

Running head: ACCOUNTABILITY MODELS

Identifying Effective and Ineffective Schools for Accountability Purposes:

A Comparison of Four Generic Types of Accountability Models

By

Fen Yu, Eugene Kennedy, Charles Teddlie  
Louisiana State University

and

Mindy Crain  
Louisiana Department of Education

### 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 was defined as one that produces school effectiveness indices that have low correlation with various student background variables (Webster, Mendro, Orsak, & Weerasinghe, 1998). The sample of this study consisted of all 299 K-5 schools in Louisiana. The results indicated 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.

## Introduction

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<sup>1</sup> 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 below the proficiency level are reduced by 10% from

---

<sup>1</sup> 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).

the previous year, regardless of the 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 (Goldschmidt, Roschewski, Choi, Auty, Hebbler, Blank et al., 2005). 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). For example, some improvement models compare the same grades in two different years (i.e., 3<sup>rd</sup> graders in 2005 vs. 3<sup>rd</sup> graders in 2006, 4<sup>th</sup> graders in 2005 vs. 4<sup>th</sup> graders in 2006). It is like comparing apples to oranges because the 3<sup>rd</sup> graders of one year can be quite different from the 3<sup>rd</sup> 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, 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).

The purpose of this study is 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 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 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 authors of this study was chosen as the representative of value-added models. The AYP model was selected because it is the primary indicator of school performance under NCLB. Louisiana's current accountability model was selected 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.

It should be noted, however, that it was not quite possible to follow strictly what these four models do. What we did was really trying to follow the underlying concept upon which these models are built. The places where we had to make a detour are discussed along the rest of this paper.

## Literature Review

### Criticisms on NCLB

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, Goldschmidt, & Yamashiro, 2006; Choi, Seltzer, Herman, Yamashiro, 2004; Doran & Izumi, 2004; Meyer, 1995; Webster & Mendro, 1998). First, AYP based on simple measures provides misleading and invalid results (Doran & Izumi, 2004). Since the 1960s, school effectiveness researchers in U.S. have consistently demonstrated that the majority of the variance in student achievement between schools are attributable to the socioeconomic status (SES) of the student body, which is usually measured by percentage of students receiving free or reduced price lunch (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).

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 (Ryan, 2004).

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).

## 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.

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 can be 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 method is very much similar to the regression models used in school effects research (Teddlie, Reynolds, & Sammons, 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. In response to the Secretary's call for growth models, many states (about 20) 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. Up until November 2006, the U.S. Department of Education (USDE) had approved growth models proposed by Tennessee, North Carolina, Delaware, Arkansas, and Florida (U.S. Department of Education, 2006).

Louisiana Department of Education 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). Louisiana adopted her 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 (Councils of Chief State School Officers, 2003). Therefore, Louisiana’s current plan is to keep the extant accountability system while trying to accommodate NCLB requirements (Linn, Baker, & Herman, 2005). This creates a dual accountability system that could potentially put the majority of schools in school improvement/corrective actions in the long run.

At present, Louisiana has managed to keep most of its schools out of corrective actions through various “common practices”, 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 ( $n > 10$ ) (CCSSO, 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 as it gets close to the deadline for *all* students to reach proficiency approaches (i.e., 100% proficiency rate by 2013-14). Using mathematics as an example, 36.9% of the students (by school and subgroup) were required to make 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). As we can see, much lower growth rates are required in earlier years while much higher growth in proficiency rates are required in the later years. It is obvious that more and more schools will be identified as needing school improvement (Bulletin 111, 2006; Linn, Baker, & Herman, 2005).

#### Four Generic Types of Accountability Models

There are many different ways to categorize the numerous accountability models that have been used and proposed. In this study, the typology by Goldschmidt et al. (2005) was employed, in which accountability models were assigned into four categories: status models, improvement models, growth models, and value-added models. The characteristics of each of the four general types of the accountability models are discussed in detail below.

##### *Status Models*

A status model takes a snapshot of the performance of a subgroup/school/LEA at one point in time, and compares the observed 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). For example, the percentage of 4<sup>th</sup> graders making proficiency in 2000 (i.e., status in 2000) can be compared to the percentage of 4<sup>th</sup> 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 subgroup fails to make AYP in a particular year but the students who scored below proficiency were reduced by 10% from the previous year's comparable group (Goldschmidt et al, 2005; Riddle, 2005).

#### *Growth Model*

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 4<sup>th</sup> 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. The expected growth can be either data-driven or policy-driven (Riddle, 2005).

### *Value-added Models*

Value-added models are a very special class of growth models (Lissitz, Doran, Schafer, and Willhoft, 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 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). As Lissitz et al. (2006) remarked:

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 (Lissitz et al., 2006, p. 8).

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, 1994). 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 et al, 2005, p. 16)

### Comparisons of Different School Accountability Models

Different types of accountability models are based on different theoretical frameworks. When the stakes associated with accountability results are high, it is crucial to examine the validity of these 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 (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.

#### Research Questions

- 1) According to the Status Model, how many K-5 schools in Louisiana will be categorized as making AYP for the school year of 2004 - 05?
- 2) According to the Improvement Model, how many K-5 schools in Louisiana will be categorized as effective for the school year of 2004 – 05?
- 3) According to the Growth Model, how many K-5 schools in Louisiana are projected to make AYP for the school year of 2004 - 05?

- 4) According the value-added model, which schools add the most “value” to student learning?
- 5) Do the four models generate consistent/inconsistent conclusions regarding the effectiveness of K-5 schools in Louisiana?
- 6) Which model provides the most fair school effectiveness indices?

### Methodology

The preceding sections introduced four generic types of accountability models: status models, improvement models, growth models, and value-added models. In this section, the methodology components are discussed separately for each model in the order of (1) model specification, (2) sample, (3) measurement, and (4) data analysis. We chose to discuss the research components separately for each model because the four types of models are so different in terms of theoretical frameworks and complexity that it is unlikely that they are going to share similar samples, data analysis methods, etc. Therefore, we thought that it might be clearer to present them one by one.

#### Phase I: Model Specification for the Status Model

The 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 for Math for the school year of 2004-05 in the state of Louisiana (i.e., the independent variable for this study) is for all subgroups/schools/LEAs to reach a proficiency rate of 41.8% (Bulletin 111, 2006).

#### Phase I: The Sample of the Status Model

The status model sample consisted of all 24,241 students in Grade 4 for the school year of 2005-06 in all 299 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.<sup>2</sup> Only the fourth graders were included in the Phase I study because Grade 4 was the only grade that had 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 24,241 students included in the Phase I study, about 47.63% were females while 52.37% were males. African American students accounted for about 53.69% of the total student population statewide, with White, Hispanic, Asian, and Native American students accounting for 42.52%, 2.17%, 1.27%, and 0.35%, respectively. Students were not evenly distributed in terms of race/ethnicity, since some schools (7 schools) had 0% African American students while others (31 schools) had 100% African American

---

<sup>2</sup> One of the four models, value-added models, uses student background variables (i.e., gender, ethnicity) as predictor variables. Therefore, students with missing information on these variables were eliminated from analysis. In order to keep practice consistence with the rest three models, students with missing background information were also eliminated for the other three models.

students. Students were not evenly distributed in terms of SES, either. On average, 72.56% 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 (8 schools) had more than 40% special education students. The average percentage of special education students was 21.80%.

#### Phase I: The Measurements of the Status Model

LEAP 21 Math was used as the measurement instrument. The Louisiana Educational Assessment Program for the 21<sup>st</sup> Century (LEAP 21) is a criterion-referenced testing (CRT) program that was designed to measure how well a student has mastered state content standards in subject areas of English Language Arts (ELA), Mathematics, Science, and Social Studies. The LEAP 21 is administered at grades 4 and 8, and it has five achievement ratings: Advanced, Mastery, Basic, Approaching Basic, and Unsatisfactory (Louisiana Department of Education, 2005a). For test scores used in this study (i.e., fourth grade math), a score between 419 and 500 is within the Advanced level; a score between 370 and 418 is within the Proficiency (or Mastery) level; a score between 315 and 369 is within the Basic level; a score between 282 and 314 is within the

Approaching Basic level; and a score between 100 and 281 is within the Unsatisfactory level (Louisiana Department of Education, 2005).

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 for the Status Model

Phase I data analysis was relatively simple, since it didn’t involve complicated computations. Math scores equal to or greater than 315 will be assigned a code of 1 while math scores less than 315 will be assigned a score of 0. Then the percentage of students coded as proficient was counted, which was compared to the expected level of proficiency rate (41.8%). Schools that had more than 41.8% of students reaching proficiency were labeled as making AYP and vice versa.

#### Phase II: Model Specifications of the Improvement Model

Louisiana’s current school 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

Louisiana's 2014 goal (Bulletin 111, 2006; Public Affairs Research Council of Louisiana, 2004).

The SPS score is determined using a weighted composite index derived from three indicators of student achievement: (1) criterion referenced tests (CRT) for grade 4, (2) norm-referenced tests (NRT) for grades 3 and 5, and (3) student attendance rate. The CRT scores count for 60%, the NRT scores count for 30%, and attendance rates count for the remaining 10% (The Accountability Technical Advisory Committee, 1998; Public Affairs Research Council of Louisiana, 1999; Bulletin 111, 2006). In this study, only the CRT (for 4<sup>th</sup> graders) and NRT (for 3<sup>rd</sup> and 5<sup>th</sup> graders) scores were used to calculate the SPS for all 299 K-5 schools because the other three models (i.e., status models, growth models, and value-added models) only use test scores as indicators of school performance. Therefore, attendance was eliminated from analysis, and the weights of CRT and NRT scores were boosted to 66.7% and 33.3%, respectively.

The CRT and NRT indicators come from grade-level indices, the calculation of which is shown in Formula (1) – (3) (Accountability Technical Advisory Committee, 1998)<sup>3</sup>. In this study, every student in Grades 3 and 5 were given a grade-level performance index as calculated by Formula (1) or (2). Then an average performance index of all students in Grades 3 and 5 was obtained to form the NRT indicator. Similarly, every student in Grade 4 was given a grade level performance index as calculated by Formula (3). Then an average performance index of all students in Grade 4 was obtained to form the CRT indicator.

---

<sup>3</sup> 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.

$$\text{Grade 3 Index} = (4.167 * SS) - 679.2 \quad (1)$$

$$\text{Grade 5 Index} = (2.941 * SS) - 544.1 \quad (2)$$

$$\text{Grade 4 Index} = (N_{\text{advanced}} * 200 + N_{\text{proficient}} * 150 + N_{\text{basic}} * 100 + N_{\text{approaching\_basic}} * 50 + N_{\text{unsatisfactory}} * 0) / N_{\text{total}} \quad (3)$$

Where

SS = scale scores on the ITBS for 4<sup>th</sup> graders or scale scores on the LEAP for 3<sup>rd</sup> and 5<sup>th</sup> graders;

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

$N_{\text{proficient}}$  = the number of students scoring at the proficient level

$N_{\text{basic}}$  = the number of students scoring at the basic level

$N_{\text{approaching\_basic}}$  = the number of students scoring at the approaching basic level

$N_{\text{unsatisfactory}}$  = the number of students scoring at the unsatisfactory level

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 2005, schools with SPS lower than 60 were labeled as “academically unacceptable”; schools with SPS between 60 and the state average were labeled as “academically below the state average”; schools with SPS between the state average and 99.9 were labeled as “academically above the state average”; schools with SPS between 100 and 124.9 were labeled as “school of academic achievement”; schools with SPS between 125 and 149.9 were labeled as “school of academic distinction”; and finally schools with SPS 150 or above were labeled as “school of academic excellence” (Reeves, 2003).

School Performance Score is also used to assign a “growth” label for each school.

For 2005, the average SPS in 2003 and 2004 was used as baseline, which was compared to the SPS of 2005 to get the actual growth. The actual growth is then compared to the target growth to see whether schools have met or exceeded the growth target. The calculation of the growth target is detailed in Formula (4) below (Louisiana Register, 2002):

$$\begin{aligned} &\text{Growth Target} \\ &= [\text{PropRE} * (120 - \text{SPS}) / N] + [\text{PropSE} * (120 - \text{SPS}) / 2N] \\ &+ \text{PropLEP} * (120 - \text{SPS}) / N \end{aligned} \quad (4)$$

Where

PropRE = the proportion of regular education students in the school

SPS = School Performance Score

N = the number of remaining years in the 12-year period

PropSE = the proportion of special education students

PropLEP = the proportion of students with limited English proficiency

Again, the “growth” here was actually what has been defined as “improvement” by the authors of this study. The performance of 2005 (i.e., the status of 2005) was compared against the average of the performance in 2003 and 2004 (i.e., the combined status of 2003 and 2004). The result would be an improvement (i.e., status change) as opposed to growth, since the data used were not panel data (i.e., individual students had not been tracked over time).

Based on Louisiana’s current school accountability system, 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.

#### Phase II: The Sample of the Improvement Model

The Phase II sample consisted of all 63,553 students enrolled in Grades 3, 4, and 5 for 2005-06 in all 299 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. Of the 63,553 students included in the Phase II study, about 48.32% were females while 51.68% were males. African American students accounted for about 50.29% of the total student population statewide, with White, Hispanic, Asian, and Native American students account for 45.48%, 2.48%, 1.34%, and 0.41% respectively. About 67.83% of the students received either free or reduced-price lunch, and about 19.65% of the students were students with disabilities. There were a very small percentage of students with limited English proficiency, since LEP students account for only 1.46% of the total student population.

#### Phase II: The Measurements of the Improvement Model

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 is administered at grades 3, 5, 6, 7, 9 in subject areas of Reading, Language, Mathematics, Science, Social Studies, and Sources of Information. The ITBS has been used in Louisiana since 1998 (Louisiana Department of Education, 2005b).

#### Phase II: Data Analysis of the Improvement Model

The first step in Phase II analysis involved calculating the CRT index for Grade 4, and the NRT index 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 are categorized as “academically unacceptable”. Improvement in SPS was examined, and the actual improvement was compared against the expected improvement.

#### Phase III: Model Specifications of the Growth Model

North Carolina’s Modified ABCs Model was chosen as the representative of growth models. North Carolina calculates a four-year<sup>4</sup> 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).

Three concepts are important in understanding how the Modified ABCs Model works. These concepts are (1) calculating a four-year growth trajectory, (2) calculating

---

<sup>4</sup> In this study, a five-year growth trajectory was used.

the c-scores, and (3) determining whether the student is on track towards proficiency (NCDPI, 2006a, 2006b, 2006c).

#### *Calculating a Four-Year Growth Trajectory*

The first thing one must do 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 will be calculated, the calculation of which is discussed in the next sub-section (NCDPI, 2006a, 2006b, 2006c).

#### *Calculating C-Scores*

One of the main concepts of the North Carolina's Modified ABCs model is "change score", which is also called "c-score". "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" (NCDPI, 2006a, p. 2). The calculation of c-score is as follows:

$$\text{C-Scale} = (\text{scale score} - \text{state mean}) / \text{standard deviation} \quad (5)$$

Where

State mean = the state mean of the standard setting year

Standard deviation = the standard deviation of the standard setting year

*Determining Whether Students Are on Track towards Proficiency*

Whether a student is on track towards proficiency is determined by baseline c-score, c-score for the current year, the number of years the student has been in the state public schools, and the c-score needed to be proficient four years from the baseline year. Using an example provided by the North Carolina Department of Public Instruction (2006), a student enters North Carolina in the 3<sup>rd</sup> grade and remains in North Carolina public schools for 4<sup>th</sup> and 5<sup>th</sup> 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 seventh grade. Assume a score of 252 is needed in order to be proficient in the seventh grade. Using Formula (5), we can obtain a c-score of  $-1.00$  for the score of 252. In other words, the student needs a c-score of  $-1.00$  when he is in the 7<sup>th</sup> grade.

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 7<sup>th</sup> grade proficiency c-score is 2.08 ( $3.08$  minus  $1.00$ ). During the four years between 4<sup>th</sup> grade and 7<sup>th</sup> 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-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 7<sup>th</sup> grade. Since the student is in the 5<sup>th</sup> grade now, the student will need a c-score of  $-2.04$  to be counted on track towards proficiency. The student will be counted as proficient, although a c-score of  $-2.04$  may be lower than the c-score needed for a regular 5<sup>th</sup> grader to be counted as proficient.

### Phase III: The Sample for the Growth Model

The sample of Phase III analysis consisted of all students in Grades 3, 4, and 5 in all 299 K-5 schools in the state of Louisiana that had at least two data points (about 76.84% of the students). After eliminating students with less than 2 data points, a total number of 41,203 students were left for Phase III data analysis. Of the 41,203 students included in the Phase III study, about 48.88% were females while 51.12% were males. African American students accounted for about 50.79% of the total student population statewide, with White, Hispanic, Asian, and Native American students accounting for 45.44%, 2.15%, 1.25%, and 0.37%, respectively. About 66.97% of the students received either free or reduced-price lunch, and about 20.27% of the students were students with disabilities. There were a very small percentage of students with limited English proficiency, since LEP students accounted for only 1.21% of the total student population. Since Phase III analysis used linked data (i.e., tracking the same students over time), a total number of 103,706 records were used.

### Phase III: Data Analysis of the Growth Model

The first step in Phase III data analysis was determining the number of years the non-proficient student had been in Louisiana public schools. A five-year growth trajectory was used in this study. In other words, if the student had been in Louisiana state public schools for five years or more and was not proficient for the school year of 2004-05, 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. One then

calculates a c-score for the baseline. The second step in Phase III data analysis involved 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).<sup>5</sup> The third step involved determining whether the c-score difference between the baseline year and the target year had been minimized by 20% each year. If the student was able to reduce the difference by at least 20% each year during the four years, the student was counted as on trajectory towards proficiency.

#### Phase IV: Model Specifications for the Value-Added Model

A value-added model calculates how much “value” a school adds to student learning. The value-added model in Phase IV analysis was based on a three-level Hierarchical Linear Model (Raudenbush & Bryk, 2002), in which measurement occasions were viewed as nested within students, and students were viewed as nested within schools. Specifically, the model is specified as follows:

$$\text{Level-1 Model: } Y_{tij} = \pi_{0ij} + \pi_{1ij}a_{tij} + e_{tij} \quad (6)$$

Where

$Y_{tij}$  = the achievement outcome measure of student  $i$  in school  $j$  at time  $t$ ;

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

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

$a_{tij}$  = a measure of time ( 2003 = -1, 2004 = 0, and 2005 = 1)

$e_{tij}$  = the level 1 random error term for student  $i$  in school  $j$

---

<sup>5</sup> Note that a five-year growth trajectory was used in this study. This is different from North Carolina’s four-year growth trajectory.

Level-2 Model:

$$\pi_{0ij} = \_00j + \_01j (\text{SES}) + \_02j (\text{Gender}) + \_03j (\text{Ethnicity}) \quad (7)$$

$$+ \_04j (\text{LEP}) + r_{0ij}$$

$$\pi_{1ij} = \_10j + \_11j (\text{Ethnicity}) + r_{1ij} \quad (8)$$

$$\text{Level-3 Model: } \_00j = \_000 + \_00j \quad (9)$$

$$\_01j = \_010$$

$$\_02j = \_020$$

$$\_03j = \_030$$

$$\_04j = \_040$$

$$\_10j = \_100 + \_10j \quad (10)$$

$$\_11j = \_110$$

where

$\_000$  = the overall mean (grand mean of schools) for initial status;

$\_010$  = the effect of SES on initial achievement status;

$\_020$  = the effects of gender on initial achievement status;

$\_030$  = the effects of ethnicity on initial achievement status;

$\_040$  = the effects of LEP on initial achievement status;

Similarly,

$\_100$  = the overall mean (grand mean of schools) for academic year achievement growth rate;

$\_110$  = the effect of SES on achievement growth rate;

#### Phase IV: The Sample of the Value-Added Model

The sample of Phase IV study consisted of all students that were enrolled in Grade 5 for the school year of 2005-06 in all 299 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 (2004-2005) and Grade 3 (2003-2004) were retrieved based on their unique student IDs. Students who did not have test records in all three grades were deleted, which reduced the total number of students from 20,231 to 14,632. 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.

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 students with disabilities. 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. Since Phase IV analysis used linked data (i.e., tracking the same students over time), a total number of 35,991 records were used.

#### Phase IV: Variables of the Value-Added Model

Various student background variables are 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 of the Value-Added Model

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 is 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 we code the measure of time,  $a_{tij}$ , as -1, 0, and 1 for 2003, 2004, and 2005 respectively, the slope represents achievement gains during the three-year period of primary school.

$$\text{Level-1 Model: } Y_{tij} = \pi_{0ij} + \pi_{1ij}a_{tij} + e_{tij} \quad (11a)$$

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, and LEP status were added later on in the conditional model.

$$\pi_{0ij} = \_00j + r_{0ij} \quad (11b)$$

$$\pi_{1ij} = \_10j + r_{1ij} \quad (11c)$$

$$\_00j = \_000 + \_00j \quad (11d)$$

$$\_10j = \_100 + \_10j \quad (11e)$$

The level-3 model also started with a fully unconditional model, and then school-level variables were added. 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 (Teddle & Reynolds, 2000). Once the predictors at each level were specified, a value-added index could be calculated for each school.

## Results

The results from the four models are presented in this section, which consists of five subsections. Each of the first four subsections concerns findings from one model. Specifically, in each of the first four subsections, a school effectiveness index was calculated for every school based on the student population as a whole. 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)

The proficiency rates of the 299 K-5 schools in Louisiana were calculated, and the observed proficiency rates were compared against the target proficiency rate of 41.8%. It was found that 249 out of 299 (83.28%) 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 early years 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 ( $n > 10$ ) (CCSSO, 2003; Linn, Baker, & Herman, 2005).

Methods (1) and (2) have already been applied to the data analysis of Research Question # 1, the results of which were presented earlier. Method (4) doesn’t apply because the sample size is always greater than 10 when the student population as a whole

is examined. Therefore, the only method that applies but has not yet been applied is Method (3), calculating a confidence interval around each school's observed proficiency rate. 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 286 out of 299 (95.65%) of the schools were able to hit the target proficiency rate. There was a 12.37% increase of schools that made AYP. Confidence interval makes it easier for schools to hit the target. In the case of Louisiana, a school whose observed proficiency rate was 27.66% (13 out of 47 students) was regarded as making AYP because the upper bound of the confidence interval was 42.86% (higher than the required 41.8% proficiency rate).

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, all schools were categorized as making AYP for the subgroup of all students. In this sense, AYP was incapable of identifying ineffective schools since all most all schools had made AYP. As schools in Louisiana were subjected to both NCLB requirements and state accountability requirements, what were really functioning were the state accountability requirements only.

### Results from the Improvement Model (Phase II)

The SPS of the 299 K-5 schools in Louisiana were calculated. The observed SPS ranged from 25.56 to 147.07. A number of schools ( $n = 22$ ) had a SPS of 120 or higher, which meant that these schools had already reached Louisiana's academic goal of 2013-14. When compared against the minimum requirement of 60, it was found that 247 out of 299 (82.61%) schools were able to reach or exceed the minimum requirement. However, there were still a significant number of schools ( $n = 52$ ) 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.

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 1** includes detailed information regarding the performance labels of the 299 K-5 schools in Louisiana.

It is worthwhile to point out that the performance labels assigned to the 299 K-5 schools in this study were somewhat different from the actual performance labels they got from the Louisiana Department of Education. The actual performance labels were based on four subject areas (i.e., English Language Arts, Math, Science, and Social Studies) 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 correlation analysis was

conducted between the SPS calculated in this study and those published by LDOE 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.

Table 1. Performance Labels Based on SPS

Performance Label	SPS	Number of Schools Identified in This Study	Number of Schools Identified by Louisiana Department of Education
Five Stars	140.0 or above	1	0
Four Stars	120.0 – 139.9	21	12
Three Stars	100.0 – 119.9	65	66
Two Stars	80.0 – 99.9	86	106
One Star	60.0 – 79.9	74	80
Academically Unacceptable	Below 60	52	33

Schools in Louisiana are also assigned growth labels. The growth target was set based on Louisiana’s 2014 goal of all schools reaching a SPS of 120. Specifically, the average SPS in 2003 and 2004 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 299 K-5 schools. Of the 299 schools examined, 67 (22.41%) of them showed “exemplary academic growth”, 83 (27.76%) of them showed “recognized academic growth”, 45 (15.05%) of them showed “minimum

growth”, 22 (7.36%) of them showed “no growth”, and 63 (21.07%) of them were in decline. **Table 2** showed detailed information regarding the growth labels of all 299 schools.

Similar with the performance labels, the growth labels calculated in this study were somewhat different from the actual performance labels schools got from the Louisiana Department of Education. The actual growth labels were based on four subject areas (i.e., English Language Arts, Math, Science, and Social Studies) 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 very high correlation ( $r=0.85$ ) with the SPS based on four subjects.

Table 2. Growth Labels of the Schools

Growth Labels	Number of Schools Identified in This Study	Number of Schools Identified by Louisiana Department of Education
Exemplary Academic Growth	67 (22.41%)	128 (43.10%)
Recognized Academic Growth	83 (27.76%)	61 (20.54%)
Minimum Academic Growth	45 (15.05%)	46 (15.49%)
No Growth	22 (7.36%)	25 (8.42%)
In Decline	62 (20.74%)	22 (7.41%)
No Label <sup>6</sup>	20 (6.69%)	15 (5.05%)

<sup>6</sup> 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 were identified and an individual growth trajectory was depicted. Based on the individual growth trajectory, students who were not proficient for the current year but were projected to be proficient five years into the future were re-categorized as proficient for the current year.

A total number of 4997 (out of 15,197) 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 1225 out of 4997 (24.51%) students could be re-categorized as proficient because they were on track towards proficiency five years into the future.

Similarly with what was done with fourth graders, a total number of 6190 (out of 11248) students who scored below the 55<sup>th</sup> percentile<sup>7</sup> on ITBS Math were identified and individual five-year growth trajectories were calculated. It was found that 1108 out of 6190 (17.90%) students could be re-categorized as above the 55<sup>th</sup> percentile because they were on track towards the 55<sup>th</sup> percentile five years into the future. Individual growth trajectories were not calculated for third graders because they had only the baseline (i.e., one year's data), and no growth tendency could be obtained.

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

---

<sup>7</sup> The 55<sup>th</sup> percentile on the NRT is largely equivalent with the "basic" level of academic achievement on the CRT. Louisiana's ten – year goal was for every student to reach "basic" on the CRT and 55<sup>th</sup> percentile on the NRT.

# 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, 38 schools had a proficiency rate lower than the target proficient rate. In other words, 12 out of 50 schools could be re-categorized as making AYP.

#### Results from the Value-Added Model (Phase IV)

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 & Raudenbush, 2002). The unconditional model of this study was presented in Formula 12a – 12d. 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 ( $\_00$ ) and the random variation within the school ( $r_0$ ); and  $\pi_1$  was model as a function of average growth in standardized score for the school ( $\_10$ ) and random variation in growth in standardized score within the school ( $r_1$ ). Finally, at level-3,  $\_00$  was modeled as the average standardized score across all schools statewide ( $\_000$ ) and the random variation in school initial status ( $\_00$ ) between schools; and  $\_10$  was modeled as the average growth in standardized score across all schools statewide ( $\_000$ ) and the random variation in school growth between schools ( $\_10$ ).

$$\text{Level 1 Model: } Z\text{MATH} = \pi_0 + \pi_1 (\text{YEAR}) + e \quad (12a)$$

$$\text{Level 2 Model: } \pi_0 = \_00 + r_0$$

$$\pi_1 = \_10 + r_1 \quad (12b)$$

$$\text{Level 3 Model: } \_00 = \_000 + \_00$$

$$\_10 = \_100 + \_10 \quad (12c)$$

$$\text{Mixed Model: } Z\text{MATH} = \_000 + \_100 * \text{YEARCD} + r_0 + r_1 * \text{YEARCD}$$

$$+ \_00 + \_10 * \text{YEARCD} + e \quad (12d)$$

### Results from the Unconditional Model

The results of the unconditional model are presented in **Table 3**. The estimated mean intercept,  $\_000$ , and mean growth rate,  $\_100$ , 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  $\_100$ . 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 ( $\chi^2 = 1061.24$ ,  $P < .001$ ).

### Results form 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 13a – 13d.

$$\text{Level 1 Model: } ZMATH = \pi_0 + \pi_1 (\text{YEAR}) + e \quad (13a)$$

$$\begin{aligned} \text{Level 2 Model: } \pi_0 &= \_00 + \_01 (\text{GENDER}) + \_02 (\text{ETHNIC}) + \_03 (\text{SES}) \\ &+ \_04 (\text{LEPFLG}) + r_0 \\ \pi_1 &= \_10 + \_11 (\text{ETHNIC}) + r_1 \end{aligned} \quad (13b)$$

$$\begin{aligned} \text{Level 3 Model: } \_00 &= \_000 + \_00 \\ \_01 &= \_010 \\ \_02 &= \_020 \\ \_03 &= \_030 \\ \_04 &= \_040 \\ \_10 &= \_100 + \_10 \\ \_11 &= \_110 \end{aligned} \quad (13c)$$

$$\begin{aligned} \text{Mixed Model: } ZMATH &= \_000 + \_010 * \text{GENDER} + \_020 * \text{ETHNIC} \\ &+ \_030 * \text{SES} + \_040 * \text{LEPFLG} + \_100 * \text{YEAR} \\ &+ \_110 * \text{ETHNIC} * \text{YEAR} + r_0 + r_1 * \text{YEAR} \\ &+ \_00 + \_10 * \text{YEAR} + e \end{aligned} \quad (13d)$$

The results of the conditional model are shown in **Table 4**. 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 ethnical backgrounds on a yearly basis. The middle panel of **Table 4** 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  $_{00}$  were fine, but the reliability for  $_{10}$  was very low. According to Bryk and Raudenbush (2002), low reliability on certain coefficient didn't discredit the whole HLM model.

#### The Value-Added Component

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  $_{100} + _{10}$  (i.e.,  $_{10}$ ) (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  $\gamma_{10}$  values. These schools were viewed as adding the most value to student learning based on the value-added model.

Table 3. Three-Level Unconditional Model

<i>Fixed Effect</i>	<i>Coefficient</i>	<i>se</i>	<i>t Ratio</i>	
Average initial status, $\gamma_{000}$	- 0.074585	0.022774	-3.275*	
Average yearly growth rate, $\gamma_{100}$	- 0.000019	0.006565	-0.003	
<i>Random Effect</i>	<i>Variance</i>	<i>df</i>	<i>P Value</i>	
Level 1: Within-student variation, $e$	0.24193			
Level 2 (students within schools)				
Individual initial status, $r_{0ij}$	0.62744	10949	44898.78784	0.000
Individual growth rate, $r_{1ij}$	0.00057	10949	10653.83110	0.000
Level 3 (between schools)				
School mean status, $\gamma_{00j}$	0.12939	298	2367.84631	0.000
School mean growth rate, $\gamma_{10j}$	0.00878	298	1061.24279	0.000

Table 4.6. Three-Level Conditional Model

<i>Fixed Effect</i>	<i>Symbol</i>	<i>df</i>	<i>Coefficient</i>	<i>Standard Error</i>
Grand Mean	$\gamma_{000}$	298	0.351292*	0.025223
Gender	$\gamma_{010}$	11243	0.085962*	0.015734
Ethnic	$\gamma_{020}$	11243	-0.515583*	0.021397
SES	$\gamma_{030}$	11243	-0.306860*	0.019929
LEP Status	$\gamma_{040}$	11243	-0.311688*	0.072036
Year (slope)	$\gamma_{100}$	298	0.020767*	0.007098
Ethnic	$\gamma_{110}$	11246	-0.040707*	0.007499
<i>Random Effect</i>	<i>Symbol</i>	<i>df</i>	<i>Variance</i>	<i>% Variance</i>
School Status	$\gamma_{00}$	298	0.04862*	5.62%
School Slope	$\gamma_{10}$	298	0.00885*	1.02%
Student Status	$r_0$	10945	0.56522*	65.35%
Student Year (Slope)	$r_1$	10948	0.00053	0.0006%
Within Student	$e$		0.24167	27.94%
<i>Reliability Estimates</i>	<i>Symbol</i>	<i>Reliability Estimates</i>		
Student	$\pi_0$	0.875		
Year (Slope)	$\pi_1$	0.004		
School	$\gamma_{00}$	0.689		
School Year (Slope)	$\gamma_{10}$	0.683		

Figure 1. The “Value” Added by Schools

L3ID	NK	EB00	EB10	EC00	EC01	EC02	EC03	EC04	EC10
018005	21	0.436168512	0.271580308	0.787460978	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.292347548
010005	35	-0.06293366	0.200184827	0.288358803	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.220952067
017072	44	0.011483759	0.197810983	0.362776225	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.219578223
005019	51	-0.05717149	0.161266927	0.294120972	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.182034168
009028	24	-0.07105155	0.151125024	0.280240921	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.171892265
017011	31	-0.14154864	0.149131453	0.209743826	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.169896693
009053	38	0.310103779	0.148777769	0.661396244	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.169545009
023030	34	-0.1424428	0.1454931	0.208849664	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.16626034
008012	30	-0.02470058	0.144312942	0.32659189	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.165080182
026009	60	-0.15699182	0.139204766	0.194300645	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.159972006
010029	24	0.22609847	0.139200855	0.577390936	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.159968095
026069	26	0.061455505	0.137519395	0.412747971	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.158286635
051018	24	0.09669443	0.133414309	0.447986896	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.154181549
010024	35	0.274183103	0.132240294	0.625475569	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.153007534
009061	16	0.184944402	0.126651418	0.536236868	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.147418658
051002	11	0.196336414	0.118227521	0.54762888	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.138994761
028024	41	0.287748422	0.117510527	0.639040888	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.138277767
040031	31	0.118497766	0.117148627	0.469790232	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.137915867
040026	47	-0.03511706	0.115838793	0.31617541	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.136606033
017062	19	0.107372215	0.1131259	0.458664681	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.13389314
017027	31	0.053137609	0.108601398	0.404430075	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.129368538
017018	37	0.173630087	0.108453699	0.524922553	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.129220939
017019	37	0.66970947	0.107312895	1.021001936	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.128080135
008018	38	0.041904874	0.105905184	0.39319734	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.126672424
032004	39	0.223827591	0.105271058	0.575120057	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.126038298
052017	36	0.113352878	0.103896813	0.464645344	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.124664053
026003	65	0.059562334	0.098125656	0.4108548	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.118892896
017022	56	-0.00577863	0.097863446	0.34551384	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.118630686
017032	20	-0.02817908	0.095211787	0.32311339	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.115979027
017100	13	0.050553525	0.094361797	0.401845991	0.085962178	-0.51558296	-0.30685994	-0.31168787	0.115129037

Results from Research Questions 5 and 6

The four models produced four indicators of school effectiveness for every school. Research Questions 5 and 6 concern the extent to which these four indicators point to the same direction regarding the effectiveness of the same schools. A correlation analysis was conducted, the results of which were shown in **Table 5**. M1, M2, M3, and M4 stand for the 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.,  $\_10$  values), 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 5. Comparing the Four Models

	M1	M2	M3	M4
M1	1.000			
M2	0.994231*	1.000		
M3	0.994414*	0.89925*	1.000	
M4	0.11186	0.11424	0.15284*	1.000

Since the four models reached different conclusions regarding the effectiveness 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), 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 6** presents the 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 meant 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 6. Using “Fairness” as the Additional Criterion

	Gender	Ethnicity	SES	LEP	SPED
M1	-0.06298	-0.66871	-0.17735	-0.03424	-0.07758
M2	-0.07910	-0.74376	-0.21683	-0.04830	-0.08182
M3	-0.00995	-0.62437	-0.18792	-0.05672	-0.03164
M4	0.04172	0.08200	0.00431	-0.01849	0.08171

### Discussion

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. In this study, four generic types of accountability models of varying theoretical framework

and complexity were examined and it was found that the four models reached inconsistent conclusions regarding the effectiveness of the same set of schools. The school effectiveness indices produced by the status model, the improvement model, and the growth model diverged significantly from the value-added model but converged largely among themselves.

From a policy point of view, 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 were compared in terms of “fairness” and it was found that school effectiveness indices produced by the value-added model had the lowest correlation with various student background variables.

Reference

- Accountability Technical Advisory Committee. (1998, October 26). Recommendations regarding the Louisiana school accountability system. Retrieved October 15, 2006, from [http://www.nciea.org/publications/LASchlDesign\\_TAC98.pdf](http://www.nciea.org/publications/LASchlDesign_TAC98.pdf)
- Bryk, S. A., Thum, M. Y., Easton, Q. J., & Luppescu, S. (1998). *Academic productivity of Chicago public elementary schools*. Chicago, IL: Consortium on Chicago School Research.
- Bulletin 111 – Please refer to Louisiana Department of Education (2006) below.
- Choi, K., Goldschmidt, P., & Yamashiro, K. (2006). Exploring models of school performance: From theory to practice (CSE Rep: No. 673). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2004). *Children left behind in AYP and Non-AYP schools: Using student progress and the distribution of student gains to validate AYP* (CSE Rep: No. 637). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED483404)
- Council of Chief State School Officers. (2003). *State Support to Low-Performing Schools*. Washington, DC: Author.
- CTB/McGraw-Hill LLC. (2002). Accountability and educational progress: A primer on the assessment provisions of the new ESEA. Retrieved November 17, 2005, from [http://www.ctb.com/media/articles/pdfs/general/ctb\\_esea\\_primer.pdf](http://www.ctb.com/media/articles/pdfs/general/ctb_esea_primer.pdf)
- Doran, H., & Izumi, L. (2004). Putting education to the test: A value-added model

- for California. Retrieved August 11, 2006, from  
<http://www.heartland.org/pdf/15626.pdf>
- Goldschmidt, P., Roschewski, P. Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005). *Policymakers' guide to growth models for school accountability: How do accountability models differ?* A paper commissioned by the Councils of Chief State School Officers, Washington, DC.
- Hanushek, E., & Raymond, M. (2002). Improving educational quality: How best to evaluate our schools? Retrieved September 30, 2006, from  
<http://edpro.stanford.edu/hanushek/admin/pages/files/uploads/accountability.BostonFed.final%20publication.pdf>
- Herman, J., Baker, E., & Linn, R. (2005). Chickens come home to roost. CRESST Line: Newsletters of the National Center for Research on Evaluation Standards, and Student Testing, pp. 1, 3, 7-8.
- Linn, R. L., Baker, E. L., & Herman, J. L. (2002, Fall). From the directors: Minimum group size for measuring adequate yearly progress. The CRESST Line, 1, 4-5. Retrieved November 24, 2005, from  
[http://www.cresst.org/products/newsletters\\_set.htm](http://www.cresst.org/products/newsletters_set.htm)
- Linn, R. L & Haug, C. (2002). *Stability of school building accountability scores and gains* (CSE Rep: No. 561). Los Angeles, CA: Center for the Study of Evaluation.
- Lissitz, R., Doran, H., Schafer, W., and Willhoft, J. (2006). Growth modeling, value-added modeling and linking: An introduction. In R. Lissitz (Ed.), *Longitudinal and Value-Added Models of Student Performance*. Maple Grove, Minnesota: JAM Press.

Lockwood, A. (2005). Educators grapple with NCLB's Demands for Accountability.

Retrieved September 24, 2006, from

<http://www.nwrel.org/nwedu/10-04/brief/cloak/brief.pdf>

Louisiana Department of Education. (2005a). The Louisiana Educational Assessment

Program. Retrieved December 12, 2006, from

<http://www.doe.state.la.us/lde/saa/2273.html>

Louisiana Department of Education. (2005b). Louisiana Statewide Norm-Referenced

Testing Program: The Iowa Test. Retrieved December 12, 2006, from

<http://www.doe.state.la.us/lde/saa/2273.html>

Louisiana Department of Education. (2006). Louisiana School, District, and State

Accountability System (Bulletin 111). Baton Rouge, LA: Author.

Massachusetts Department of Education (2006). Massachusetts charter school

achievement comparison study: An analysis of 2001-2005 MCAS performance

(Appendix G: Statistical methods information). Retrieved April 1, 2007, from

<http://www.doe.mass.edu/charter/reports/datastudy/appendixG.pdf>

McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, S. (2003). *Evaluating value-*

*added models for teacher accountability*. Santa Monica, CA: RAND.

Meyer, R. (1995). *Educational performance indicators: A critique* (Rep. No: 1052-94).

Madison, Wisconsin: Institute for Research on Poverty Discussion

National Association of State Boards of Education. (2002). The No Child Left Behind

Act: What states need to know? Retrieved November 17, 2005, from

[http://www.nasbe.org/Front\\_Page/NCLB/NCLBFeedback\\_validation.html](http://www.nasbe.org/Front_Page/NCLB/NCLBFeedback_validation.html)

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

Retrieved October 11, 2005, from

<http://www.ed.gov/policy/elsec/leg/esea02/index.html>

North Carolina Department of Public Instruction (2006a, April). North Carolina's proposal to pilot the use of a growth model for AYP purpose in 2005-06.

Retrieved Nov 14, 2006, from

<http://www.ed.gov/admins/lead/account/growthmodel/nc/ncayp.doc>

North Carolina Department of Public Instruction. (2006b). The ABCs Model for 2005-06.

Retrieved November 18, 2006, from

<http://www.dpi.state.nc.us/docs/accountability/reporting/growth/theabcsmodelfor2005.pdf>

North Carolina Department of Public Instruction. (2006c, November). Technical Notes.

Retrieved January 11, 2007, from

<http://www.ncpublicschools.org/docs/accountability/reporting/growth/technotesnewformulas.pdf>

Louisiana Register. (2002). Vol. 28, No. 07.

Olson, L. (2005, May 18). States hoping to "grow" into AYP success. *Education Week*, 24(37), pp. 15-20.

Porter, A. C., Linn, R. L., & Trimble, S. (2005). The effects of state decisions about NCLB Adequate Yearly Progress targets. *Educational Measurement: Issues and Practice*, 24(4), 32-39.

Public Affairs Research Council of Louisiana. (1999). Educational accountability and the role of school choice.

Public Affairs Research Council of Louisiana. (2004). NCLB: A steep climb ahead

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd Edition). Thousand Oaks, CA: Sage Publications
- Reeves, C. (2003). State support to low-performing schools. Council of Chief State School Officers, Washington, D. C.
- Riddle, W. C. (2005, August 15). *Adequate yearly progress (AYP): Might growth models be allowed under the No Child Left Behind Act?* Washington, DC: Congressional Research Service. Note: This paper was updated March 8, 2006, and retitled *Adequate Yearly Progress (AYP): Growth Models Under the No Child Left Behind Act*.
- Rumberger, R. W., & Palardy, G. J. (2003). Multilevel models for school effectiveness research. In D. Kaplan (Ed.) *Handbook of Quantitative Methodology for the Social Sciences* (pp.235-258). Thousand Oaks, CA: Sage Publications.
- Sanders, W., Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation*, 9, 299-311.
- Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie and D. Reynolds (Eds.), *The International Handbook of School Effectiveness Research* (pp.55-133). New York: Falmer Press.
- Teddlie, C., & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research*. New York: Falmer Press.
- U. S. Department of Education. (2006, May 31). Growth models: Press releases and

letters, guidance and fact sheets, state applications. Retrieved October 12, 2006,  
from <http://www.ed.gov/admins/lead/account/growthmodel/index.html>

U. S. Government Accountability Office. (2006). No Child Left Behind: States face  
challenges measuring academic growth.

Webster, W., Mendro, R., Orsak, T., & Weerasinghe. (1998). An application of  
hierarchical linear modeling to the estimation of school and teacher effect. Paper  
presented at the annual meeting of the American Educational Research  
Association.