571428</td>
<td>0</td>
<td>0.289078843</td>
</tr>
<tr>
<td>18</td>
<td>040045</td>
<td>30</td>
<td>11</td>
<td>0.3666666667</td>
<td>0</td>
<td>0.213333333</td>
</tr>
<tr>
<td>19</td>
<td>017029</td>
<td>73</td>
<td>29</td>
<td>0.3708696076</td>
<td>0</td>
<td>0.291856928</td>
</tr>
<tr>
<td>20</td>
<td>017011</td>
<td>57</td>
<td>21</td>
<td>0.384210526</td>
<td>0</td>
<td>0.363964787</td>
</tr>
</tbody>
</table>

Figure 4.1
Before and After Applying Confidence Interval: Status Model

(Note: “site05” – six digit school identification number; “total” – the total number of 4th graders in each school; “totalprof” – the total number of 4th graders who scored at or above proficiency; “prate” – proficiency rate; “AYP” – whether the school made AYP; “amoci” – the lower bound of the AMO after confidence interval was applied; “AYP99CI” – whether schools passed AYP after the 99% confidence interval was applied)
African Americans

The subgroup of African American students was examined for 249 schools. The other 48 schools were considered as making AYP automatically because they had less than 10 African American students. Of the 249 schools examined, 165 (66.27%) of them were able to reach the target proficiency rate of 41.8% while 84 (33.73%) of them were not. After the 99% confidence intervals were applied, 233 out of 249 (93.57%) schools were able to hit the target. Due to the application of confidence interval, a school with an observed proficiency rate as low as 20% (3 out of 15 African American Students) was considered as making AYP. Until now, a total number of 16 schools were identified as failing AYP for the subgroup of African American students. The final opportunity for these schools to make AYP was through Safe Harbor. Once Safe Harbor (with a 75% confidence interval) was applied to the 16 schools, only five schools were identified as failing AYP.

Economically Disadvantaged Students

The subgroup of economically disadvantaged students was examined for all 297 schools. Of the 297 schools examined, 232 (78.11%) of them were able to reach the target proficiency rate of 41.8%. After the 99% confidence intervals were applied, 282 out of 297 (94.95%) schools were able to make AYP. Due to the application of confidence interval, a school with an observed proficiency rate of 25% (4 out of 16 students) was considered as making AYP. Further more, Louisiana sets the minimal subgroup reporting size to 10, which means that a subgroup will be considered as making AYP automatically if it has less than 10 students. Once the subgroup reporting size was applied, no more schools could be considered as making AYP because all of the 15 schools failing schools had more than 10 students in this subgroup. Until now, a total number of 15 schools were identified as not making AYP for the subgroup of economically disadvantaged students.
disadvantaged students. The final opportunity for these schools to make AYP is through Safe Harbor. Once Safe Harbor (with a 75% confidence interval) was applied to the 15 schools, only one school was considered as failing AYP.

**Students with Limited English Proficiency**

The subgroup of LEP students was examined for 9 schools. The other 288 schools were considered as making AYP automatically for this subgroup because they had either no LEP students or less than 10 LEP students. Of the 9 schools examined, 8 (88.89%) of them were able to reach the target proficiency rate of 41.8%. After the 99% confidence interval was applied, all of them (100%) were able to hit the target. Due to the application of confidence intervals, a school with an observed proficiency rate of 29.41 (5 out of 17 students) was considered as making AYP. Safe Harbor didn’t apply here because the proficiency rate is 100% already.

**Student with Disabilities**

The subgroup of special education students was examined for 235 schools. The other 62 schools were considered as making AYP automatically because they had either no or less than 10 special education students. Of the 235 schools examined, only 117 (49.79%) of them were able to reach the target proficiency rate of 41.8%. The next opportunity for the failing schools to make AYP is through confidence interval. After the 99% confidence interval was applied, 190 out of 235 (80.85%) schools were able to hit the target. Due to the application of confidence interval, a school with an observed proficiency rate of 18.89% (2 out of 11 students) was considered as making AYP. Until now, a total number of 45 schools were identified as failing AYP for the subgroup of special education students. The final opportunity for these schools to make AYP is through Safe Harbor. Once Safe Harbor was applied to these 45 schools, only 15 schools were identified as failing AYP.
Summary of the Findings from Research Questions 1 and 2

The performances of five subgroups (all students, African American students, economically disadvantaged students, LEP students, special education students) were examined in Phase I study. Of the 297 schools examined, only 20 schools were identified as failing AYP. Other schools were able to reach the target proficiency rate of 41.8% through various “common practices”. In this sense, AYP was incapable of identifying ineffective schools since all most all schools had made AYP.

Since the calculations in this study were based on Math alone, the findings in this study were compared against the results published at the official website of the Louisiana Department of Education (LDE), which were based on both ELA and Math and all subgroups. Of the 297 schools included in this study, LDE only identified 5 of them as failing AYP. In other words, the status model under NCLB played an even lighter role in reality. More schools were identified in this study probably because this study didn’t take students’ “full academic year status” into consideration. Table 4.1 illustrates the processes through which subgroups were examined.

Table 4.1
Performance of the Five Subgroups

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>Proficiency Rate</th>
<th>99% Confidence Interval</th>
<th>Safe Harbor</th>
<th>Schools Failing AYP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>with n &lt; 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Students</td>
<td>0</td>
<td>247 (83.16%)</td>
<td>284 (95.62%)</td>
<td>295 99.33%</td>
</tr>
<tr>
<td>African Americans</td>
<td>48</td>
<td>165 (71.71%)</td>
<td>281 (94.61%)</td>
<td>292 98.31%</td>
</tr>
<tr>
<td>Economically</td>
<td>0</td>
<td>232 (78.11%)</td>
<td>282 (94.95%)</td>
<td>296 99.66%</td>
</tr>
<tr>
<td>Disadvantaged</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEP Students</td>
<td>288</td>
<td>296 (99.66%)</td>
<td>297 (100%)</td>
<td>297 (100%)</td>
</tr>
<tr>
<td>SPED Students</td>
<td>62</td>
<td>179 (60.27%)</td>
<td>252 (84.85%)</td>
<td>282 (94.95%)</td>
</tr>
</tbody>
</table>

16 In AYP calculation, most states, including Louisiana, only include students that have been in the school for a full academic year.
Results from the Improvement Model (Phase II)

As discussed in Chapter Three, Louisiana’s current school and district accountability model was selected as the representative of improvement models. It was a rather sophisticated model in which an assessment index was calculated all grades that were administered the Norm-Referenced Test (NRT) (i.e., grades 3 and 5 in this case) and all grades that were administered the Criterion-Referenced Test (CRT) (i.e. grade 4 in this case), and then the NRT and CRT indices were weighted and combined so that a single school performance score (SPS) can be calculated for each school. The SPS score ranges from 0.0 to 120.0 and beyond. A score of 120.0 or beyond symbolizes that the school has reached Louisiana’s 12-year goal of academic achievement, while a score less than 60 is academically unacceptable (Bulletin 111, 2006; Public Affairs Research Council of Louisiana, 2004).

The following is a restatement of the three research questions of the Phase II study, which were presented earlier in Chapters One and Three.

3) According to the improvement model, how many K-5 schools in Louisiana can be categorized as making AYP for the school year of 2004-05?

4) If a K-5 school is categorized as making AYP based on the student population as a whole, will each of the following subgroups also make AYP?
   
   e) African American students
   
   f) Economically disadvantaged students
   
   g) Students with limited English proficiency
   
   h) Students with disabilities

5) According to the Improvement Model, how many K-5 schools in Louisiana have met or exceeded the expected improvement?
Results from Research Question #3

The SPS of the 297 K-5 schools in Louisiana were calculated. The observed SPS ranged from 28.02 to 133.83. A number of schools (n = 15) had a SPS of 120 or higher, which meant that these schools had already reached Louisiana’s academic goal of 2013-14. (See Figure 4.2) When compared against the minimum requirement of 60, it was found that 227 out of 297 (76.43%) schools were able to reach or exceed the minimum requirement. However, there were still a significant number of schools (n = 70) that had a SPS lower than 60. These schools would face various types of sanctions, depending on how many times they failed to hit the target previously.

<table>
<thead>
<tr>
<th>site05</th>
<th>spedtotal</th>
<th>leptotal</th>
<th>total</th>
<th>SPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>037010</td>
<td>17</td>
<td>0</td>
<td>101</td>
</tr>
<tr>
<td>2</td>
<td>017096</td>
<td>18</td>
<td>7</td>
<td>105</td>
</tr>
<tr>
<td>3</td>
<td>028048</td>
<td>58</td>
<td>2</td>
<td>374</td>
</tr>
<tr>
<td>4</td>
<td>052006</td>
<td>96</td>
<td>1</td>
<td>350</td>
</tr>
<tr>
<td>5</td>
<td>009060</td>
<td>58</td>
<td>1</td>
<td>208</td>
</tr>
<tr>
<td>6</td>
<td>037011</td>
<td>49</td>
<td>0</td>
<td>236</td>
</tr>
<tr>
<td>7</td>
<td>037007</td>
<td>58</td>
<td>1</td>
<td>243</td>
</tr>
<tr>
<td>8</td>
<td>037047</td>
<td>50</td>
<td>0</td>
<td>245</td>
</tr>
<tr>
<td>9</td>
<td>022037</td>
<td>37</td>
<td>0</td>
<td>272</td>
</tr>
<tr>
<td>10</td>
<td>010046</td>
<td>67</td>
<td>0</td>
<td>291</td>
</tr>
<tr>
<td>11</td>
<td>017082</td>
<td>26</td>
<td>1</td>
<td>249</td>
</tr>
<tr>
<td>12</td>
<td>017094</td>
<td>67</td>
<td>1</td>
<td>190</td>
</tr>
<tr>
<td>13</td>
<td>037004</td>
<td>58</td>
<td>0</td>
<td>296</td>
</tr>
<tr>
<td>14</td>
<td>010019</td>
<td>37</td>
<td>0</td>
<td>206</td>
</tr>
<tr>
<td>15</td>
<td>008033</td>
<td>48</td>
<td>1</td>
<td>328</td>
</tr>
</tbody>
</table>

Figure 4.2
Schools That Have Reached Louisiana’s Academic Goal of 2014

(Note: “schooled” – the six-digit school identification number; “spedtotal” – total number of special education students; “leptotal” – total number of LEP students; “total” – total number of students; “SPS” – the school performance score)

Schools in Louisiana are assigned six types of performance labels based on their SPS. The “five stars” label is assigned to schools with a SPS of 140 or above; the “four stars” label is assigned to schools with a SPS of 120 or above; the “three star” label is assigned to schools with
a SPS of 100 or above; the “two stars” label is assigned to schools with SPS of 80 or above; and the “one star” label was assigned to schools with a SPS of 60 or above. Schools with a SPS below 60 are assigned the label of “academically unacceptable”. Table 4.2 includes detailed information regarding the performance labels of the 297 K-5 schools in Louisiana.

Table 4.2
Performance Labels Based on SPS

<table>
<thead>
<tr>
<th>Performance Label</th>
<th>SPS</th>
<th>Number of Schools</th>
<th>Comparison to LDE Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four Stars</td>
<td>120.0 – 139.9</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Three Stars</td>
<td>100.0 – 119.9</td>
<td>54</td>
<td>66</td>
</tr>
<tr>
<td>Two Stars</td>
<td>80.0 – 99.9</td>
<td>80</td>
<td>106</td>
</tr>
<tr>
<td>One Star</td>
<td>60.0 – 79.9</td>
<td>78</td>
<td>80</td>
</tr>
<tr>
<td>Academically Unacceptable</td>
<td>Below 60</td>
<td>70</td>
<td>33</td>
</tr>
</tbody>
</table>

It is worthwhile to point out that the performance labels assigned to the 297 K-5 schools in this study were somewhat different from the actual performance labels they got from LDE. The actual performance labels were based on four subject areas (i.e., English Language Arts, Math, Science, and Social Studies) and student attendance data while the performance labels in this study were based on Math only. In order to check the validity of the findings based on Math alone, a Pearson-Product-Moment correlation analysis was conducted between the SPS calculated in this study and those published by LDE and it was found the SPS based on Math alone had a very high correlation ($r=0.96$) with the SPS based on four subjects.

Results from Research Question # 4

Louisiana’s current accountability model was originally designed to examine the performance of “all students” (i.e., the student population as a whole). Since NCLB (2001) emphasizes subgroup performance, this model has been used to calculate a SPS for each subgroup in each school. Although there is no absolute target (e.g., $SPS > 60$), all subgroups are
required to improve 2 points. However, how well the subgroups perform doesn’t influence the performance label of schools, and it only influences schools’ growth label a little bit (Bulletin 111, 2006). For example, in order for schools to be labeled as “exemplary academic growth”, all subgroups need to grow at least 2 points, in additional to other conditions. If all subgroups don’t grow at least 2 points, the schools will be labeled as “recognized academic growth”, if all other conditions stay the same. The other growth labels (i.e., minimal growth, no growth, and school in decline) are not influenced by subgroup performance.

African American Students

The subgroup of African American students was examined for 295 schools. The other 2 schools didn’t have African American students for all three years (i.e., 2003, 2004, and 2005), so their growth labels were not influenced by the performance of this subgroup. Of the 295 schools examined, African America students improved 5.65 points on average. However, there was huge variation in terms of how much schools improve (or decline) (SD = 11.88). When compared against the target growth of 2 points, 174 (58.98%) of the 295 schools had improved at least 2 points from the weighted average of two previous years (i.e., the school years of 2002-03 and 2003-04).

Economically Disadvantaged Students

The subgroup of economically disadvantaged students was examined for 297 schools. Of the 297 schools examined, low SES students (i.e., students who received free or reduced-price lunch) improved 5.08 points on average, but there was huge variation in terms of how much schools improve (or decline) (SD = 9.83). When compared against the target growth of 2 points, 183 (61.62%) of them had improved at least 2 points from the average of two previous years (i.e., the school years of 2002-03 and 2003-04).
Students with Limited English Proficiency

The subgroup of LEP students was examined for 130 schools. The other 167 schools didn’t have LEP students for all three years (i.e., the school years of 2002-03, 2003-04, and 2004-05), so their growth labels were not influenced by the performance of this subgroup. Of the 130 schools examined, LEP students improved 3.91 points on average, but there was huge variation in terms of how much schools improve (or decline) (SD = 18.89). When compared against the growth target of 2 points, only 19 of them (14.62%) improved more than two points.

Students with Disabilities

The subgroup of special education students was examined for 297 schools. This subgroup improved 3.77 points on average, but there was huge variation in terms of how much schools improve (or decline) (SD = 18.84). When compared against the growth target of 2 points, 154 (51.85%) of them had improved at least 2 points from the average of two previous years (i.e., 2003 and 2004).

Summary of the Four Subgroups

Growth in SPS was analyzed for four subgroups. All four subgroups have to grow at least 2 points in order for a school to be considered (together with other conditions) as “exemplary academic growth”. Of the 297 schools examined, 207 of them failed to grow 2 points in either one or more subgroups. Therefore, these schools can not be labeled as “exemplary growth” even if all other conditions are met. Figure 4.3 is a sample of 20 such schools. Columns that are entitled “diffblk”, “diffses”, “difflep”, and “diffsped” are the SPS growth (or decline) for the subgroups of African American students, economically disadvantaged students, LEP students, and SPED students, respectively. As shown in Figure 4.3, most of the schools had only one subgroup failed to meet the growth target of two points.
Results from Research Question # 5

In Research Question # 3, a SPS was calculated for each school and schools were assigned 6 types of performance labels based on some absolute standards (i.e., a school with a SPS of 140 or above was labeled “five stars” school). In Research Question # 4, A SPS was calculated for each subgroup in each school and the subgroups were also compared against an absolute standard (i.e., 2 points minimum growth). In this research question, however, the growth target was set based on Louisiana’s 2014 goal of all schools reaching a SPS of 120. Specifically, the (weighted) average school performance indices in the school years of 2002-03 and 2003-04 was used as baseline, and schools were required to grow enough in 2005 so that
they would be on track towards a SPS of 120 in 2014. Schools that started low were required to grow at a faster pace and therefore they would have higher growth target.

Growth in SPS was examined for all 297 K-5 schools. Of the 297 schools examined, 65 (21.89%) of them showed “exemplary academic growth”, 75 (25.25%) of them showed “recognized academic growth”, 62 (20.88%) of them showed “minimum growth”, 23 (7.74%) of them showed “no growth”, and 61 (20.54%) of them were in decline. Table 4.3 showed detailed information regarding the growth labels of all 297 schools.

Similar with the performance labels, the growth labels calculated in this study were somewhat different from the actual performance labels. The actual growth labels were based on four subject areas (i.e., English Language Arts, Math, Science, and Social Studies) and student attendance data while the growth labels in this study were based on Math only. In order to check the validity of the findings based on Math alone, a correlation analysis was conducted and it was found the SPS based on Math alone had a high correlation ($r=0.85$) with the SPS based on four subjects.

Table 4.3
Growth Labels of the Schools

<table>
<thead>
<tr>
<th>Growth Label</th>
<th>Number of Schools</th>
<th>Comparison of LDE Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exemplary Academic Growth</td>
<td>65</td>
<td>128</td>
</tr>
<tr>
<td>Recognized Academic Growth</td>
<td>75</td>
<td>61</td>
</tr>
<tr>
<td>Minimum Academic Growth</td>
<td>62</td>
<td>46</td>
</tr>
<tr>
<td>No Growth</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>In Decline</td>
<td>61</td>
<td>22</td>
</tr>
<tr>
<td>No Label$^{17}$</td>
<td>11</td>
<td>15</td>
</tr>
</tbody>
</table>

$^{17}$ Schools with a SPS of 105 or above were assigned “no label” if they were in decline.
Results from the Growth Model (Phase III)

As discussed in Chapter Three, North Carolina’s Modified ABCs Model was selected as the representative of growth models. In this model, students who scored below the proficient level of academic achievement are identified and an individual growth trajectory is depicted. Based on the individual growth trajectory, students who are not proficient for the current year but are projected to be proficient four years into the future are re-categorized as proficient for the current year. The purposes are (1) “to encourage schools to put individual students who have yet to reach proficiency on accelerated paths to meeting state achievement standards” and (2) “encourage schools to identify and provide appropriate interventions to students who are at risk of falling below proficiency” (Tennessee Department of Education, 2005, p. 1) The following is a restatement of the research questions for Phase III study, which were presented earlier in Chapters One and Three.

6) According to the growth model, how many K-5 schools in Louisiana can be categorized as making AYP?

7) If a K-5 school in Louisiana is categorized as making AYP based on the population as a whole, will each of the following subgroups also make AYP?
   i) African American students
   j) Economically disadvantaged students
   k) Students with limited English proficiency
   l) Students with disabilities

Results from Research Question # 6

A total number of 5967 fourth graders who scored below the proficient level of academic achievement were identified and their individual five-year growth trajectories were calculated. It
was found that 1376 out of 5967 (23.06%) students could be re-categorized as proficient because they are on track towards proficiency five years into the future. Figure 4.4 is a sample of the output for Research Question # 6. Columns 1 and 2 are the school and student identification number, respectively. Column 3, which is entitled “C2004”, is the c-score of the student when he or she was in Grade 3 in 2004 (i.e., baseline). Column 4, which is entitled “C2005”, is the c-score of the student when he or she was in Grade 4 in 2005 (i.e., the current status). Column 5, which is entitled “C2009”, is the c-score needed for a student to be counted as proficient in Grade 8. Since a five-year growth trajectory was used, this cohort of students (i.e., fourth graders in 2005) would be in Grade 8 in 2009. Column 6, which is entitled “grow”, is the growth in c-score when the students moved from the third grade to the fourth grade. Column 7, which is entitled “growneed”, is the amount of c-score growth that is need each year in order for the students to be counted as “on track towards proficiency five years into the future”. If the actual growth (i.e., the column entitled “growth”) is bigger than the growth needed (i.e., the column entitled “growneed”), a code of “1” is assigned (i.e., the column entitled “code”). Otherwise a code of “0” is assigned. Students who got a code of “1” were re-categorized as being proficient.

Similarly with what was done with fourth graders, a total number of 7608 students who scored below the 55th percentile on ITBS Math were identified and individual five-year growth trajectories were calculated. It was found that 740 out of 7608 (9.73%) students could be re-categorized as proficient because they were on track towards proficiency five years into the future.

In growth modeling students who were not proficient but were on track towards proficiency were re-categorized as being proficient. Therefore, the proficiency rate of schools generally increased with the application of growth models. In Research Question # 1, 50 schools
had a proficiency rate lower than 41.8% before confidence interval and Safe Harbor were applied. Once students who were on track towards proficiency were added to the total number of proficient students, 34 schools had a proficiency rate lower than the target proficient rate.

Figure 4.4
Students Who Are Projected To Be Proficient

(Note: “schooled” - the school identification number; “studentid”- student identification number; “C2004” - the c-score of the student when he or she was in Grade 3 in 2004; “C2005” - c-score of the student when he or she was in Grade 4 in 2005 (i.e., the current status). “C2009”, is the c-score needed for a student to be counted as proficient in Grade 8 in 2009; “grow”, is the growth in c-score when the students moved from the third grade to the fourth grade; “growneed”, is the amount of c-score growth that is need each year in order for the students to be counted as on track towards proficiency five years into the future.)

Results from Research Question #7

The effect of growth modeling on subgroup performance was examined in Research Question # 7. The purpose was to see whether all four sub groups benefit equally from the application of the growth model.

African American Students

In the fourth grade, 4206 African American students who scored below proficiency were identified and their individual growth trajectories were depicted. It was found that 978 (23.25%)
of them were eligible for re-categorization because they were on track towards proficiency five years into the future. Similarly, 4343 students were identified in the fifth grade, and 383 (8.82%) of them were re-categorized as proficient.

Economically Disadvantaged Students

In the fourth grade, 4964 economically disadvantaged students who scored below proficiency were identified and their individual five-year growth trajectories were depicted. It was found that 1146 (23.09%) of them were eligible for re-categorization as proficient because they were on track towards proficiency five years into the future. Similarly, 5544 economically disadvantaged students who scored below proficiency were identified in the fifth grade, and 533 (9.61%) of them were re-categorized as proficient.

Students with Limited English Proficiency

In the fourth grade, 80 LEP students who scored below proficiency were identified and their individual five-year growth trajectories were depicted. It was found that 20 (25.00%) of them were eligible for re-categorization because they were on track towards proficiency five years into the future. Similarly, 100 LEP students who scored below proficiency were identified in the fifth grade, and 20 (20.00%) of them were re-categorized as proficient because they were on track towards proficiency five-years into the future.

Students with Disabilities

In the fourth grade, 1641 special education students who scored below proficiency were identified and their individual growth trajectories were depicted. It was found that 362 (22.06%) of them were eligible for re-categorization because they were on track towards proficiency five years into the future. Similarly, 1118 students were identified in the fifth grade, and 102 (9.02%) of them were re-categorized as proficient.
Summary of the Four Subgroups

The growth model had largely unbiased effects on subgroup performance. As shown in Table 4.4, there was approximately 23% increase of proficient students for the fourth grade across all four subgroups, and there was approximately 10% increase of proficient students for the fifth grade across all four subgroups. None of the subgroups, except for the LEP students, benefited more from the growth model than others, although a higher percentage of students were eligible for re-categorization in grade four than in grade five.

Table 4.4
The Effects of the Growth Model on Subgroup Performance

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>African American Students</th>
<th>Economically Disadvantaged Students</th>
<th>Students with Limited English Proficiency</th>
<th>Students with Disabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourth Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(# of students who were re-categorized)</td>
<td>978 out of 4206 (23.25%)</td>
<td>1146 out of 4964 (23.09%)</td>
<td>20 out of 80 (25.00%)</td>
<td>362 out of 1641 (22.06%)</td>
</tr>
<tr>
<td>Fifth Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(# of students who were re-categorized)</td>
<td>383 out of 4343 (8.82%)</td>
<td>533 out of 5544 (9.61%)</td>
<td>20 out of 100 (20.00%)</td>
<td>102 out of 1118 (9.02)</td>
</tr>
</tbody>
</table>

Results from the Value-Added Model (Phase IV)

In Phase IV of this study, school effectiveness was modeled as the amount of “value” that a school added to student learning. The three-level hierarchical linear modeling technique was used, in which individual student growth trajectories were represented by the level-one model, variations in growth parameters among children were represented by the level-two model, and the variations among schools were represented by the level-three model (Bryk & Raudenbush, 2002).
The following is a restatement of the research questions for Phase IV study, which were presented earlier in Chapters One and Three.

8) According to the value-added model, which schools add the most “value” to student learning?

9) Do schools that are categorized as making AYP by the status model add more “value” to student learning than those categorized as not making AYP?

10) Do schools that are categorized as making AYP by the improvement model add more “value” to student learning than those categorized as not making AYP?

11) Do schools that are categorized as making AYP by the growth model add more “value” to student learning than those categorized as not making AYP?

12) Do the four models generated consistent or inconsistent conclusions regarding the effectiveness of K-5 schools in Louisiana?

Results from Research Question # 8

Hierarchical linear modeling is a step-by-step process. It usually starts with what is called the “unconditional model”, in which no predictor variables are incorporated, and then moves progressively to various types of “conditional models” (Bryk and Raudenbush, 2002). The unconditional model of this study was presented in Formula 4.1 – 4.4. At level-1, individual student achievement (ZMATH) was modeled as a function of average standardized score for the student ($\pi_0$) across three years, the effect of year the student took the test ($\pi_1$), and a random within-student error ($e$). At level-2, $\pi_0$ was then modeled as a function of the average standardized score for the school ($\beta_{00}$) and the random variation within the school ($r_0$); and $\pi_1$ was modeled as a function of average growth in standardized score for the school ($\beta_{10}$) and random variation in growth in standardized score within the school ($r_1$). Finally, at level-3, $\beta_{00}$ was
modeled as the average standardized score across all schools statewide ($\gamma_{000}$) and the random variation in school initial status ($\mu_{00}$) between schools; and $\beta_{10}$ was modeled as the average growth in standardized score across all schools statewide ($\gamma_{000}$) and the random variation in school growth between schools ($\mu_{10}$).

Level 1 Model:
$$ZMATH = \pi_0 + \pi_1 \text{ (YEAR)} + e$$

(4.1)

Level 2 Model:
$$\pi_0 = \beta_{00} + r_0$$
$$\pi_1 = \beta_{10} + r_1$$

(4.2)

Level 3 Model:
$$\beta_{00} = \gamma_{000} + \mu_{00}$$
$$\beta_{10} = \gamma_{100} + \mu_{10}$$

(4.3)

Mixed Model:
$$ZMATH = \gamma_{000} + \gamma_{100} \times \text{YEARCD} + \mu_{00} + \mu_{10} \times \text{YEARCD} + e$$

(4.4)

Results from the Unconditional Model

The results of the unconditional model are presented in Table 4.5. The estimated mean intercept, $\gamma_{000}$, and mean growth rate, $\gamma_{100}$, were $-0.074585$ and $-0.000019$, respectively. As all test scores were converted to Z-scores before the analysis, it was not quite meaningful to interpret $\gamma_{100}$ and $\gamma_{100}$. This is a limitation imposed by the test instruments that were used to collect student performance data. Since the tests were not vertically scaled, Z scores were used. This made it difficult to interpret $\gamma_{100}$ and $\gamma_{100}$. The estimates for the variance of individual growth parameters $\pi_0$ and $\pi_1$ were $0.62744$ and $0.00057$, respectively. This tells us that students vary significantly in terms of their initial math achievement ($\chi^2 = 44898.79, P < .001$), and there was also significant variation in their achievement growth rates ($\chi^2 = 10653.83, P < .001$). Similarly, there was significant difference in school mean status ($\chi^2 = 2367.85, P < 0.001$) and school mean growth rate ($\chi^2 = 1061.24, P < .001$).
Table 4.5
Three-Level Unconditional Model

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>T-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average initial status, $\gamma_{00}$</td>
<td>-0.074585</td>
<td>0.022774</td>
<td>-3.275*</td>
</tr>
<tr>
<td>Average yearly growth rate, $\gamma_{100}$</td>
<td>-0.000019</td>
<td>0.006565</td>
<td>-0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Variance</th>
<th>Df</th>
<th>$\chi^2$</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-student variation, $e$</td>
<td>0.24193</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Students within schools)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual initial status, $r_{0ij}$</td>
<td>0.62744</td>
<td>10949</td>
<td>44898.79</td>
<td>0.00</td>
</tr>
<tr>
<td>Level 2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Students within schools)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual growth rate, $r_{1ij}$</td>
<td>0.00057</td>
<td>10949</td>
<td>10653.83</td>
<td>0.00</td>
</tr>
<tr>
<td>Level 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Between schools)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School mean status, $\mu_{00j}$</td>
<td>0.12939</td>
<td>298</td>
<td>2367.85</td>
<td>0.00</td>
</tr>
<tr>
<td>Level 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Between schools)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School mean growth rate, $\mu_{10j}$</td>
<td>0.00878</td>
<td>298</td>
<td>1061.24</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Results from the Conditional Model

The conditional model was specified by adding various student/school background variables to the unconditional model. Many exploratory analyses were needed before the conditional model can be finalized. In this study, it was found that factors such as GENDER, ETHNICITY, SES, and LEP STATUS were significant predictors of initial student status (i.e., $\pi_0$) while ETHNICITY was the only significant predictor for student achievement growth (i.e., $\pi_1$). Therefore, the final conditional model was specified as Formula 4.5 – 4.7.

Level 1 Model: $ZMATH = \pi_0 + \pi_1 \cdot (YEAR) + e$  

(4.1)

Level 2 Model: $\pi_0 = \beta_{00} + \beta_{01} \cdot (GENDER) + \beta_{02} \cdot (ETHNIC) + \beta_{03} \cdot (SES) + \beta_{04} \cdot (LEPFLG) + r_0$

$\pi_1 = \beta_{10} + \beta_{11} \cdot (ETHNIC) + r_1$  

(4.5)
Level 3 Model: \( \beta_{00} = \gamma_{000} + \mu_{00} \)
\[
\beta_{01} = \gamma_{010} \\
\beta_{02} = \gamma_{020} \\
\beta_{03} = \gamma_{030} \\
\beta_{04} = \gamma_{040} \\
\beta_{10} = \gamma_{100} + \mu_{10} \\
\beta_{11} = \gamma_{110} \tag{4.6}
\]

Mixed Model: \( ZMATH = \gamma_{000} + \gamma_{010} \times GENDER + \gamma_{020} \times ETHNIC \\
+ \gamma_{030} \times SES + \gamma_{040} \times LEPFLG + \gamma_{100} \times YEAR \\
+ \gamma_{110} \times ETHNIC\times YEAR + \tau_0 + \tau_1 \times YEAR \\
+ \mu_{00} + \mu_{10} \times YEAR + e \tag{4.7} \)

Results from the Conditional Model

The results of the conditional model are shown in Table 4.6. From the top panel we can see that GENDER, ETHNIC, SES, and LEP STATUS were significant predictors of student initial achievement (\( \pi_0 \)). The associated coefficients are 0.085962, -0.515583, -0.306860, and -0.311688, respectively. Possible inferences are: (1) on average the Z-score on math of a male student was 0.085962 higher than a female, (2) on average the Z-score on math of an African American student was 0.515583 lower than that of a student of other ethnical backgrounds, (3) on average the Z-score on math of low SES student was 0.30680 lower than a student of middle or high SES, and (4) on average the Z-score on math of a LEP student was 0.311688 lower than a student whose native language was English. The predictor ETHNICITY had an effect not only on student initial status but also student achievement growth (\( \pi_1 \)). The associated coefficient for ETHNICITY was –0.040707, which means that on average African American students grew 0.040707 slower than students of other ethical backgrounds on a yearly basis. The middle panel
of Table 4.6 provides useful information on various random effects. The majority of the variance came from differences in initial student status (65.35%) and within-student error. Although between school variation in initial status and in achievement growth accounted for only a small portion of the total variance, but their effects were significant (i.e., both p-values were smaller than .001). The bottom panel concerns reliability. The reliability for $\pi_0$, $\pi_1$, and $\beta_{00}$ were fine, but the reliability for $\beta_{10}$ was very low. According to Bryk and Raudenbush (2002), low reliability on certain coefficient didn’t discredit the whole HLM model.

Table 4.6
Three-Level Conditional Model

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Symbol</th>
<th>DF</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand Mean</td>
<td>$\gamma_{000}$</td>
<td>298</td>
<td>0.351292*</td>
<td>0.025223</td>
</tr>
<tr>
<td>Gender</td>
<td>$\gamma_{010}$</td>
<td>11243</td>
<td>0.085962*</td>
<td>0.015734</td>
</tr>
<tr>
<td>Ethnic</td>
<td>$\gamma_{020}$</td>
<td>11243</td>
<td>-0.515583*</td>
<td>0.021397</td>
</tr>
<tr>
<td>SES</td>
<td>$\gamma_{030}$</td>
<td>11243</td>
<td>-0.306860*</td>
<td>0.019929</td>
</tr>
<tr>
<td>LEP Status</td>
<td>$\gamma_{040}$</td>
<td>11243</td>
<td>-0.311688*</td>
<td>0.072036</td>
</tr>
<tr>
<td>Year (Slope)</td>
<td>$\gamma_{100}$</td>
<td>298</td>
<td>0.020767*</td>
<td>0.007098</td>
</tr>
<tr>
<td>Ethnic</td>
<td>$\gamma_{110}$</td>
<td>11246</td>
<td>-0.040707*</td>
<td>0.007499</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Symbol</th>
<th>DF</th>
<th>Variance</th>
<th>% Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Status</td>
<td>$\mu_{00}$</td>
<td>298</td>
<td>0.04862*</td>
<td>5.62%</td>
</tr>
<tr>
<td>School Slope</td>
<td>$\mu_{10}$</td>
<td>298</td>
<td>0.00885*</td>
<td>1.02%</td>
</tr>
<tr>
<td>Student Status</td>
<td>$r_0$</td>
<td>10945</td>
<td>0.56522*</td>
<td>65.35%</td>
</tr>
<tr>
<td>Student Year (Slope)</td>
<td>$r_1$</td>
<td>10948</td>
<td>0.00053</td>
<td>0.0006%</td>
</tr>
<tr>
<td>Within Student</td>
<td>E</td>
<td>10948</td>
<td>0.24167</td>
<td>27.94%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Symbol</th>
<th>Reliability Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>$\pi_0$</td>
<td>0.875</td>
</tr>
<tr>
<td>Year (Slope)</td>
<td>$\pi_1$</td>
<td>0.004</td>
</tr>
<tr>
<td>School</td>
<td>$\beta_{00}$</td>
<td>0.689</td>
</tr>
<tr>
<td>School Year (Slope)</td>
<td>$\beta_{10}$</td>
<td>0.683</td>
</tr>
</tbody>
</table>
School Effectiveness Indices

The purpose of running the HLM model was to obtain a value-added school effectiveness index for every school. In this study, the “value” added by each school was quantitatively defined as $\gamma_{10} + \mu_1$ (i.e., $\beta_{10}$) (Massachusetts Department of Education, 2006). The software HLM produced this value for every school in the level-3 residual file. Figure 4.5 is a sample of schools with the highest $\beta_{10}$ values. These schools were viewed as adding the most value to student learning based on the value-added model.

![Figure 4.5](Image)

The “Value” Added by Schools

Results from Research Questions 9, 10, 11, and 12

The four models produced four indicators of school effectiveness for every school.

Research Questions 9 – 12 concerns the extent to which these four indicators point to the same direction regarding the effectiveness of the same schools. A Pearson-Product-Moment
correlation analysis was conducted, the results of which were shown in Table 4.7. M1, M2, M3, and M4 stands for the school effectiveness indicators produced by (1) the status model, (2) the improvement model, (3) the growth model, and (4) the value-added model, respectively. The school effectiveness indicators produced by M4 had a very low and insignificant correlation with those produced by M1 and M2, and had a significant but low correlation with those produced by M3.

Table 4.7
Comparing the Four Models

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>0.994*</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>0.994*</td>
<td>0.899*</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>0.111</td>
<td>0.114</td>
<td>0.153*</td>
<td>1.000</td>
</tr>
</tbody>
</table>

An Overall Comparison of the Four Models

Since the four models reached different conclusions regarding the effectiveness of schools, additional criteria are needed to judge the credibility of the models. “Fairness” was chosen as the additional criterion here (Webster et al., 1998), and a good model should produce school effectiveness indicators that had low correlations with student background variables. In other words, the school effectiveness indicators should not simply be a reflection of various student/school background variables.

In order to test the “fairness” of the models, the four school effectiveness indicators were correlated with aggregated student background variables such as gender, ethnicity, SES, LEP status, and SPED status. Table 4.8 presents the Pearson-Product-Moment correlation results. It
was found that background variables such as GENDER, LEP STATUS, and SPED STATUS had a low correlation with all four indicators. Therefore, they were inadequate in differentiating the four models and therefore were eliminated from further analysis. The focus was then on SES and ETHNICITY. As we can see, the school effectiveness indicators produced by the status model, improvement model, and growth model had a high correlation with Ethnicity ($r = -0.66871, -0.74376, -0.62437$, respectively), which means that these school effectiveness indicators were biased against schools that serve African American students. Similarly, the school effectiveness indicators produced by the three models also had significant correlations with SES ($r = -0.17735, -0.21683, -0.18792$, respectively), which means that these school effectiveness indicators were biased against schools serving low SES students.

On the other hand, the school effectiveness indicators produced by value-added model had both low and insignificant correlations with ethnicity ($r = 0.08200, p = 0.1573 > 0.05$) and SES ($r = 0.00431, p = 0.9408 > 0.05$), which means that it produces the most fair school effectiveness indicators.

Table 4.8
Using “Fairness” as the Additional Criterion

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Ethnicity</th>
<th>SES</th>
<th>LEP</th>
<th>SPED</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>-0.06298</td>
<td>-0.66871</td>
<td>-0.17735</td>
<td>-0.03424</td>
<td>-0.07758</td>
</tr>
<tr>
<td>M2</td>
<td>-0.07910</td>
<td>-0.74376</td>
<td>-0.21683</td>
<td>-0.04830</td>
<td>-0.08182</td>
</tr>
<tr>
<td>M3</td>
<td>-0.00995</td>
<td>-0.62437</td>
<td>-0.18792</td>
<td>-0.05672</td>
<td>-0.03164</td>
</tr>
<tr>
<td>M4</td>
<td>0.04172</td>
<td>0.08200</td>
<td>0.00431</td>
<td>-0.01849</td>
<td>0.08171</td>
</tr>
</tbody>
</table>
CHAPTER FIVE: CONCLUSION

This study was conducted within the context of NCLB, which sets the ambitious (some say unattainable) goal of 100% proficiency for all students by the school year of 2013-14. Educational institutions are held accountable for not making adequate yearly progress (AYP) towards the goal, with AYP being defined by the simplistic status model (CTB/McGraw-Hill, 2002; National Association of School Boards of Education, 2002; Public Affairs Research Council of Louisiana, 2004). NCLB poses serious problems for many states. In Louisiana, for example, about 75% of all public schools will eventually be identified as failing AYP by the status model (PARCL, 2004). Therefore, it becomes imperative to explore the possibility of other models through which school performance can be measured.

The four models examined in this study were a status model, an improvement model, a growth model, and a value-added model. One of the purposes was to see what alternative conclusions can be drawn on the performance of the same set of schools when different accountability models are applied. The other purpose was to compare the four models in terms of “fairness”, with a fair model being defined as one that produces school effectiveness indices that have lower correlations with various student background variables (Webster et al., 1998).

Although this study was limited by the type of assessments/data available, it was the first to examine all four generic types of accountability models based on data collected from Louisiana schools and students. This study could shed some light on a potential alternative to the current accountability system in Louisiana. This chapter is divided into two sections. In Section One, major findings from each of the four models are discussed. Section two concerns the methodological and policy implications of these findings.
Major Findings from the Four Models

Major Findings from the Status Model

The status model under NCLB is an overly simplistic accountability model that defines school effectiveness as the percentage of students scoring at or above the proficient level of academic achievement. Coupled with the unattainable goal of 100% proficiency by 2014, the status model will eventually identify the majority of American schools as failing. The problem with identifying an unmanageable number of schools for school improvement will be less severe for Louisiana in the beginning years of NCLB, since a relatively low percentage of students (e.g., 30.1% for Math and 36.9% for English Language Arts in 2002) are required to reach proficiency. However, in some states where the assessment standards are high (e.g., California), thousands of schools were identified for school improvement at the very beginning. As a response to this problem, the U. S. Department of Education has granted states some flexibilities, which include (1) calculating confidence intervals around AMO, (2) counting “basic” level of achievement as “proficient”, (3) using a back-loaded trajectory (i.e., requiring only modest increases of proficiency rate in early years and much larger increase of proficiency rate later on), and (4) specifying the minimum subgroup reporting size ($n > 10$) (CCSSO, 2003; Linn, Baker, & Herman, 2005).

In this study, the status model was applied to all 297 K-5 schools in Louisiana. For the subgroup of all students, 247 out of the 297 schools (83.16%) were able to reach the target proficiency rate of 41.8%; for the subgroup of African American students, 165 out of 297 schools (66.27%) were able to reach the target; for the subgroup of economically disadvantaged students, 232 out of 297 schools (78.11%) were able to reach the target; for the subgroup of students with limited English proficiency, 296 out of 297 schools (99.67%) were able to reach
the target; and for the subgroup of special education students, 179 out of 297 schools (60.27%) were able to hit the target. According to NCLB provisions, a school is considered as failing AYP if any one of the subgroups fails to make AYP. Therefore, more than half of the 297 schools would be identified as failing AYP before the NCLB flexibility procedures are applied. If we took all 1376 public schools in Louisiana into consideration, hundreds of schools would have been identified as failing AYP before the NCLB flexibility procedures are applied. The condition can be much worse in the years closer to 2014 because the target proficiency rate becomes higher and higher every two years. For example, the target proficiency rate for math was 41.8% in 2005, and it will become 53.5% in 2008, 65.2% in 2011, 76.1% in 2012, 88.6% in 2013, and ultimately 100% in 2014.

After the NCLB flexibility procedures were applied, all but 20 schools were categorized as making AYP. Therefore, one of the major effects of NCLB flexibility procedures on Louisiana schools is that AYP becomes incapable of identifying less effective schools. This also creates a dilemma: without the flexibility procedures about half of the schools would have been identified as failing AYP; with the flexibility procedures only twenty schools were identified as failing AYP. NCLB flexibility procedures will not work wonders as the time gets closer to the end year of 2014. At that time, the state will go back to identify an unmanageable number of schools as failing.

Due the controversies generated by NCLB, many people expect that fundamental changes will happen to NCLB. As a matter of fact, 25 states across the nation still keep their own statewide accountability system while trying to accommodate NCLB requirements (EdSource, 2005). A common characteristic among these states is that all of them had already adopted a much more sophisticated accountability model before NCLB imposed the overly simplistic status.
model on top of them. In these states, educational institutions are subjected to a dual accountability system, but for some states (e.g., Louisiana) what really holds the schools accountable is the statewide accountability system rather than NCLB.

Major Findings from the Improvement Model

Louisiana’s current school and district accountability system is a rather sophisticated system that aligns many of content standards, assessments, and Louisiana’s 10-year and 20-year goals of academic achievement. The accountability model is based on simple statistical concepts, but it can get complicated in actual calculations. One complication is caused by the fact that different tests were used in different grades. Students in Grades 3, 5, 6, 7, and 9 were given the NRT tests (i.e., ITBS tests), and students in 4, 8, 10, and 11 were given the CRT tests (i.e., LEAP tests). Louisiana used to weight the two components\(^{18}\) to come up with single school performance score until 2006, when a new test called ILEAP was introduced at all grades.

Another relevant complication came from differences in grade configurations. In order for both the CRT and NRT components to be present, a school would need to have at least one grade taking the CRT tests and one grade taking the NRT tests. In other words, a school would need to have least one grade in grades 3, 5, 6, 7, and 9 and at least one grade in 4, 8, 10, and 11. Schools that didn’t have this type of grade configuration would need to “borrow” students from other schools\(^{19}\). In addition to all these issues, different weightings were used for different subjects and grades, bonus points were given to students who took the tests for the second time in summer schools, and students with varying disabilities were given LAA1 (Louisiana Alternative Assessment 1) and LAA2 (Louisiana Alternative Assessment 2), all of which add to

\(^{18}\) CRT was weighted 60%, NRT was weighted 30%, attendance was weighted 10% for K-6 and 5% for 7-12, and dropout was weighted 5% for 7-12.

\(^{19}\) Interested readers can refer to a LDE publication called Bulletin 111 and information on the LDE website for more information.
the complexity of the calculation (Bulletin 111, 2006). Despite of all these complications, Louisiana’s model is understood by the majority of school administrators with proper training provided by LDE.

In this study, Louisiana’s current accountability model was applied to all 297 K-5 schools in Louisiana. Each school was given a pair of labels: performance label and the growth label. There are six types of performance labels. Of the 297 schools examined, 15 schools (5.05%) got the “four stars” label, 54 schools (18.18%) got the “three stars” label, 80 schools (26.94%) got the “two stars” label, 78 schools (26.26%) got the “one star” label, and 70 schools (23.57%) got the label of “academically unacceptable”. Schools that were labeled as “academically unacceptable would face various types of corrective actions, depending on how many years they were labeled as “academically unacceptable” and how much they improved. There are also six types of growth labels. Of the 297 schools studied, 65 schools (21.88%) got the label of “exemplary academic growth”, 75 schools (25.25) got the label of “recognized academic growth”, 62 schools (20.88%) got the label of “minimum academic growth”, 23 schools (7.74%) got the label of “no growth”, 61 schools (20.54%) got the label of “in decline”, and 11 schools (3.70%) got the label of “no label” because they had a SPS of 105 or higher but were in decline.

It is worthwhile to emphasize again that the performance labels and growth labels assigned in this study were somewhat different from the actual labels assigned by the LDE because only math scores were used. In order to check the validity of the findings in this study, a correlation analysis was conducted between the SPS obtained in this study and those produced by LDE, and it was found that the correlation was very high (0.96). Additionally, a correlation analysis between the Growth SPS obtained in this study and those produced by the LDE was also conducted, and the correlation between the two was also moderately high (0.85).
Louisiana’s current accountability model was also used to examine subgroup performance. Each subgroup was required to improve 2 points. However, how well the subgroups perform doesn’t influence the performance label of schools, and it only influences schools’ growth label very little. Specifically, in order for a school to be labeled as “exemplary academic growth”, all subgroups need to grow at least 2 points, in addition to all other conditions. If all subgroup doesn’t growth at least 2 points, the schools will be labeled as “recognized academic growth”, if all other conditions stay the same. Of the 297 schools examined, 90 of them grew at least 2 points all four subgroups (African Americans, LEP students, SPED students, and economically disadvantaged students).

Louisiana’s model has many advantages. First, it uses test scores from all four subject areas, so schools can not just focus their attention on reading/English language arts and mathematics only. Second, it gives bonus points to students who scored below “basic” in previous tests but improved a level in later tests. This gives school incentives to help students falling behind. Thirdly, the model is understood by most school administrators.

However, Louisiana’s model also has all the disadvantages associated with using cross-sectional data. For example, the accountability results are contaminated by student mobility. Each year schools will have a slightly (and in some cases largely) different student population than the year before. Therefore, improvement or decline in school performance scores may be because of changes in student population rather than what the teachers and school administrators do. Another associated disadvantage is that individual student-level data are aggregated to the school-level, so a lot of detailed and useful information on individual students is lost. Thirdly, the accountability results are also contaminated by student background variables. For example, the
correlation between the school performance score generated by the Louisiana model and the percentage of African American students was -0.74.

Major Findings from the Growth Model

Growth models have been proposed as a supplement to the status model (i.e., the AYP model) under NCLB. The growth model proposed by North Carolina is intended for students scoring below proficiency only. It depicts an individual growth trajectory for each student. If the student is projected to be proficient four years into the future, the student is re-categorized as proficient for the current year even if his/her test score is lower than the cut point for proficiency.

In this study, 5968 fourth graders scoring below proficiency were identified, and an individual growth trajectory was depicted for each of them. It was found that 1376 of them (23.06%) could be re-categorized as proficient because they were on track towards proficiency five years into the future. This reduced the number of schools failing AYP from 50 to 34. In other words, 16 more schools could be relabeled as AYP schools due to the application of the growth model. It is worthwhile to point out that a five-year rather than a four-year growth trajectory was used because Louisiana doesn’t implement the same test on all grades like North Carolina does. Fourth graders take the CRT test in 2005, and they will take the CRT test again five years later when they are in eighth grade in 2009. If a four-year growth trajectory were used, the students would take the NRT test instead because they would be in seventh grade in 2008.

Growth models have many advantages. First, they provide incentives for educational institutions to put students who have yet to reach proficiency on accelerated path towards meeting state academic standards (Tennessee Department of Education, 2006). Second, they are based on a stronger theoretical perspective. In growth modeling, individual student performance is monitored over time, so we are no longer comparing apples to oranges. Third, one of the main
purposes of growth modeling under NCLB is to identify more schools as AYP schools. As students who are not proficient but projected to be proficient are re-categorized as proficient for the current year, schools generally see an increase in proficiency rate. This, of course, eases the burden of making the target proficiency rate.

Growth models also have some disadvantages. For example, growth models require at least two data points for each student. Sometimes this simple condition is hard to satisfy in schools and districts where student mobility rate is high. Additionally, growth models require a vertically scaled test to be implemented across all grade levels so that student performances can be compared over time. Louisiana didn’t have such a test until 2006 when ILEAP replaced ITBS as the statewide test. Since the data used in this study were from 2003-2005, z-scores were computed to facilitate cross-tests comparison. Thirdly, incorporating growth model as part of the NCLB accountability system will not solve the problem of identifying an unmanageable number of schools of failing schools. In this study, 16 out of 297 schools could be re-categorized as AYP schools. That represents a meaningful increase, but it won’t help much in easing the NCLB burden. As a matter of fact, nothing could help without changing the unattainable goal of 100% proficiency by 2014.

Major Findings from the Value-Added Model

Value-added models are a very special class of growth models. In value-added modeling, an empirical estimation of a school’s effectiveness is obtained by decomposing the variance in student test scores into portions that are explained by student inputs, and into other portions that are presumably caused by the school. Schools are held accountable only for what they have control of. Value-added models inevitably set lower standards for disadvantaged students. This contradicts NCLB provisions that are intended to hold all students to the same standards.
Therefore, value-added models have not yet approved by USDE for NCLB accountability purposes.

The value-added model employed in this study was a three-level hierarchical model in which measurement occasions were viewed as nested within students, who in turn were viewed as nested within schools. The “value” added by each school was quantitatively defined as $\gamma_{10} + \mu_{10}$ (i.e., $\beta_{10}$) (Massachusetts Department of Education, 2006). It indicates the difference between a school’s actual growth and expected growth. The HLM software program produced this value for every school in the level-3 residual file.

In order to compare the value-added model to the other three models, a Pearson-Product-Moment correlation analysis was conducted between school effectiveness indices produced by (1) the status model (i.e., proficiency rate), (2) the improvement model (i.e., the SPS score), (3) the growth model (i.e., the proficiency rate after adding students who were not proficient but were projected to be proficient five years into the future), and (4) the value-added model (i.e., $\beta_{10}$ values). It was found that the school effectiveness indicators produced by the value-added model had a very low and insignificant correlation with those produced by the status model and the improvement model, and had a significant but low correlation with those produced by the growth model.

Since the four models reached different conclusions regarding the effectiveness of the same set of schools, additional criteria were needed to judge the credibility of the models. “Fairness” was chosen as the additional criterion here (Webster et al., 1998), such that a good model should produce school effectiveness indicators that have low correlations with various student background variables. In other words, the school effectiveness indicators should not simply be a reflection of various student/school background variables.
In order to test the “fairness” of the models, the four school effectiveness indicators were correlated (i.e., Pearson-Product-Moment correction) with aggregated student background variables such as gender, ethnicity, SES, LEP status, and SPED status. It was found that background variables such as GENDER, LEP STATUS, and SPED STATUS had a low correlation with all four indicators. Therefore, they were inadequate in differentiating the four models and therefore were eliminated from further analysis.

The focus was then on SES and ETHNICITY. As detailed in Chapter 4, the school effectiveness indicators produced by the status model, improvement model, and growth model had a high correlation with Ethnicity ($r = -0.67, -0.74, -0.62$, respectively), which meant that these school effectiveness indicators were biased against schools serving African American students. Similarly, the school effectiveness indicators produced by the three models also had significant correlations with SES ($r = -0.18, -0.22, -0.19$, respectively), which means that these school effectiveness indicators were biased against schools serving low SES students. On the other hand, the school effectiveness indicators produced by value-added model had both low and insignificant correlations with ethnicity ($r = 0.08200, p = 0.1573 > 0.05$) and SES ($r = 0.00431, p = 0.9408 > 0.05$), which means that it produces the most fair school effectiveness indicators.

Methodological and Policy Implications

Methodological Implications

The findings of this study have important methodological implications. First, different accountability models define school effectiveness in different ways and therefore are likely to hold a different set of schools accountable. As demonstrated in this study, school effectiveness can be defined as the current status of student performance (e.g., the status model), an improvement from one status to another (e.g., the improvement model), growth in student
learning (e.g., the growth model), and the amount of “value” that a school adds to student learning (e.g., the value-added model). When the four types of accountability models were tested on all K-5 schools in Louisiana, it was found that the school effectiveness indices produced by the status model, the improvement model, and the growth model diverged significantly from those produced by the value-added model but converged highly among themselves. Second, growth models and value-added models require longitudinal data that are harder to obtain. Ideally, student test scores across multiple years can be easily tracked through students’ social security numbers. In reality, however, various types of problems can occur that can lead to a low matching rate. The effects of low matching rate are hard to estimate. Third, growth models and value-added models require a vertically-scaled measurement instrument. The data used in this study didn’t come from a vertically-scaled measurement instrument, so test scores were transformed into z-scores. This made it difficult to interpret some of the results.

Policy Implications

The findings of this study have important policy implications. First, the status model (i.e., the AYP model) imposed by NCLB doesn’t work very well on the schools in Louisiana. It is driven by the unrealistic goal of 100% by 2014, which could eventually identify the majority of the schools in Louisiana as failing. Second, the NCLB flexibilities on the AYP model can not solve the problem (i.e., the problem of identifying an unmanageable amount of schools as failing) in the long run. At present, a large number of schools are saved by the NCLB flexibilities, but it becomes more like a number game. Third, USDE called for proposals for growth models. However, growth models, as narrowly defined in this study within the context of NCLB, do not help much in reducing the number of schools identified as failing. Fourth, schools in Louisiana, like schools in many other states, are subject to a dual accountability system. The
dilemma is that NCLB imposes an accountability model that is impossible to follow in reality. Therefore, Louisiana, like many other states, wants to keep her own accountability system while trying to accommodate NCLB requirements. It is hoped that fundamental changes will be made to NCLB.
REFERENCES


Goldschmidt, P., Roschewski, P. Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams,


Louisiana Department of Education. (2006c). Louisiana School, District, and State Accountability System, Bulletin 111,


Tolbert, L. (2003). The Impact of Louisiana’s School and District Accountability System on Students’ Performance on the State Mandated Criterion Referenced Test. Unpublished Dissertation from the Department of Educational Leadership, Research and Counseling, Louisiana State University.


## APPENDIX I

### STATUS, GROWTH, AND VALUE-ADDED MODELS

<table>
<thead>
<tr>
<th>General considerations by accountability model</th>
<th>Status models</th>
<th>Growth models</th>
<th>Value added models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Currently approved by ED for NCLB</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Underlying purpose</td>
<td>Rank/rate schools based on current performance</td>
<td>Rank/rate schools based on performance change</td>
<td>Rank/rate schools based on performance changes different from expected</td>
</tr>
<tr>
<td>Major issues for consideration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results aligned with AYP</td>
<td>Very likely</td>
<td>Less likely</td>
<td>Less likely</td>
</tr>
<tr>
<td>Ratings generally understood</td>
<td>Very likely</td>
<td>Likely</td>
<td>Less likely</td>
</tr>
<tr>
<td>Inferences same as AYP</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Requires more than 1 year of data</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Unique student ID required</td>
<td>No</td>
<td>Generally</td>
<td>Generally</td>
</tr>
<tr>
<td>Potentially confounds student &amp; school effects</td>
<td>Yes</td>
<td>Less likely</td>
<td>Less likely</td>
</tr>
<tr>
<td>Implementation time</td>
<td>Quick</td>
<td>Moderate/varies</td>
<td>High/varies</td>
</tr>
<tr>
<td>Implementation process</td>
<td>Simple</td>
<td>Moderate/varies</td>
<td>High/varies</td>
</tr>
<tr>
<td>Optimal testing requirements</td>
<td>None</td>
<td>Annual/same content</td>
<td>Annual/same content</td>
</tr>
<tr>
<td>Estimate teacher effects</td>
<td>No</td>
<td>Possible</td>
<td>Possible</td>
</tr>
<tr>
<td>Possible to measure within school inequities in performance</td>
<td>Limited</td>
<td>Possible</td>
<td>Possible</td>
</tr>
<tr>
<td>Costs</td>
<td>Low</td>
<td>Moderate</td>
<td>Potentially high</td>
</tr>
<tr>
<td>Simultaneously suitable for program evaluation</td>
<td>Unlikely</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Measures change for individual students</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Absolute</td>
<td>-</td>
<td>Possible</td>
<td>Possible</td>
</tr>
<tr>
<td>-Relative to standard</td>
<td>-</td>
<td>Possible</td>
<td>Possible</td>
</tr>
<tr>
<td>-Requires equal interval scale</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Requires vertically equated scale score</td>
<td>No</td>
<td>Varies</td>
<td>Varies</td>
</tr>
<tr>
<td>Successful school profile</td>
<td>High average achievement, or exceeds % proficient target</td>
<td>High average achievement growth given average student enrollment</td>
<td>Higher than expected achievement growth given average student enrollment</td>
</tr>
<tr>
<td>Intended consequences</td>
<td>Reward high performing school</td>
<td>Rewards growth</td>
<td>Rewards better than expected growth</td>
</tr>
<tr>
<td>Unintended consequences</td>
<td>-Fosters status quo -Ignores within school inequities -Rewards schools with &quot;favorable&quot; enrollment -Does not reward student achievement growth (school improvement) -Reduce incentives for high quality teachers to teach</td>
<td>-May ignore high achieving schools -May ignore within school inequities -Perceived different standards for different sub-groups</td>
<td>-May ignore high achieving schools -May ignore within school inequities -Perceived different standards for different sub-groups</td>
</tr>
</tbody>
</table>

(Source: Goldschmidt et al., 2005, p. 13)
### APPENDIX II

**CHARACTERISTICS OF VALUE-ADDED MODELS**

<table>
<thead>
<tr>
<th>Characteristic of Model</th>
<th>Type of Value-Added Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TVAAS Model</td>
</tr>
<tr>
<td>Underlying purpose</td>
<td>Implicit accounts for initial status</td>
</tr>
<tr>
<td>Current locations</td>
<td>Tennessee</td>
</tr>
<tr>
<td><strong>Major tradeoffs</strong></td>
<td></td>
</tr>
<tr>
<td>Results aligned with AYP</td>
<td>Less likely</td>
</tr>
<tr>
<td>Ratings generally understood</td>
<td>Less likely</td>
</tr>
<tr>
<td>Inference same as AYP?</td>
<td>No</td>
</tr>
<tr>
<td>Potentially confounds student &amp; school effects</td>
<td>Less likely</td>
</tr>
<tr>
<td>Implementation time</td>
<td>High/Varies</td>
</tr>
<tr>
<td>Estimate teacher effects</td>
<td>Yes</td>
</tr>
<tr>
<td>Teacher effect models in use</td>
<td>Yes</td>
</tr>
<tr>
<td>Possible to measure within school inequalities</td>
<td>Limited</td>
</tr>
<tr>
<td>Costs</td>
<td>High</td>
</tr>
<tr>
<td>Simultaneously suitable for program evaluation</td>
<td>Possible, not practical</td>
</tr>
<tr>
<td>Successful student profile</td>
<td>High output trends compared to input trends</td>
</tr>
<tr>
<td>Individual value added estimates inappropriate</td>
<td>Higher than expected gain compared to norms</td>
</tr>
<tr>
<td>Successful school profile</td>
<td>Higher than expected average gain compared to norms</td>
</tr>
<tr>
<td>Type of growth examined</td>
<td>Layered gains</td>
</tr>
<tr>
<td>Methodological advantages</td>
<td>Multiple years initial status implicit</td>
</tr>
<tr>
<td>Methodological disadvantages</td>
<td>Extreme complex-convergence</td>
</tr>
<tr>
<td>Software</td>
<td>Proprietary</td>
</tr>
</tbody>
</table>

(Source: Goldschmidt et al., 2005, p. 21)
APPENDIX III
A COMPARISON OF THE SAMPLE USED IN EACH OF THE FOUR MODELS

<table>
<thead>
<tr>
<th></th>
<th>Status Model</th>
<th>Improvement Model</th>
<th>Growth Model</th>
<th>Value-Added Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort 1</td>
<td>Cohort 2</td>
<td>Total # of Records</td>
<td>Total # of Students</td>
</tr>
<tr>
<td>Schools (N)</td>
<td>297</td>
<td>297</td>
<td>13,411*3</td>
<td>298</td>
</tr>
<tr>
<td>Students (n)</td>
<td>24,159</td>
<td>63,279</td>
<td>17,612</td>
<td>11,997</td>
</tr>
<tr>
<td>Male (1)</td>
<td>52.39%</td>
<td>51.70%</td>
<td>51.24%</td>
<td>48.58%</td>
</tr>
<tr>
<td>Native American (1)</td>
<td>0.35%</td>
<td>0.41%</td>
<td>0.43%</td>
<td>0.48%</td>
</tr>
<tr>
<td>Asian (2)</td>
<td>1.26%</td>
<td>1.34%</td>
<td>1.15%</td>
<td>1.33%</td>
</tr>
<tr>
<td>African American (3)</td>
<td>53.67%</td>
<td>50.26%</td>
<td>48.00%</td>
<td>41.77%</td>
</tr>
<tr>
<td>Hispanic (4)</td>
<td>2.17%</td>
<td>2.48%</td>
<td>2.16%</td>
<td>2.43%</td>
</tr>
<tr>
<td>White (5)</td>
<td>42.55%</td>
<td>45.51%</td>
<td>48.25%</td>
<td>54.00%</td>
</tr>
<tr>
<td>Limited English Proficiency</td>
<td>1.34%</td>
<td>1.47%</td>
<td>1.04%</td>
<td>1.13%</td>
</tr>
<tr>
<td>Students with Disabilities</td>
<td>21.24%</td>
<td>19.65%</td>
<td>19.77%</td>
<td>16.51%</td>
</tr>
<tr>
<td>Economically Disadvantaged</td>
<td>70.42%</td>
<td>67.91%</td>
<td>67.56%</td>
<td>58.87%</td>
</tr>
</tbody>
</table>
VITA

Fen Yu was born in Wuhan, China, on May 26, 1978. After earning a bachelor of arts in English education from Wuhan University in 2000, she came to the United States and entered the graduate program in the Department of Curriculum and Instruction at Kent State University where she earned the master of education degree in secondary education in 2003. She is now a doctoral candidate in the Department of Educational Theory, Policy, and Practice at Louisiana State University. She is expected to receive the Doctor of Philosophy degree in educational leadership and research in August 2007.