USING A YARDSTICK TO MEASURE A METER: GROWTH, PROJECTION, AND VALUE-ADDED MODELS IN THE CONTEXT OF SCHOOL ACCOUNTABILITY

Michael Joseph Weiss

A DISSERTATION in Education

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2008

Margaret E. Spritz
Supervisor of Dissertation

Graduate Group Chairperson
COPYRIGHT

Michael Joseph Weiss

2008
Acknowledgements

This dissertation has been greatly enhanced because of the support and guidance I have received from many individuals.

First, I am especially appreciative of my dissertation committee for their skilled guidance and expertise. To Peg Goertz, my Chair, who kept me “on track,” organized, and whose policy expertise greatly improved the quality of my work. To Henry May, my statistical guru, for his methodological knowledge, in-depth editing of my work, and statistically significant entertainment value ($p < .01$). To Rebecca Maynard, for helping reign in the scope of this work and for her support of my development as a researcher. To Ruth Curran Neild, for her persistent focus on the practical importance of this work. As I have said before, “best dissertation committee ever.”

Thank you to my colleagues and friends at the Consortium for Policy Research in Education at the University of Pennsylvania for making the work environment both intellectually stimulating and an enjoyable place to spend four years. Special thanks to Elliot Weinbaum for giving needed advice, providing endless editing support and for generally guiding me through the entire process. Thanks to Russell Cole for his shared enthusiasm in discussing statistical methods and this work with me, and giving me thoughtful and extremely helpful comments.

Special thanks to the Institute of Education Sciences (IES), U.S. Department of Education, whose IES pre-doctoral fellowship provided financial assistance to me while I was completing this research. This funding came through Grant R305C050041-05 to the University of Pennsylvania; however, the opinions expressed in this research are those of the authors and do not represent views of the U.S. Department of Education.
Finally, I want to express my deepest gratitude to my family, for their love and support, as well as their understanding and patience.
ABSTRACT

USING A YARDSTICK TO MEASURE A METER:
GROWTH, PROJECTION, AND VALUE-ADDED MODELS
IN THE CONTEXT OF SCHOOL ACCOUNTABILITY

Michael Joseph Weiss
Margaret E. Goertz

As test-based educational accountability has moved to the forefront of national and state educational policy, so has the desire for better measures of school performance. Past debates focused on the potential benefits of using value-added models to measure school performance. These statistical models represent the most sophisticated technique for assessing the unique contribution of schools to the learning gains of their students. However, recent national policy has shifted the discussion. The federally created Growth Model Pilot Program (GMPP) permits states to use projection models in their accountability systems. Most states involved in the pilot program are using projection models to give schools “credit” for those students who have not yet achieved proficiency, but have made sufficient learning gains such that they appear on track to become proficient in the near future. While value-added measures of school performance attempt to compare schools’ relative effectiveness, the proposed projection models are used to assess whether schools are effective at getting their students up to a fixed learning target - proficiency.

This dissertation has two components. The first is an empirical comparison of different state projection models, answering the questions: How accurate are states’ projection models? Are their goals realistic? The second is a comparison of three
approaches to measuring school performance: the status model (used under No Child Left Behind (NCLB)), the projection models (used under the GMPP), and value-added models (used in some state accountability systems).

Findings suggest that the GMPP's projection models are not very accurate. Moreover, even if they were perfectly accurate, the GMPP's measures of school performance are so similar to NCLB's status measure that this new program is unlikely to have significant impact on the way we assess schools' performances. In contrast, using value-added models would dramatically shift the focus of our education accountability system. Highlighted in this work is the inherent tension in the desire to compare schools' relative effectiveness while simultaneously holding schools accountable for bringing all students up to high achievement levels.
Table of Contents

Acknowledgements ........................................................................................................ iii
Abstract .......................................................................................................................... v
List of Tables ................................................................................................................ ix
List of Figures ................................................................................................................ x

Chapter 1: Introduction ............................................................................................... 1
Research Questions ........................................................................................................ 14

Chapter 2: Background and Literature ................................................................. 16
The Federal Growth Model Pilot Program (GMPP) ................................................ 16
Florida – Simple Linear Projection ............................................................................. 21
Arizona – Regression Adjusted Linear Projection ..................................................... 25
North Carolina – Linear Growth after Standardizing Scale Scores and Proficiency Cut Scores ........................................................................................................... 28
Arkansas – Proportional Growth ................................................................................ 33
Tennessee – Longitudinal Statistical Model (EVAAS®) .............................................. 37
The (Potential) Added Value of Value-Added Models ................................................ 41
Value-Added Models and Causal Inference ................................................................. 45

Chapter 3: Data and Methods ................................................................................... 52
Data .............................................................................................................................. 52
Methods ...................................................................................................................... 58
Florida’s Projection Model ......................................................................................... 59
Arizona’s Projection Model ......................................................................................... 60
North Carolina’s Projection Model ............................................................................. 62
Arkansas’ Projection Model ......................................................................................... 64
Tennessee’s Projection Model (EVAAS®) .................................................................. 66

Chapter 4: Analyses and Results .............................................................................. 71
Research Question 1: How accurate are projection models at the student level? .... 71
Research Question 2: How accurate are projection models at the school level? ..... 87
Research Question 3: Are growth expectations realistic? ......................................... 95
Research Question 4: How do status, projection, and value-added measures of school performance compare to each other? ................................................................. 102

Chapter 5: Discussion ............................................................................................... 108
Projection Model Accuracy ........................................................................................ 108
Realistic Goals ............................................................................................................ 123
Comparing Measures of School Performance .......................................................... 126

Chapter 6: Study Limitations ................................................................................... 133

Chapter 7: Conclusions ........................................................................................... 138
List of Tables

Table 1.1 Consequences of Failing to Make Adequate Yearly Progress.......................... 2
Table 2.1 Projection Model Descriptions ........................................................................... 21
Table 2.2 Florida's Growth Model Applied to a Hypothetical Student ............................... 22
Table 2.3 Arizona's Growth Model Applied to a Hypothetical Student ............................. 26
Table 2.4 North Carolina's Growth Model Applied to a Hypothetical Student ............... 30
Table 2.5 Arkansas' Growth Model Applied to a Hypothetical Student ............................ 34
Table 3.1 Frequency Counts for the 3rd Grade Cohort of 2001-2002 ................................. 54
Table 3.2 Demographic Characteristics of the 3rd Grade Cohort of 2001-2002 ............... 55
Table 3.3 Mathematics FCAT Scale Score Descriptive Statistics for the Study District. .... 56
Table 4.1 Projection Model Accuracy (Student Level) ....................................................... 75
Table 4.2 Concordance between Current Status and Future Proficiency (Student Level) 77
Table 4.3 Descriptive Statistics for Projection Model Residuals (School Level) ............. 91
Table 4.4 Descriptive Statistics by 3rd Grade Decile ....................................................... 97
Table 4.5 Correlations Between Various Measures of School Performance .................. 104
Table 5.1 Percent Change in Proficiency Cut Score vs. Observed Scale Scores.............. 117
List of Figures

Figure 2.1 Florida’s Developmental Scale Proficiency Cut Scores and on Track to Become Proficient Growth Targets for a Hypothetical Student .................................................. 23

Figure 2.2 North Carolina’s C-Scale Proficiency Cut Scores and on Track to Become Proficient Growth Targets for a Hypothetical Student ................................................. 30

Figure 2.3 A Comparison of Florida and North Carolina Growth Targets on Florida’s Developmental Scale for a Hypothetical Student ......................................................... 33

Figure 2.4 Arkansas’ Developmental Scale Proficiency Cut Scores and on Track to Become Proficient Growth Targets for a Hypothetical Student ............................................. 35

Figure 2.5 A Comparison of Florida and Arkansas Growth Models on Florida’s Developmental Scale for a Hypothetical Student ............................................................... 36

Figure 2.6 A Comparison of the Required Gains under a Status Model (NCLB) vs. a Projection Model (GMPP) ........................................................................................................ 40

Figure 4.1 Box Plot of Residuals from One-year Projections (Student Level) .......... 82

Figure 4.2 Box Plot of Residuals from Two-year Projections (Student Level) ......... 84

Figure 4.3 Plot of Residuals vs. Projected Values for Arkansas’ Two-year Projections (Student Level) ............................................................................................................. 85

Figure 4.4 Box Plot of Residuals from One-year Projections (School Level) .......... 88

Figure 4.5 Box Plot of Residuals from Two-year Projections (School Level) ......... 90

Figure 4.6 Plot of Residuals vs. Projected Values for Arkansas’ One-year Projection Model (School Level) ........................................................................................................ 93

Figure 4.7 Plot of Residuals vs. Projected Values for Arizona’s Two-year Projection Model (School Level) .................................................................................................... 94

Figure 4.8 Percent Proficient by 4th Grade (Required under NCLB) ..................... 99

Figure 4.9 Percent Proficient by 6th Grade (Required under GMPP) ...................... 100

Figure 5.1 – Arizona’s Reading Proficiency Cut Scores by Grade .......................... 113

Figure 5.2 – Status and Growth .............................................................................. 127

Figure 5.3 - Percent Proficient vs. Percent Who Become Proficient (School Level)..... 128
Figure 5.4 - Percent Proficient vs. Value-Added Score (School Level) ........................................ 129

Figure B.1 Plot of Residuals vs. Projected Values for Arizona’s One-year Projection Model (Student Level) .................................................................................................................. 154

Figure B.2 Plot of Residuals vs. Projected Values for Arizona’s Two-year Projection Model (Student Level) ............................................................................................ 155

Figure B.3 Plot of Residuals vs. Projected Values for Arkansas’ One-year Projection Model (Student Level) ................................................................................................. 156

Figure B.4 Plot of Residuals vs. Projected Values for Arkansas’ Two-year Projection Model (Student Level) ................................................................................................. 157

Figure B.5 Plot of Residuals vs. Projected Values for Florida’s One-year Projection Model (Student Level) ....................................................................................................... 158

Figure B.6 Plot of Residuals vs. Projected Values for Florida’s Two-year Projection Model (Student Level) ................................................................................................. 159

Figure B.7 Plot of Residuals vs. Projected Values for North Carolina’s One-year Projection Model (Student Level) ...................................................................................... 160

Figure B.8 Plot of Residuals vs. Projected Values for North Carolina’s Two-year Projection Model (Student Level) ...................................................................................... 161

Figure B.9 Plot of Residuals vs. Projected Values for Tennessee’s One-year Projection Model (Student Level) ............................................................................................. 162

Figure B.10 Plot of Residuals vs. Projected Values for Tennessee’s Two-year Projection Model (Student Level) ............................................................................................. 163

Figure B.11 Plot of Residuals vs. Projected Values for Arizona’s One-year Projection Model (School Level) ................................................................................................. 164

Figure B.12 Plot of Residuals vs. Projected Values for Arizona’s Two-year Projection Model (School Level) ................................................................................................. 165

Figure B.13 Plot of Residuals vs. Projected Values for Arkansas’ One-year Projection Model (School Level) ................................................................................................. 166

Figure B.14 Plot of Residuals vs. Projected Values for Arkansas’ Two-year Projection Model (School Level) ................................................................................................. 167

Figure B.15 Plot of Residuals vs. Projected Values for Florida’s One-year Projection Model (School Level) ................................................................................................. 168
Figure B.16 Plot of Residuals vs. Projected Values for Florida’s Two-year Projection Model (School Level) ........................................................................................................ 169

Figure B.17 Plot of Residuals vs. Projected Values for North Carolina’s One-year Projection Model (School Level) .................................................................................. 170

Figure B.18 Plot of Residuals vs. Projected Values for North Carolina’s Two-year Projection Model (School Level) .................................................................................. 171

Figure B.19 Plot of Residuals vs. Projected Values for Tennessee’s One-year Projection Model (School Level) ....................................................................................... 172

Figure B.20 Plot of Residuals vs. Projected Values for Tennessee’s Two-year Projection Model (School Level) ....................................................................................... 173
Chapter 1: Introduction

In 2001, the United States Congress reauthorized the Elementary and Secondary Education Act of 1965, now commonly referred to as the No Child Left Behind Act (NCLB). Under the revised law, in order for states to receive Title I funding they are required to test every student annually in reading and mathematics in grades three through eight and at least once again in high school (Olson, 2004). In 2006, Title I appropriations totaled 12.7 billion dollars, making Title 1 the largest federal education program. Funds are distributed across all 50 states, Washington D.C. and Puerto Rico (U. S. Department of Education, 2006). Title 1 funding (and federal education funding in general) represents a small portion of the overall education spending which totaled 403 billion dollars in 2003-04 (Johnson, 2006). Nevertheless, even though states are not required to meet the terms of NCLB, the threat of losing Title 1 funds has been great enough such that, thus far, all states have opted to comply.

The state-wide exams required under NCLB are used in each state’s accountability system to determine whether schools make Adequate Yearly Progress (AYP). The use of exams as part of a state accountability system did not start with NCLB; in fact, prior to 2001 state accountability systems were quite common. In 1999, the majority of states (36) used the results of statewide exams to produce school report cards, and between 10 and 20 states used exam results to allocate rewards, sanctions, and/or assistance (Hess & Petrilli, 2006). While education accountability systems are not new, NCLB has ratcheted up the consequences for Title 1 funded schools. Table 1.1 (below), summarizing information from the United States Department of Education
(USDOE) website, depicts the consequences associated with failing to make AYP, which range from somewhat minor to quite severe.

<p>| Table 1.1 Consequences of Failing to Make Adequate Yearly Progress¹ |
|-----------------|----------------------------------|</p>
<table>
<thead>
<tr>
<th>Year</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No consequence</td>
</tr>
<tr>
<td>2</td>
<td>School identified as needing school improvement. School officials must receive help and technical assistance. These schools must develop a two-year plan to turn around the school. Every student in the school must be given the option to transfer to another public school in the district. Students from low-income families must be offered &quot;supplemental educational services&quot; such as free tutoring services or additional academic help for students provided outside of the regular school day. Either: Replace school staff, implement a new scientifically based curriculum, decrease management authority at the school level, extend the school day or year, appoint an outside expert to advise the school, or reorganize the school entirely.</td>
</tr>
<tr>
<td>3</td>
<td>Plan for school restructuring.</td>
</tr>
<tr>
<td>4</td>
<td>Implement alternative governance plan.</td>
</tr>
</tbody>
</table>

After a school fails to make AYP for two consecutive years, students must be given the option of public school choice, a consequence that is currently of little practical impact, since only around one percent of all eligible students take advantage of this opportunity (Hoff, 2003). However, after five years of failing to make AYP schools are required to “restructure,” a result that may seriously change the everyday operations of a school.

For schools, failing to make AYP over several years can have significant consequences. So how does a school make AYP? AYP is determined in part through the

Table of Contents

Acknowledgements ........................................................................................................ iii
Abstract ............................................................................................................................... v
List of Tables ......................................................................................................................... ix
List of Figures ......................................................................................................................... x

Chapter 1: Introduction ....................................................................................................... 1
  Research Questions ............................................................................................................. 14

Chapter 2: Background and Literature ............................................................................ 16
  The Federal Growth Model Pilot Program (GMPP) .......................................................... 16
    Florida – Simple Linear Projection .................................................................................. 21
    Arizona – Regression Adjusted Linear Projection ............................................................. 25
    North Carolina – Linear Growth after Standardizing Scale Scores and Proficiency Cut
      Scores ............................................................................................................................... 28
    Arkansas – Proportional Growth ...................................................................................... 33
    Tennessee – Longitudinal Statistical Model (EVAAS®) ..................................................... 37
  The (Potential) Added Value of Value-Added Models ....................................................... 41
  Value-Added Models and Causal Inference ........................................................................ 45

Chapter 3: Data and Methods ............................................................................................ 52
  Data ..................................................................................................................................... 52
  Methods ............................................................................................................................... 58
    Florida’s Projection Model ............................................................................................... 59
    Arizona’s Projection Model .............................................................................................. 60
    North Carolina’s Projection Model .................................................................................. 62
    Arkansas’ Projection Model ............................................................................................. 64
    Tennessee’s Projection Model (EVAAS®) ......................................................................... 66

Chapter 4: Analyses and Results ....................................................................................... 71
  Research Question 1: How accurate are projection models at the student level? ........... 71
  Research Question 2: How accurate are projection models at the school level? ............. 87
  Research Question 3: Are growth expectations realistic? ............................................... 95
  Research Question 4: How do status, projection, and value-added measures of school
    performance compare to each other? .............................................................................. 102

Chapter 5: Discussion ....................................................................................................... 108
  Projection Model Accuracy ............................................................................................... 108
  Realistic Goals .................................................................................................................. 123
  Comparing Measures of School Performance ............................................................... 126

Chapter 6: Study Limitations ............................................................................................. 133

Chapter 7: Conclusions ..................................................................................................... 138
The percent proficient measure, defined above, is intuitive, informative, and easy to understand. Assuming a state’s standardized exam is aligned with the state’s standards, this measure provides a single number that represents the percentage of students at a school who have mastered the requisite material to be considered proficient. The safe harbor measure is also relatively transparent, essentially asserting that if a school can increase the percentage of proficient students from one year to the next, then the school must be “improving.” However, both of these measures have come under increasing levels of criticism (Aitkin & Longford, 1986; Alliance for Fair and Effective Accountability, 2004; Betts & Danenberg, 2004; Choi, Goldschmidt, & Yamashiro, 2005; Kane & Staiger, 2002; Linn, 2004; Lissitz, Doran, Schafer, & Willhoft, 2006). In the next two sections I describe the major criticisms of status and safe harbor measures. This is followed by an overview of an alternative approach to measuring school performance.

**Status (percent proficient) as a Measure of School Performance**

The status, or percent proficient measure, provides useful information about a school; however, it may tell us little about the impact the school has on the learning gains of its students (Lissitz et al., 2006). There are two major reasons for this. First, students enter schools at varying initial achievement levels, so year-end achievement results are largely influenced by the achievement levels of students at the beginning of the year. Second, factors other than school and students’ prior achievement influence current achievement.
To elaborate on the first criticism of status as a measure of school performance, imagine a middle school that boasts that 80 percent of its students passed the 6th grade math exam in 2006. At first, this may seem impressive, but suppose that 78 percent of these same students were capable of passing the 6th grade exam a year earlier, when they were in 5th grade and attending another school. Suddenly the school’s strong performance seems questionable. Perhaps these students attended a phenomenal elementary school and the middle school is ineffective. By measuring a school’s performance using a single snapshot in time, the school is credited for learning that occurred prior to students attending the school. As a result, a school whose students enter the school with high initial levels of achievement will appear to perform well even if it is relatively ineffective, and a school whose students enter the school with particularly low initial levels of achievement will appear to perform poorly even if it is relatively effective. In essence, status measures do not attempt to determine the contribution of the school to the learning gains of its students; rather status measures describe the achievement level of students at a single point in time.

Not only do status measures fail to take into account students’ prior achievement, they also attribute all achievement to the school, ignoring other contextual factors that influence student learning. While some knowledge gained is attributable to a school’s influence, some knowledge gained is likely attributable to peer-interactions, knowledge acquired at home, gains in knowledge resulting from prior schooling experiences (e.g., development of good study habits, interest in school, etc.). These and other outside influences are ignored when using status measures. (McCaffrey, Lockwood, Koretz, & Hamilton, 2003)
As a result, an accountability system that focuses solely on status measures holds schools to very different effectiveness standards. A school whose students’ initial achievement levels are low and whose students’ experiences outside school are less conducive to academic achievement is responsible for achieving the same outcomes as a school with advanced students whose experiences outside of school are highly supportive of academic achievement. Consequently, two schools that contribute precisely the same amount to the learning gains of their students likely will be rated differently.

By setting the same end of year achievement level goals for all students, the required learning gains vary greatly by student. As a result, expectations are easily attainable for some students, while they are unrealistic for others. In a policy brief regarding fixing NCLB, Linn states that NCLB’s most serious problem is that the “expectations for student achievement have been set unrealistically high” (Linn, 2005, p. 2). This is largely because “no goal can simultaneously be challenging to and achievable by all students across the entire achievement distribution” (Rothstein, Jacobsen, & Wilder, 2006, p. 2). At the aggregate level, Linn (2005) demonstrates that the goal that 100 percent of students will be proficient by 2013-2014 is unrealistic. The federal Growth Model Pilot Program, discussed in detail later in this proposal, attempts to address the unrealistic goals of NCLB at the student level.

While the above points are worthy criticisms of the current status measure used under NCLB, use of this measure is not without a theory of action. It is used in part, as George W. Bush repeated during his 2000 presidential campaign, to combat the “soft bigotry of low expectations” (Hess & Petrilli, 2006, p. 22). Schools are not allowed to use the initial achievement levels of their students as an “excuse” for non-proficient
performance. While the expected learning gains may not be the same for all student groups, NCLB expects and demands that all students achieve a certain standard (Linn, 2004; Rigney & Martineau, 2006).

"Safe Harbor" as a Measure of School Performance

In contrast with the status measure, NCLB’s safe harbor provision examines school-level achievement scores at two time points rather than one. By comparing the percentage of students in a school who are proficient this year to the percentage of students who were proficient last year, safe harbor attempts to assess “annual improvement for schools” (NCLB, 2001a). A theoretical definition of what safe harbor (and “improvement”) intends to capture cannot be found in NCLB. While the Act does clearly define the measure of school improvement (the change in percent proficient from last year to this year) as well as a minimum standard (a reduction of non-proficient students by at least 10 percent), from the Act alone it is difficult to decipher what “improvement” means conceptually.

Safe harbor’s assessment of school “improvement” has been criticized for a variety of reasons. I focus here on four central critiques:

1) Safe harbor compares different cohorts of students to each other,

2) Safe harbor uses aggregate school-level data rather than individual-level data,

3) Safe harbor’s measure is statistically unreliable, and

4) Safe harbor focuses on the proficiency cut score threshold.

---

1 Some states use more than two time points to calculate safe harbor. Doing so should improve the reliability of the safe harbor measure.
First, NCLB's safe harbor provision has been criticized for comparing different cohorts of students. Within any given school safe harbor compares the percentage of proficient students in the current year to the percentage of proficient students in the previous year. In a typical K-5 elementary school, students in grades three through five are tested and a certain percentage of them are deemed proficient. The previous year’s 5th graders are no longer in the school, and the current year’s 3rd graders were not counted towards the school’s percent proficient the previous year. As a result, in a school with $x$ tested grades, a minimum of approximately $1/x$ of the tested students are from a cohort that was not tested for accountability purposes in that school the previous year.

Comparing aggregate achievement scores across different cohorts of students can be problematic because of variation among cohorts of students (Linn, 2004). This variation results from natural sampling variability, changes in student demographics over time, inter-school student mobility, retention in grade, etc. Changes in the percent proficient from one year to the next may not reflect changes in practices at the school level, but rather changes in the characteristics of the enrolled students.

These problems would not be resolved by simply tracking the same cohort of students over time. Not only are different cohorts of students dissimilar, but within a school even the same cohort of students is not the same over time. That is, within a given school the current 4th graders may be a very different set of individuals than the 3rd graders from the previous year. This occurs because of student mobility and grade retention. In schools where large numbers of students enter and leave, whether during the school year or between school years, making aggregate comparisons from year to year compares different groups of students to each other. Likewise, if a student is retained in
grade, then she switches cohorts. In general, changes in aggregate school-level scores may reflect differences between the new set of aggregated students, not necessarily any real gains or improvement. To avoid these problems, measures of school performance must track the progress of individual students from year to year, rather than using aggregated data.

In part because different cohorts of students are different and in part because of measurement challenges, the safe harbor measure of school improvement leads to statistically unreliable conclusions (Linn, 2004). Empirical evidence reveals that “...if one were to look for signs of improvement by closely tracking changes in school-level scores from one year to the next, most of what one observed would be temporary – either due to sampling variability or some other nonpersistent cause” (Kane & Staiger, 2002, p. 97). In other words, shifts in the percentage of proficient students at a school from one year to the next are more likely to be a result of random noise than actual school improvement. Kane and Staiger (2002) find that the unreliability of estimates of school improvement are exacerbated in smaller schools, where sampling variability plays an even larger role in obscuring the assessment of school improvement. As a result, using a safe harbor type of measure, small schools will be over identified as showing dramatic improvement or decline.

Finally, the safe harbor measure (and the status measure as well) focuses on whether students score above or below a proficiency cut score. As a result, any gains or improvements that do not result in students crossing the proficiency threshold go unacknowledged. Using a threshold measure may result in schools focusing resources on borderline students, to the detriment of the lowest (and highest) performing students.
Does NCLB Intend to Compare Schools “Fairly”?

The above criticisms of NCLB’s two measures of AYP are valid critiques of these measures’ ineffectiveness at creating a “fair” accountability system for schools. However, it should be acknowledged that it may not be the intent of NCLB to create a “fair” accountability system, where schools are all held to the same effectiveness standards. The ultimate goal of NCLB is to get all students up to proficiency and to close achievement gaps. In fact, one of NCLB’s explicitly stated goals is to ensure that all children “reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments” (NCLB, 2001b). The framers of NCLB may acknowledge, yet find no problem with, the fact that the learning gains required to reach AYP vary from school to school. Rather than focusing on schools’ relative contributions to the learning gains of their students, NCLB focuses on an absolute goal. Consequently, schools whose students’ initial achievement levels are low must produce larger learning gains than schools whose students’ initial achievement levels are high. Schools are not measured against each other, but rather against a proficiency standard.

While this is the intent of the law, teachers and the public may have a different interpretation. When one school is labeled as failing and another is labeled as making AYP, it is difficult not to assume that the failing school is less effective. These comparisons are not limited to people simply misinterpreting the law’s intentions. While the absolute nature of NCLB’s objectives is clear, the inclusion of the school choice requirements for students attending Schools In Need of Improvement (SINI) implies that some schools are more effective than others and thus encourages the comparison of
Schools. School choice is advocated based on the assumption that it will create a more competitive, market-driven school system (Chubb & Moe, 1986; Colvin, 2004; Hess & Petrilli, 2006), and competition is inherently relative. If a market based system relies on survival of the fittest, then the measure of fitness needs to be accurate and relative. Under the current system a student might leave a school that contributes a lot to the learning gains of its students, but fails to make AYP, in order to attend a school that is contributing relatively little to the learning gains of its students, but is able to make AYP because of its students’ initially high achievement levels. Markets are most effective when information is clear and accurate (i.e. the assumption of perfect, full, information). Comparisons of schools based on current AYP measures would not lead to an optimal equilibrium. Regardless of intention or interpretation, the AYP measures used under NCLB are not an accurate way of measuring a school’s contribution to the learning gains of its students.

**An Alternative Approach**

Such critiques have led to the desire for alternative measures of school performance. The leading option is a class of models known as growth models, which measure change in individual students’ achievement over time. In 2005, the federal government announced the commencement of the Federal Growth Model Pilot Program (GMPP) that would allow states to apply to pilot growth models as part of their school accountability systems (U. S. Department of Education, 2005). Participating states use growth models to give schools “credit” for students who have made sufficient learning gains such that they are on track to be proficient in the near future. By allowing states to
use growth models that track individual students over time, rather than school-level aggregated data, it is hoped that schools will receive credit for the progress they are making.

The terms “growth model” and “value-added model” are often used interchangeably (for an example see U. S. Department of Education (2008b)). I wish to clarify their definitions and introduce a definition of a special class of growth models called “projection models.” In this research I focus on individual-level growth models (as opposed to school-level growth models), so unless otherwise indicated, the term growth model will be used to refer to individual-level growth models. Borrowing from Lissitz, Doran, Schafer, and Willhoft (2006) growth models refer to the entire class of models that utilize longitudinal data to track individual student’s achievement over time. This broad categorization allows for the inclusion of models designed for many purposes, including measuring the academic progress of individual students over time, making projections regarding students’ future exam scores based on their past learning gains, measuring the impacts of teachers or schools on student achievement using longitudinal data, etc.

Under this broad definition, projection models are a subset of growth models. Typically, projection models utilize historical achievement data for the specific purpose of predicting (or “projecting”) students’ unknown future achievement scores and/or proficiency status (Wright et al., 2006). Projection models can be used to assess whether students have made sufficient learning gains in the past such that they appear to be on track to be proficient in the near future. Using this measure, schools can be given credit for those students who have not yet achieved proficiency, but appear to be on track to
become proficient. This measure can also be used to identify students who are currently proficient, but based on their limited learning gains appear unlikely to remain proficient. Projection models have become increasingly relevant in the national education policy arena with the commencement of the federal GMPP. Under the GMPP, as of the 2007-2008 school year, seven states are piloting a variety of different projection models for use in their accountability systems. Of note is the fact that projection models do not attempt to measure schools' effectiveness relative to other schools' effectiveness.

Like projection models, value-added models are a subset, or a specific type of growth model. Value-added models refer to those growth models that attempt to measure teachers' or schools' relative 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). Value-added models are those growth models whose purpose is to attempt to measure the causal impact of teachers or schools on the learning gains of their students (Raudenbush, 2004).

In this research I examine the application of growth models, both projection models and value-added models, to the measurement of school performance in education accountability systems. Specifically, I empirically assess how accurately the GMPP’s projection models do what they intend to do, which is to forecast future proficiency. In addition, I empirically compare different measures of school performance: the NCLB status model, the GMPP projection model, and the value-added model. These analyses serve several practical purposes. They help us know how much confidence we should have in the projection models used under the GMPP. They also inform us how different
the new GMPP measures are from the old NCLB status measure, and how different value-added models are from both the old NCLB status measure and the new GMPP measures. The answers to these questions have significant implications for the design of future education accountability systems.

Research Questions

The following general research questions are addressed in this work:

1. How accurate are states’ projection models at forecasting whether individual students will become proficient at a set point in the future? Do certain state-proposed projection models forecast future proficiency more accurately than others?

2. How accurately do states’ projection models forecast the percentage of students who will become proficient at the school level?

3. For the lowest performing students, are growth expectations realistic under a status model (NCLB)? Under a projection model (GMPP)?

4. How similar or different are assessments of schools’ performance under a status model, a projection model, and a value-added model?

In order to answer these questions, I have organized this dissertation into seven chapters. In Chapter 2, Background and Literature, I describe the recently enacted GMPP, explain the projection models used under the GMPP, and detail how value-added models measure school performance. In Chapter 3, Data and Methods, I provide a full description of the data used in this study and mathematically describe the five unique state projection models being piloted as part of the GMPP as of 2007-2008. Chapter 4,
Analyses and Results, explains the analyses used to answer the research questions and introduces the results of these analyses. In Chapter 5, Discussion, I detail the implications of the analyses and suggest ways to improve the projection models. In Chapter 6, Study Limitations, I identify and acknowledge the limitations of this research, and the impact of these limitations on potential uses of this research. Finally, in Chapter 7, Conclusions, I summarize the relevance of this research from a broader policy perspective. In this final chapter, I discuss the implications of my findings for policymakers, giving special attention to the design of future accountability systems.
Chapter 2: Background and Literature

This chapter begins by providing an overview of the Federal Growth Model Pilot Program (GMPP) and the projection models used to measure school performance under the program. Next, it summarizes five states’ proposed projection models currently used under the GMPP. The chapter then transitions to discuss another method for measuring school performance, value-added models, describing their purpose and reviewing the literature on the causal inferences one might try to make when utilizing value-added models.

The Federal Growth Model Pilot Program (GMPP)

In response to requests by educators and policymakers that states be allowed to use growth models to recognize the progress schools are making, in 2005 the U.S. Department of Education announced a plan to allow states to submit proposals to pilot growth models as part of their state accountability systems (U. S. Department of Education, 2005). Proposals were required to maintain the basic tenets of NCLB. To ensure these tenets were upheld, the federal government designated seven core principles (called the “bright-line principles”) which had to be met in order for a state’s growth model to be approved. These principles (verbatim) are as follows:

1. Ensure that all students are proficient by 2014 and set annual state goals to ensure that the achievement gap is closing for all groups of students;

2. Set expectations for annual achievement based upon meeting grade-level proficiency and not upon student background or school characteristics;
3. Hold schools accountable for student achievement in reading/language arts and mathematics;

4. Ensure that all students in tested grades are included in the assessment and accountability system, hold schools and districts accountable for the performance of each student subgroup, and include all schools and districts;

5. Include assessments, in each of grades 3 through 8 and high school, in both reading/language arts and mathematics that have been operational for more than one year and have received approval through the NCLB standards and assessment review process for the 2005-06 school year. The assessment system must also produce comparable results from grade to grade and year to year;

6. Track student progress as part of the state data system; and

7. Include student participation rates and student achievement as separate academic indicators in the state accountability system.

(U. S. Department of Education, 2007)

The most notable of these seven core principles is that the models must ensure that all students are proficient by 2013-14. This requirement aligns well with the initial intent of NCLB, which was to bring all students in the nation up to proficiency. As a result of this principle, in a school where students were not proficient last year, one year of student progress is not sufficient for one year of instruction, since some students will remain below proficient (U. S. Department of Education, 2005). Also as a result of this core principle, value-added models (discussed in section “The (Potential) Added Value of Value-Added Models”) are not allowed under the pilot program. Growth models must measure growth with respect to the proficiency standards, generally asking the question
“are students on track to become proficient in the near future?” For example, as explained in Tennessee’s GMPP proposal,

Of Tennessee’s two growth models – a value-added model that estimates district, school, and teacher effect scores and a projection model that estimates individual students’ projected scores on future assessments – only one is appropriate for the NCLB growth model pilot program. The value-added model, which measures whether districts, schools, and teachers provide sufficient instruction for their students as a group to make one year of progress each year, is an innovative mechanism to drive academic progress for all students but is clearly not aligned with NCLB’s precise goal that each individual student will reach proficiency. The projection model, meanwhile, by predicting each student’s future achievement relative to state standards, holds great promise as a mechanism to guide education policy and practice under NCLB (Tennessee Department of Education, 2006, p. 2).

The other core principles of the GMPP are similarly aligned with the accountability model established by NCLB: separate accountability decisions should be made for math and language arts, all students in the tested grades should be included in the analyses, and schools are accountable for the performance of subgroups (i.e. racial subgroups, English language learners, socio-economic status, etc.).

Projection Models

Since the 2005 announcement of the federal GMPP eight states have had their proposals accepted (North Carolina, Tennessee, Delaware, Arkansas, Florida, Ohio, Ohio,
Alaska, and Arizona). One proposal (Delaware) measures growth using value tables where schools are given some form of extra credit for students who move from a lower non-proficient level to a higher non-proficient level in achievement categorization (e.g. from below basic to basic) (Delaware Department of Education, 2006). The remaining seven states utilize projection models, where they attempt to assess, based on an individual’s growth, whether she is on track to become proficient at a specified time point in the future. The seven states use six different projection methodologies (Tennessee and Ohio use the same method) varying from Florida’s simple linear trajectory to Tennessee’s more complicated statistical model (Alaska Department of Education & Early Development, 2006; Arizona Department of Education, 2007b; Arkansas Department of Education, 2006; Florida Department Of Education, 2006; North Carolina Department of Education, 2006; Ohio Department of Education, 2006; Tennessee Department of Education, 2006).

While the methods for calculating projections vary by state, the overall policy implementation is fairly similar across states. Generally, states first assess whether a school makes AYP using the old status and safe harbor measures. If a school does not make AYP using either of these measures, then growth is calculated as a third way for a school to make AYP. As such, a school cannot fare any worse under the new system since growth is only examined if a school fails to meet the status and safe harbor provisions. The growth component works as follows: the state projects whether each student has made sufficient growth such that she appears to be on track to become

---

5 As of February 5, 2008.
6 Florida and Alaska use very similar methods, although they are not precisely the same. Because their methodology is so similar, I do not examine Alaska’s method.
proficient several years into the future. Schools are then given “credit” for those students who are on track to become proficient. The same rules that apply to the AYP status measure are then applied to the percentage of students who are on track to become proficient. That is, if the status model for AYP required 54 percent of students to be proficient on the mathematics exam in 2006, then the growth model requires that 54 percent of students be on track to be proficient. Generally, under the GMPP students are given 3-4 years to achieve proficiency, after which time they must actually be proficient in order for a school to receive credit for them. Since growth cannot be assessed for students who are taking a state exam for the first time, no projection is made for such students. For a school to receive credit for a student taking an exam for the first time in that state, the student must currently be proficient, i.e. the GMPP does not apply to them.

As noted earlier, the major difference between growth model proposals is that the projection methodologies differ by state. Each state uses a different method to determine whether a student is on track to be proficient. Table 2.1 (below) provides an overview of each unique projection model. Descriptions of Ohio’s and Alaska’s models are not included since their models are very similar to Tennessee’s and Florida’s models respectively.

---

7 For the growth component, some states (Tennessee and Ohio) only give schools credit for those students who are on track to be proficient. Other states give schools credit for those students who currently are proficient plus those students who are not currently proficient but are on track to become proficient. This distinction is important because under a projection model a student who is currently proficient may not be on track to remain proficient in the future.
Table 2.1 Projection Model Descriptions

<table>
<thead>
<tr>
<th>State</th>
<th>Model Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Florida</td>
<td>Sets linear growth targets</td>
</tr>
<tr>
<td>Arizona</td>
<td>Sets linear growth targets, attempts to account for regression toward the mean</td>
</tr>
<tr>
<td>North Carolina</td>
<td>Sets linear growth targets, uses standardized scores (like z-scores)</td>
</tr>
<tr>
<td>Arkansas</td>
<td>Sets growth targets which mimic the shape of proficiency cut scores</td>
</tr>
<tr>
<td>Tennessee</td>
<td>Complex statistical model, mimics the shape of previous cohort’s growth trajectories</td>
</tr>
</tbody>
</table>

The five described models are organized from the simplest (Florida) to the most complicated (Tennessee). The first four models are all similar in that they set yearly growth targets that a student must exceed in order to be deemed on track to be proficient at a pre-specified point in the future. Tennessee’s model observes a student’s achievement scores and then projects a future score; a student is then labeled on track to be proficient if her projected score exceeds the proficiency cut score in the projected year. Found below is a more detailed description of each state’s proposed model. A technical (mathematical) description of the implementation of these models is included in the Methods section of Chapter 3: Data and Methods.

*Florida – Simple Linear Projection*

Under Florida’s projection model, schools receive credit for those students who are on track to become proficient. A student is considered on track to become proficient if, from her grade of first enrollment (or her first instate exam), her average score gain is
great enough such that if she continues to make the same gains in the future she will become proficient within 3 years (Florida Department Of Education, 2006).

Table 2.2 provides an example of Florida’s calculation for a hypothetical student. In grade three this hypothetical student scored 1170 on the state mathematics exam, which is below the required proficiency cut score of 1269. Since no prior data are available for a projection, the status measure is used and this student is deemed not to be on track to become proficient (i.e. the school receives no credit for this student). In 4th grade, this student’s scale score (1370) remains below the proficiency cut score (1444). Under the status model she is not proficient and the school does not receive credit for her. According to Florida’s projection model, she is on track to become proficient and thus the school would receive credit for her. This determination can be understood in two ways, both of which will always yield identical results with respect to projected proficiency.

<table>
<thead>
<tr>
<th>Table 2.2 Florida's Growth Model Applied to a Hypothetical Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothetical Student’s Mathematics Scale Score</td>
</tr>
<tr>
<td>Required Scale Score for Proficiency</td>
</tr>
<tr>
<td>Growth Target</td>
</tr>
<tr>
<td>Is student on track to become proficient?</td>
</tr>
</tbody>
</table>

First (as demonstrated in Table 2.2, and as described by the state), one can calculate how many scale score points this student needed to gain in order to become proficient by 6th grade. Given that the 6th grade proficiency cut score is 1692, the hypothetical student needed to gain 522 points (1692–1170) over 3 years in order to become proficient by 6th grade. Therefore, under Florida’s model, which assumes linear
growth (i.e., the same gain each year), this student had to gain an average of 174 points \((522/3)\) per year in order to be labeled on track to become proficient by 6th grade. As such, this student needed to score at least 1344 \((1170+174)\) in 4th grade in order to be on track to become proficient by 6th grade. This score, 1344, can be thought of as the student’s 4th grade “growth target.” Since the hypothetical student scored above her 4th grade growth target she is labeled on track to become proficient in 4th grade. Her 5th grade growth target is 1518 \((1170+174+174)\). Figure 2.1 below provides a visual depiction of Florida’s mathematics proficiency cut scores (diamonds) and the minimum score (or growth targets) our hypothetical student needed to attain in order to be on track to become proficient, given that her 3rd grade scale score was 1170.

**Figure 2.1 Florida’s Developmental Scale Proficiency Cut Scores and on Track to Become Proficient Growth Targets for a Hypothetical Student**

![Florida Growth Model (Math)](image)

Alternatively the calculation described in the last paragraph can be thought of as a true “projection” model, used to answer the following question: Given the student’s average gain score(s), is her projected 6th grade score at or above proficient? In the
hypothetical example, the student’s 3rd grade score was 1170, and her 4th grade score was 1370. Accordingly, her average yearly gain score was $200 \left( \frac{1370 - 1170}{1} \right)$ and her linearly projected 6th grade score is 1770 ($1370 + 200 \times 2$). This projected score assumes the hypothetical student will maintain the same rate of gains she demonstrated from 3rd to 4th grade, over the following two years. Since 1770 is above the required 6th grade proficiency cut score (1692), the hypothetical student would be labeled on track to become proficient. In 5th grade she scored 1524, so her updated average yearly gain score is $177 \left( \frac{1524 - 1170}{2} \right)$, yielding a projected 6th grade score of 1701 ($1524 + 177 \times 1$). Since 1701 exceeds the 6th grade proficiency cut score (1692) she is labeled on track to become proficient. Both calculations will always yield identical results with respect to whether students are on track to become proficient. The first description represents the methodology as described and implemented in Florida. The second description applies the underlying assumption of Florida’s model, linear growth, in order to obtain projected 6th grade scores and projected 6th grade proficiency status. Again, both approaches will always yield precisely the same results with respect to projected proficiency status. The first approach is used in the majority of the analyses in this dissertation; the second approach is used in one section of the analyses as a way to unpack why the models are as accurate (or inaccurate) as they are.

At its core Florida’s growth model allows schools 3 years to gradually bring new students up to proficiency. Of considerable note is the fact that Florida’s approach assumes a linear growth trajectory, an assumption that may not be appropriate when a developmental scale is curvilinear. Given that Florida’s mathematics proficiency cut
scores demonstrate generally decreasing gains across grades (see Figure 2.1), the assumption of linear growth is questionable.

Arizona – Regression Adjusted Linear Projection

Arizona’s GMPP projection model is a variant of Florida’s linear projection model. Much like Florida, Arizona sets linear growth targets. However, unlike Florida, Arizona does not compare a student’s observed score to the growth target in order to determine whether she is on track to become proficient. Instead, Arizona compares the lower bound of a student’s predicted score, obtained from a 95% prediction interval based upon a regression model, to the linear growth target in order to determine whether she is on track to become proficient. According to Arizona’s proposal, their regression approach is intended to adjust for regression to the mean and correct for measurement error associated with gain scores (Arizona Department of Education, 2007a). In addition, Arizona updates growth targets each year based on students’ most recent test score. For example, as is the case with Florida’s model, a student’s 4th grade growth target is a function of her 3rd grade test score and the 6th grade proficiency score. However, unlike Florida’s model, a student’s 5th grade growth target is a function of her 4th grade test score and the 6th grade proficiency score, not her 3rd grade test score.

Table 2.3 provides an example of Arizona’s calculation applied to a hypothetical student, partially borrowed from Arizona’s proposal (Arizona Department of Education, 2007b). In grade three this hypothetical student scored 362 on Arizona’s mathematics assessment, which is below Arizona’s required proficiency cut score of 420. Since no prior data are available for setting a 3rd grade growth target, the status measure is used
and the school does not receive credit for this student. In 4th grade this student’s scale score (447) remained below the required scale score for proficiency (448), but she is labeled on track to become proficient. She is deemed on track to become proficient not because her observed 4th grade score was above the growth target (although it happened to be), but because the lower bound of her predicted 4th grade score (409) was above the 4th grade growth target (407).

Table 2.3 Arizona’s Growth Model Applied to a Hypothetical Student

<table>
<thead>
<tr>
<th>GRADE</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothetical Student’s Mathematics Scale Score</td>
<td>362</td>
<td>447</td>
<td>480</td>
<td>503</td>
</tr>
<tr>
<td>Required Scale Score for Proficiency</td>
<td>420</td>
<td>448</td>
<td>476</td>
<td>496</td>
</tr>
<tr>
<td>Predicted Score (Standard Error)</td>
<td>NA</td>
<td>417</td>
<td>474</td>
<td>NA</td>
</tr>
<tr>
<td>Lower Bound of Predicted Score (Predicted score – 1.96*(Standard Error))</td>
<td>NA</td>
<td>409</td>
<td>465</td>
<td>NA</td>
</tr>
<tr>
<td>Growth Target</td>
<td>NA</td>
<td>407</td>
<td>472</td>
<td>496</td>
</tr>
<tr>
<td>Is student on track to become proficient?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

As noted, the 4th grade growth target is set based on the same linear growth assumption implemented in Florida’s model. Since the 6th grade proficiency cut score (in Arizona) is 496 and this student’s 3rd grade scale score was 362, this student needed to gain 134 points over the next three years. Arizona’s linear model requires this student to gain one third of this amount in the first year, or 45 points, resulting in a 4th grade growth target of 407 (362+45). If the lower bound of the hypothetical student’s predicted score is above 407 she is on track to become proficient.

The student’s predicted 4th grade score is calculated using a regression model with 3rd grade test scores and school fixed effects as predictor variables, and 4th grade test scores as the dependent variable. In the example in Table 2.3 the lower bound of our
hypothetical student’s 4th grade predicted score was 409 (417-1.96*4.27), which is above the growth target (407), thus this student is track to become proficient. For a more detailed description of the regression model and how the lower bound of the predicted score is calculated, see Chapter 3: Data and Methods.

Unlike Florida’s model where the 5th grade growth target is set based on a student’s 3rd grade score and the 6th grade proficiency cut score, Arizona’s model sets the 5th grade growth target based on a student’s 4th grade score and the 6th grade proficiency cut score. As such, since our hypothetical student scored 447 in 4th grade, she needs to gain 49 points to reach proficiency by 6th grade. Since she has 2 years to gain 49 points, she must gain 25 points each year, meaning her 5th grade growth target was 472 (447+25).

The hypothetical student’s predicted 5th grade score is calculated using a regression model with 4th grade test scores and school fixed effects as predictor variables, and 5th grade achievement as the dependent variable. In the example in Table 2.3, the lower bound of our hypothetical student’s 5th grade predicted score was 465 (474-1.96*4.78), which is below the growth target (472), thus this student is not on track to become proficient. Notably, this student’s observed 5th grade score (480) was above the growth target, yet she is still not on track to become proficient because the lower bound of her predicted 5th grade score was below the growth target. Again, using this predicted score rather than the observed score is intended to account for regression toward the mean and measurement error in gain scores.

Arizona’s growth model allows schools three years to gradually bring new students up to proficiency. Like Florida, Arizona sets growth targets which assume linear
growth on the developmental scale, an assumption that is not appropriate when a
developmental scale is curvilinear. Much like Florida's mathematics proficiency cut
scores, Arizona's mathematics and reading proficiency cut scores are not linear, so the
linear assumption may not be tenable in Arizona.

Unlike Florida's model, Arizona's regression approach attempts to account for
regression toward the mean. However, their approach does not fully utilize the
information regarding any individual student's test history. That is, while a student's 3rd
grade test score is used to calculate her 4th grade predicted score, her 4th grade test score
is not directly used to calculate her 4th grade predicted score. As such, two students
from the same school who score the same in 3rd grade, but vastly different in 4th grade,
will both have the same predicted 4th grade score. Consequently, while these two
students may have made very different gains, they will both be identified as "on track" or
not "on track," regardless of the differences in their gain scores.

*North Carolina – Linear Growth after Standardizing Scale Scores and Proficiency Cut Scores*

Like Florida's and Arizona's projection models, North Carolina's model uses a
linear projection approach. Unlike other models, North Carolina's growth model does
not require a vertically equated exam, but instead takes a "Standardized Scale Approach
(SSA)." In the standard setting year of North Carolina's state exam, they calculate the
mean scale score and the standard deviation of the scale scores for each grade and
subject. Using this information they convert the proficiency cut scores from the regular

---

8 It is indirectly used since the student's score is one observation from the entire population of observations used to calculate the regression model's coefficients.
scale to one on a “time-locked modified z-scale,” referred to as the C-scale, or change scale. They achieve this by subtracting the state-wide grade/subject mean from the proficiency cut score (or individual students’ score), and then they divide the difference by the state-wide grade/subject standard deviation (North Carolina Department of Education, 2006). This same calculation is computed for individual students to determine their score on the C-scale.9

For example, in order to be considered proficient on the North Carolina 7th grade reading exam a student needs to score at or above the proficiency cut score of 252 points. 2003 was the standard setting year for North Carolina’s Reading exam. In that year, the mean 7th grade scale score was 261.1 with a standard deviation of 9.06. By standardizing the 252 point proficiency cut score one can determine that, on the C-scale, a student needs to score at or above a -1.00 \((252 - 261.1)/9.06\) in order to be proficient. The top line in Figure 2.2 below represents the C-scale minimum proficiency cut scores for grades 3-8 on North Carolina’s math exam. Since the C-Scale is a grade specific standardized scale, the proficiency scores are not monotonically increasing even though achieving at the proficiency level requires knowing more as the grade level increases.

---

9 Most people would think of this as a z-score. North Carolina may have chosen to distinguish their measure, calling it a C-scale, because the mean and standard deviation used to standardize scores always come from the standard setting year, not the current cohort of students.
With the proficiency cut scores converted to the C-scale, student scores are also converted to the C-scale using the same mean and standard deviation from the standard setting year. Once a student’s scores are converted to the C-scale growth targets are calculated in the same manner as in Florida. Table 2.4 provides an example of this calculation for a hypothetical student.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Pre-Test</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothetical Student’s Mathematics C-Scale Score</td>
<td>-1.96</td>
<td>-1.65</td>
<td>-1.48</td>
<td>-1.1</td>
<td>-0.88</td>
</tr>
<tr>
<td>Required C-Scale Score for Proficiency</td>
<td>-0.75</td>
<td>-0.59</td>
<td>-1.06</td>
<td>-1.04</td>
<td>-0.92</td>
</tr>
<tr>
<td>Growth Target on C-Scale</td>
<td>NA</td>
<td>-1.70</td>
<td>-1.44</td>
<td>-1.18</td>
<td>-0.92</td>
</tr>
<tr>
<td>Is student on track to become proficient?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

On the 3rd grade pretest our hypothetical student’s C-scale score was equal to -1.96, which is below the C-scale pretest proficiency cut score of -0.75. This student was not proficient on the 3rd grade pretest. Since North Carolina has a 3rd grade pretest (unlike the other states discussed), they are able to establish growth targets for students
by the 3rd grade end of year assessment. On the 3rd grade end of year assessment this student’s C-scale score remained below the required score for proficiency, but she was on track to become proficient. One can calculate how many C-scale score points a student needs to gain in order to become proficient by 6th grade by subtracting the 6th grade C-scale cut score (-0.92) from the student’s 3rd grade pretest C-scale score. In the example our hypothetical student needed to gain 1.04 points (-1.96 - [-0.92]) over the course of 4 years. As such, using the linear model she needed to gain an average of .26 C-scale points (1.04/4) per year in order to be on track to become proficient by the 6th grade. Therefore, her 3rd grade growth target was -1.70 (-1.96+.26) in order to be on track to become proficient by 6th grade. Her 4th grade growth target was -1.44 (-1.70+.26) in order to be on track to become proficient, etc. Figure 2.2 above provides a visual depiction of North Carolina’s mathematics C-scale proficiency cut scores and the growth targets, or minimum scores our hypothetical student needed to attain in order to be on track to become proficient, given that her 3rd grade pretest C-scale score was -1.96.

Whereas Florida’s and Arizona’s projection models assume linear growth on their developmental scale, North Carolina’s projection model assumes linear growth on the C-scale, a standardized scale. Under the Florida model, given the proficiency cut score and a student’s initial scale score one can calculate growth targets. In contrast, North Carolina uses proficiency cut scores, the student’s initial scale score, plus actual observed data of students from the standard setting year. This difference highlights a fundamentally different approach to calculating growth targets. Florida relies more heavily on the assessment instrument and properties of the assessment metric (e.g. vertical equating, the standard setting procedure) whereas North Carolina’s model
focuses more on the relationship between the relative achievement scores of students across grades in the standard setting year. If a student scores 2 standard deviations below the average student at time $t$ and she scores 1.5 standard deviations below the average student at time $t + 1$, then North Carolina’s model assumes she’ll continue to score closer to the average student at time $t + 2$. If attaining the same C-scale score from one year to the next is an approximation for one year’s worth of typical growth, then North Carolina’s model assumes that a student who makes more than one year’s worth of growth at time $t$ will continue to make more than one year’s worth of growth at time $t + 1$.

We can apply both Florida’s and North Carolina’s methodologies to a hypothetical student (the same hypothetical student as in the Florida – Simple Linear Projection section of this paper) on Florida’s developmental scale, in order to better understand the differences between these two methods (and their assumptions). Figure 2.3 below compares the growth targets under the Florida and North Carolina models. Growth targets for North Carolina’s model are set using the underlying assumptions of North Carolina and applying them to data from Florida’s standard setting year. In this case North Carolina’s model sets stricter growth targets than does Florida’s model.
Arkansas – Proportional Growth

Arkansas’ growth model also attempts to assess whether individual students are on track to become proficient at a certain point in the future. Unlike Florida, Arizona and North Carolina, Arkansas does not assume linear growth, but rather assumes “proportional growth.” In order for a student to be labeled on track to become proficient she must close the gap between her current score and the future requisite proficiency cut score by a rate that is proportional to the change in the proficiency cut scores over that time period (Arkansas Department of Education, 2006).

Table 2.5 below provides an example that should help clarify this method.

---

10 Arkansas model makes projections 4 years into the future. For the sake of a more consistent comparison of projection methodologies, in my example I use a three year projection for Arkansas.
Table 2.5 Arkansas’ Growth Model Applied to a Hypothetical Student

<table>
<thead>
<tr>
<th>Hypothetical Student’s Mathematics Scale Score</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required Scale Score for Proficiency</td>
<td>471</td>
<td>535</td>
<td>589</td>
<td>636</td>
</tr>
<tr>
<td>Proportion of 3 year gain required to be on track to become proficient</td>
<td>500</td>
<td>559</td>
<td>604</td>
<td>641</td>
</tr>
<tr>
<td>Cut score needed to be on track to become proficient</td>
<td>NA</td>
<td>42%</td>
<td>32%</td>
<td>26%</td>
</tr>
<tr>
<td>Is student on track to become proficient?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Consider Arkansas’ mathematics proficiency cut scores in grades 3, 4, 5, and 6 (500, 559, 604, and 641 respectively). Between 3rd and 6th grade the proficiency cut score increases by 141 (641 – 500) points, 59 (559 – 500) of which occur between 3rd and 4th grade. As such, 42 percent (59/141) of the total increase in proficiency cut score from 3rd and 6th grade occurs between 3rd and 4th grade. Likewise, 32 percent occurs between 4th and 5th grade, and 26 percent occurs between 5th and 6th grade. Notice that the proficiency cut scores do not increase linearly, and as a result neither do the growth targets. In order to determine whether a student is on track to become proficient by the 6th grade, Arkansas calculates how many total points an individual student needs to gain from 3rd grade to 6th grade and then requires that she gain 42 percent of those points in 4th grade, 32 percent in 5th grade and 26 percent in 6th grade.

Table 2.5 above demonstrates how the growth targets are set for a hypothetical student in Arkansas. Our hypothetical student scored 471 on her 3rd grade math exam, so she needed to gain a total of 170 (641-471) points in order to become proficient on Arkansas’ 6th grade exam. According to Arkansas’ model, from 3rd to 4th grade she needed to make up 42 percent of those 170 points, or 71 (0.42 * 170) points in order to be on track to become proficient by 6th grade. As such, her 4th grade growth target is 542.
(471+71) in order to be deemed on track to proficiency. Between 4th and 5th grade she needed to make up 32 percent of the 170 points, or 55 points. So her 5th grade growth target is 597 (471+71+55) in order to remain on track to become proficient. Figure 2.4 below provides a visual depiction of how this plays out for our hypothetical student.

**Figure 2.4 Arkansas’ Developmental Scale Proficiency Cut Scores and on Track to Become Proficient Growth Targets for a Hypothetical Student**

![Graph of Arkansas Growth Model (Math)](image)

Much like the other projection models, Arkansas’ model allows schools several years to bring students up to proficiency rather than requiring them to attain this goal immediately. In contrast to the other models discussed so far, Arkansas’ model does not assume linear gains, but instead assumes gains that are proportionately related to changes in the proficiency cut scores. Under Arkansas’ model, if typical student growth trajectories do not mimic the shape of the state’s proficiency cut scores, then the model is less likely to be accurate in its projections.

By applying both Florida and Arkansas’ methodologies to a hypothetical student (the same hypothetical student as in the Florida – Simple Linear Projection section of
this paper) on Florida’s developmental scale, we can better observe the differences between these two methods (and assumptions). Figure 2.5 below provides a visual display of the differences.

**Figure 2.5 A Comparison of Florida and Arkansas Growth Models on Florida’s Developmental Scale for a Hypothetical Student**

![Florida and Arkansas Growth Models](image)

This figure depicts how Florida’s model requires a student to attain equal gains all three years whereas Arkansas requires larger gains to be made in years one and two, followed by much smaller gains in year three. In this instance, Arkansas’ growth targets are more stringent than Florida’s growth targets. Under Florida’s growth model a student does not need to reduce the gap between her scale score and the proficiency cut score until 6th grade. In grade 3 our hypothetical student was 100 points below proficient. Because of the assumption of linear gains, according to Florida’s growth model (Arizona’s model makes the same assumption) this student could actually remain as many as 114 points below proficient in 5th grade and still exceed the 5th grade growth target and be deemed on track to become proficient by 6th grade. In contrast, Arkansas’ model requires that our
hypothetical student reduce the difference between her scale score and the proficiency cut score each year. The hypothetical student must reduce the 100 point deficit to 59 points in 4th grade and 14 points in 5th grade in order to be on track to become proficient by 6th grade. Which model will more accurately project whether students will actually become proficient in 6th grade depends on what the typical growth trajectories are on the state’s developmental scale. Whereas Florida’s growth model relies heavily on the assumption that growth is linear, Arkansas’ growth model is dependent on the standard setting procedure used to determine the proficiency cut scores.

Tennessee – Longitudinal Statistical Model (EVAAS®)

Tennessee’s projection model utilizes a student’s past scores as well as historical patterns of other students’ achievement scores to project whether she will become proficient at a set point in the future (Tennessee Department of Education, 2006). Wright, Sanders, and Rivers (2006) describe the model using equation (2-1):

\[
(2-1) \text{Projected Score} = M_Y + b_1 (X_1 - M_1) + b_2 (X_2 - M_2) + ... = M_Y + X^T b
\]

\(M_Y, M_i,\) etc. are estimated mean scores for the response variable (\(Y\)) and the predictor variables (\(X_i\)). Much like the statistical method used in value-added modeling (described later in The (Potential) Added Value of Value-Added Models), Tennessee’s Educational Value-Added Assessment System (EVAAS) projection model uses information from the covariance matrix of all the predictors and the response variable in making projections. In 4th grade, their model uses a student’s 3rd and 4th grade scores to project a 6th grade score. If this projected score is above the 6th grade proficiency cut score, then the student is deemed on track to become proficient. In 5th grade, Tennessee’s
model uses a student’s 3rd, 4th and 5th grade scores to project a 6th grade score. If this projected score is above proficiency, then the student is deemed on track to become proficient.

Wright et al. (2006) describe several advantageous features of their projection model. First, much like North Carolina’s model, test scores do not need to be vertically linked (i.e. on the same scale) from grade to grade in order for this approach to be used, a feature which makes both these approaches very versatile. Second, EVAAS makes no assumptions regarding the overall shape of the growth curve. While harnessing information from the covariance matrix assumes linear relationships between pairs of scores, it makes no assumption about the overall shape of the growth curve.

Wright et al.’s (2006) paper focuses on comparing EVAAS to a Hierarchical Linear Model (HLM) when making projections; however, when comparing EVAAS to other states’ projection models two features are most salient. First, EVAAS utilizes significantly more data when making projections, and it will likely be more accurate because it does not make many unrealistic assumptions (e.g. – does not assume linear growth, does not rely on cut scores to set growth targets). While North Carolina’s projections are in part based upon the mean and standard deviations of students in the standard setting year, only EVAAS takes into account the historical patterns of achievement scores of other students. Second, EVAAS is much less readily accessible to the public since the techniques used are understood by few. It is even unclear whether the methodology is exactly replicable, since the details provided in the growth model proposal and the article by Wright et al. (2006) are somewhat limited. Contact made, for the purpose of model clarification, by the author of this dissertation with the Tennessee
Department of Education (Senior Executive Director, Tennessee Department of Education, Assessment, Evaluation, and Research on 5/2/07, 5/9/07, and 6/11/07) have not yielded information necessary for perfect replication of Tennessee’s approach. While this approach may be more accurate than others, if transparency is important to an accountability system, then EVAAS leaves something to be desired. For more details on my best approximation of this projection model see Chapter 3: Data and Methods, or see Wright et al.(2006) for their description.

**Objectives of using Projection Models**

The main objective of the GMPP’s accepted projection models is to give schools credit for those students who have not yet reached the proficiency cut score, but have made sufficient growth such that they appear on track to reach the fixed proficiency target at some point in the near future. While students (and therefore schools) are still held to different learning gains standards, the time schools have to bring students up to proficiency is longer. A school whose students were initially one year behind grade level proficiency would have had to make up the entire difference in a single year under the NCLB status model. Using projection models schools have three years to make up the difference, allowing them to make approximately one and a third years worth of progress per year over the course of three years, rather than having to make two years of progress over the course of a single year. By giving students (and thus schools) several years to get up to proficiency, it is possible that the GMPP’s objectives are more realistic than learning those under NCLB’s original measures of AYP.
Figure 2.6 below provides an illustration of how this works. In this example the hypothetical student scored 100 points below proficient in 3rd grade (i.e. she scored 1169). Under a status model, by 4th grade this student must reach proficiency (i.e. she must score at least 1444), requiring a gain of 275 scale score points. Under Florida’s projection model, this student’s 4th grade score only needs to demonstrate that she “is on track to become proficient,” requiring a gain of 175. The latter goal may be more realistic than the former.

Figure 2.6 A Comparison of the Required Gains under a Status Model (NCLB) vs. a Projection Model (GMPP)

The use of projection models in the GMPP may combat two of the major criticisms of status measures; firstly, that schools will receive no credit for students who make substantial learning gains yet still do not reach proficiency, and secondly, that the status goals are unrealistic for certain initially low performing students. Still, the GMPP does not address the larger criticism of status measures, which is that they do not attempt
to compare schools to each other based on their unique contributions to the learning gains of their students. For this, we turn to “value-added” models.

**The (Potential) Added Value of Value-Added Models**

The value-added model “is a term used to label methods of assessment of school/teacher performance that measure the knowledge gained by individual students from one year to the next” (Tekwe et al., 2004, p. 12). Value-added models attempt to mitigate many of the problems associated with the status and safe harbor measures of school performance discussed earlier in this paper. Specifically, value-added models seek to control for student prior achievement levels, account for student mobility, account for retention in grade, account for the persistent effect of past teachers/schools on current performance, rely on scale scores rather than on discrete student outcomes, adjust for classroom/school size, and implicitly or explicitly control for socioeconomic status and other background characteristics generally believed to influence student achievement. Described in the next section is one popular approach to value-added modeling.

*Tennessee Value-Added Assessment System (TVAAS)*

TVAAS, the layered value-added model described by Sanders & Horn (1994), is the most well known value-added model used to measure teacher effectiveness. TVAAS utilizes longitudinal student-level achievement data, harnessing information across multiple years, multiple cohorts, and multiple subject areas. Rather than explicitly modeling current scores based on prior achievement, TVAAS treats test scores at all time points and subject areas as a multivariate outcome variable. TVAAS accounts for prior
achievement by exploiting the variance-covariance matrix through modeling intra-student correlations over time and across subject areas. In doing so, TVAAS “implicitly control(s) for socioeconomic status and other background factors that influence initial levels of achievement” (Ballou, Sanders, & Wright, 2004, p. 37). Since current achievement scores are not explicitly predicted by past achievement scores, TVAAS methodology is able to utilize all student data, regardless of how sparse or complete (Sanders & Horn, 1997).

Expressed mathematically in equation (2-2) (Lockwood, Doran, & McCaffrey, 2003) \( y_{1i}, y_{12}, \) and \( y_{13} \) represent student \( i \)'s test scores at times one, two, and three respectively. In modeling time one test scores, a year (or grade) fixed effect, \( \mu_1 \), and a teacher random effect, \( \theta_{j(i)} \), are used as independent variables. \( \mu_1 \) can be thought of as the average test score for all students at time one. \( \theta_{j(i)} \) represents the random effect of the \( j^{th} \) teacher, student \( i \)'s teacher at time one.

In modeling time two test scores, a year (or grade) fixed effect, \( \mu_2 \), and two teacher random effects, \( \theta_{j1(i)} \) and \( \theta_{j2(i)} \) are used as independent variables. \( \theta_{j1(i)} \) represents the \( j^{th} \) teacher, which is student \( i \)'s teacher at time one; \( \theta_{j2(i)} \) represents student \( i \)'s teacher at time two. By including the random effect for the time one teacher when modeling time two test scores, this value-added model attempts to account for past teachers’ influences on all future outcomes. The same pattern can be used to understand the prediction of time three test scores.
\[ y_{it} = \mu + \theta_{jt(i)} + \epsilon_{i1} \]  
\[ y_{i2} = \mu + \theta_{jt(i)} + \theta_{j2(i)} + \epsilon_{i2} \]  
\[ y_{i3} = \mu + \theta_{jt(i)} + \theta_{j2(i)} + \theta_{j3(i)} + \epsilon_{i3} \]  

Where:

\[ \epsilon_{i1} \rightarrow \text{within student residual term.} \]

\[ (y_{i1}, y_{i2}, y_{i3}) \sim N\begin{pmatrix} 0 \\ \sigma_{2,1}^2 \\ \sigma_{2,2}^2 \\ \sigma_{3,1}^2 \\ \sigma_{3,2}^2 \\ \sigma_{3,3}^2 \end{pmatrix} \]

As noted earlier, it is through the modeling of the error structure that TVAAS implicitly controls for student prior achievement. Conditions described in (2-3) explain the error structure, \( \epsilon_{i1} \). Essentially, built into this statistical model is the expectation that the residuals for student \( i \) will be related at times one, two, and three. The exact strength of these correlations is not predetermined, but instead is estimated based on the actual data. For example, \( \sigma_{2,1} \), is the covariance between students’ scores at time one and time two, estimated using the actual data. The teacher effects therefore represent influences on student achievement after accounting for the intra-student correlations.

Equation (2-2) and the conditions expressed in (2-3) are a simplified representation of the actual TVAAS model. In its actual usage, the model utilizes data across five years, multiple subject areas and multiple cohorts. It is through the usage of many repeated measures on each individual student that TVAAS attempts to capture and separate out teacher effects from exogenous influences. Notice that the model uses only student test score data in order to estimate teacher effects.
TVAAS accounts for across year mobility and retention in grade by tracking individual students over time, rather than looking at aggregate data. TVAAS is also capable of partially accounting for within-year student mobility by attributing students’ scores to their teacher based on the proportion of the school year the student spent in each classroom. Additionally, classroom size is considered in these value-added estimates. As described by Sanders, Saxton, and Horn teachers are assumed to be the average of their school system “until the weight of the data pulls their specific estimates away from their school system’s mean. A very important consequence is that it is nearly impossible for individual teachers with small quantities of student data to have estimates measurably different from their system mean” (Sanders & Horn, 1997, p. 143). This is a result of the Bayesian procedure used in estimating the teacher random effects.

When measuring teacher/school performance using a value-added model, controlling for student background characteristics is controversial. It can lead to differential learning gain expectations for different groups of students – a fact that can be politically problematic. As implemented by Tennessee, TVAAS does not explicitly control for any student or school-level covariates. However, it has been found that when using the TVAAS approach, including individual student-level predictors has little or no influence on estimated school effects (Ballou et al., 2004). Although the model is capable of including student characteristics as predictors in the model, Ballou et al.’s (2004) findings suggests that this may not be necessary. If Ballou’s findings are correct, this would significantly reduce concerns regarding the need to control for individual-level background characteristics. However, McCaffrey et al. (2003) were unable to reach the same conclusions in their work. Regardless, this does not address concerns regarding the
need to control for larger *school-level* contextual issues (e.g., the proportion of students eligible for free and/or reduced priced lunch).

The majority of research on value-added models concentrates on estimating the contributions of teachers, not schools, to the learning gains of their students. In this work, value-added models of school performance are considered. While this difference may seem inconsequential, one cannot simply replace the word “teacher” with the word “school” to extend the findings of McCaffrey et al. and Sanders et al. The application of value-added models to measure school effectiveness rather than teacher effectiveness may significantly alter the meaning of these models because “The key feature of longitudinal achievement data for modeling teacher contributions to student achievement is the sequential regrouping of students into different classrooms with different teachers. This results in data where students who are nested under a common teacher for one measurement are not nested together for another measurement” (Lockwood, McCaffrey, Mariano, & Setodji, 2007, p. 126). In contrast, at the school level most students tend to remain grouped together from year to year. Consequently, compared to value-added models of teacher effectiveness, value-added models of school effectiveness are probably far less successful at separating out the unique contribution of schools to the learning gains of their students. The significance of this difference can not be overstated.

**Value-Added Models and Causal Inference**

Raudenbush (2004) points out that a good place to begin when considering how well value-added models work in approximating the causal effects of schools is first to clearly define the possible treatments and potential outcomes under alternative
treatments. Essentially, we first need to reflect on what it is we are trying to estimate. Earlier work by Willms and Raudenbush (1989) describes two types of effects one might be interested in estimating. They refer to these as Type A and Type B effects, each referring to a different concept which could help answer distinct policy questions and therefore serve a unique purpose.

**Type A Effect**

The Type A effect is the difference between student $i$’s outcome (achievement score) at school $s$ compared to her potential outcome had she attended the “typical school.” Raudenbush believes that this effect is “arguably estimable with tolerably small bias” (Raudenbush, 2004, p. 122) in large part because the Type A effect does not attempt to decompose variation in achievement gains into portions that are attributable to school practices versus those that are attributable to school context. In other words, the Type A effects intend to measure which school(s) a student is more likely to be “successful” at; however, they do not concern themselves with whether success at these schools is due to the schools’ practices or other factors that are beyond the control of the school such as peer influence or local community factors. As such, Type A effects would not be ideal for an accountability system if the system’s goal were to compare schools’ relative effectiveness based on factors within the control of the school. However, Type A effects could be of use to parents who are trying to decide which school their child is most likely to excel at academically.
Type B Effect

The Type B effect is basically the Type A effect after holding constant the school context. That is, the Type B effect “asks how well a particular school performed relative to other schools with similar student intakes, contextual effects, and wider social influences” (Willms & Raudenbush, 1989, p. 213). It is the Type B effect that is of most interest in an accountability system that desires to hold schools/teachers accountable for their contribution to the learning gains of their students. It is this effect that is generally sought after through the use of value-added models.

Although we can estimate Type A effects under fairly reasonable assumptions, the same cannot be said for the elusive Type B effects. The challenge lies in decomposing the variation that results from school practice and that which results from school context. In 1989 Willms and Raudenbush stated that in order to accurately estimate Type B effects “one requires measures of the relevant policy and contextual variables. Although the literature provides some guidance on what variables are relevant, we feel that the processes are complex and interactive, and that theories about school effects are in their early stage of development” (Willms & Raudenbush, 1989, p. 228). How much progress have we made since this somewhat hopeful claim? Nearly 15 years after Willms and Raudenbush (1989) first articulated the Type B effect, Raudenbush stated that “the prospects for estimating Type B effects are dim at best” (Raudenbush, 2004, p. 123). This belief stems largely from the fact that we cannot adequately measure school practice in order to statistically control for its possible correlation with student-level covariates and school-level covariates. As a result, Type B treatment assignment (school practice) is largely unknown, so the selection of schools into their treatments cannot be adequately
controlled for in a statistical model. Consequently, Type B effects are usually estimated under the unlikely assumption that schools of varying context are randomly assigned to school practices (Raudenbush & Willms, 1995).

It is not only Raudenbush and Willms (1995) who have documented the significant challenges associated with estimating Type B school/teacher effects. McCaffrey et al. (2003) note that since students are not randomly assigned to schools/classrooms it is possible that students are systematically separated into better or worse schools. As a result, “Bias can occur when students attending different schools differ in ways that are likely to affect both achievement and growth in achievement, and the context of the school (e.g., the proportion of students eligible for free and reduced price lunches) affects these outcomes” (McCaffrey et al., 2003, p. xvii). McCaffrey et al.’s research found that context does affect growth and, as a result, estimates of teacher effects may mix up contextual effects with true teacher effects.

Even Ballou, Sanders, and Wright acknowledge that it is “plausible that the make-up of the school influences achievement through peer effects” and since “the covariance structure of the TVAAS does not capture the effects of student clustering (covariances across students, even those in the same class, are zero), we cannot be confident that the TVAAS controls for contextual variables in the same way that it controls for the influence of student-level SES and demographics” (Ballou et al., 2004, p. 61). Even some of the most ardent proponents of value-added modeling for accountability purposes concede that separating out the effects of school practices from other contextual variables is not something that value-added models can yet do. Since evidence suggests that contextual variables matter (for an example, see Hanushek, et al. (2003) on how peer
ability affects student achievement), this recognized limitation to value-added models is significant.

Using a potential outcomes perspective, Rubin, Stuart, and Zanutto (2004) highlight yet another difficulty in attempting to reach causal conclusions using observational data. They note that an essential component to making comparisons using observational data is to find treatment and comparison units that look as similar as possible on background characteristics. If this is not feasible, results are bound to be “based on untestable modeling assumptions and extrapolation” (Rubin et al., 2004, p. 11). Under the condition where “school A has no students who “look like” students in the other schools, it is impossible to estimate the effect of school A relative to the comparison schools without making heroic assumptions” (Rubin et al., 2004, p. 11). The problem of extrapolating beyond the data is a serious one, and may be extended further.

Value-added models rate schools as though school effectiveness is unidimensional and linear. The multidimensional, nonlinear nature of school effectiveness is most obvious as it relates to the fact that schools do more than just try to raise the achievement scores of their students (e.g. they may introduce students to the arts, promote social and emotional well-being, etc.). But even within the realm of school/teacher effectiveness as it pertains to student achievement, there is probably not a unidimensional, linear scale that can appropriately rate school effectiveness. A school/teacher that is extremely effective at educating certain types of students may be much less effective at educating other types of students. Although an underlying “effective school” construct likely exists, it is unclear how highly correlated a school’s effectiveness might be in different contexts. If this correlation is not extremely high, then
making judgments regarding school effectiveness among schools with varying contexts is difficult to justify.

This problem highlights the importance of Raudenbush’s suggestion that we clearly define exactly what experiment we are trying to approximate when using value-added models to measure school effectiveness. In measuring school $s$’s Type B effects, we are comparing the outcomes of the students who attended school $s$ with the performance that we would have expected those students to achieve had they attended a school of “average” effectiveness in an identical context. Given the likely possibility that a school’s effectiveness depends on the context, what is the appropriate set of schools to compare school $s$ to in approximating the school of “average” effectiveness? If we are trying to approximate an experiment, then school $s$ should be compared to only those schools with a similar context. This leads to fairer comparisons, but it fails to address broader concerns that more or less effective teachers/schools tend to cluster in particular contexts. The alternative is for school $s$ to be compared to the “average” school, defined as an average of all schools; however, this is where the issue of extrapolating beyond the data becomes problematic. The dilemma appears to be irresolvable. Making comparisons among schools that share a similar context is limiting, and making comparisons among schools from different contexts may not be reasonable.

This chapter outlined the broader accountability context in which various measures of school performance lie. NCLB’s original measures of school performances serve a purpose, but are frequently criticized. In order to combat some of the criticisms and to increase flexibility, the GMPP was introduced and projection models were devised as a new way to measure school performance tracking individual students longitudinally.
The projection models used under the GMPP attempt to give schools credit for the progress students are making towards proficiency. Both NCLB and the GMPP do not allow a third method to measure school performance: value-added models. Value-added models attempt to assess the relative contribution of schools to the learning gains of their students, regardless of current or future proficiency status. In the next section I discuss the secondary data source used to examine the various measures of school performance described in this chapter.
Chapter 3: Data and Methods

This chapter describes the data source used for analyses in this research and mathematically describes the five state projection models used under the GMPP.

Data

The analyses for this research use student-level vertically scaled scores from the Florida Comprehensive Assessment Test (FCAT) in mathematics. The data are from the 2001-02 through 2004-05 school years, and they are census data from a large urban school district’s public schools. The study district’s school system was one of the 20 largest school districts in the United States, serving over 129,500 students enrolled in approximately 182 schools. The ethnic composition of the district’s students in this was 46 percent white, 43 percent black, and 5 percent Hispanic, with 49 percent of students receiving free or reduced-price lunch assistance.

The fact that FCAT developmental scale scores are vertically equated is important because several of the projection models studied (Arizona’s, Arkansas’, and Florida’s models) were designed for use on vertically scaled scores. Vertical equating, the procedure that creates vertically scaled scores, “refers to the creation of a single reporting scale extending over a number of school grades” (du Toit, 2003). Theoretically, this means that developmental scale scores from the FCAT are all on the same metric both across grades and over time. The FCAT is an excellent assessment instrument for these analyses because the exam has a long history in the state and is believed to be both reliable and valid (Florida Department Of Education, 2004).
This research focuses on the 2002 cohort of 3rd grade students, tracked from 2002-2005. These grades were selected because they are the grades most likely to be affected by the GMPP. Although the GMPP credits schools for students who are not yet proficient, the GMPP projection models require that all students attain proficiency 3-4 years after their grade of first enrollment or their first instate exam. As such, the GMPP is unlikely to have much impact after 6th or 7th grade.\textsuperscript{11}

\textit{Attrition}

This study examines 10,007\textsuperscript{12} students with 3rd grade achievement scores in spring 2002. Due to attrition, achievement data are not available for all 10,007 students in the years following 2002. Table 3.1 provides frequency counts, by grade level, for the 10,007 students from spring 2002 through spring 2005. Of the 10,007 students with mathematics achievement data in 2002, 8875 or 89 percent had mathematics achievement data in 2003. In 2005, test score data were available for 7,550 or 75 percent of the original full 3rd grade sample. Inter-district mobility is the most likely explanation for the majority of the sample attrition.

\textsuperscript{11} Most states give students 3-4 years, from their grade of first enrollment in the state, to achieve proficiency. After this time students are assessed by status only. Consequently, the largest potential impact of the GMPP is on elementary schools, i.e. the grades in which most students spend their first 3-4 years of enrollment in a state.

\textsuperscript{12} The original data file contained 11,485 3rd grade students in 2002; however, 1,478 of these students did not have achievement score data. In these analyses I only consider the 10,007 students with 3rd grade achievement data in 2002.
Table 3.1 Frequency Counts for the 3rd Grade Cohort of 2001-2002

<table>
<thead>
<tr>
<th>Grade</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10,007</td>
<td>209</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8,653</td>
<td>556</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>7,742</td>
<td>665</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>42</td>
<td>6,750</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>10</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>10,007</td>
<td>8,875</td>
<td>8,364</td>
<td>7,550</td>
</tr>
</tbody>
</table>

NA: These students did not have achievement data available this year.

Sample attrition can bias analyses if those students who leave the sample are systematically different from those who remain in the sample. Table 3.2 (below) provides a breakdown of the demographic characteristics (Limited English Proficiency, Free/Reduced priced lunch, Special Education, and Race) for the 10,007 students with exam scores in 2002, and the 7,550 students with exam scores in 2005. Attrition does not appear to have significantly changed the demographic makeup of the observed students. The largest demographic shift was in the race category where 42 percent of students were Black in 2002 and 46 percent of students were Black in 2005. This four percentage point increase corresponds with the four percentage point decrease in the percentage of White students. Outside of the race category, no other demographic shift was greater than two percentage points. In addition, the average 3rd grade FCAT score of the full sample was 1279; the subsample was 1280, nearly identical. These data suggest that attrition did not drastically change the demographic makeup of the sample.
Table 3.2 Demographic Characteristics of the 3rd Grade Cohort of 2001-2002

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>2001-2002</th>
<th>2004-2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEP</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>FRL</td>
<td>53%</td>
<td>51%</td>
</tr>
<tr>
<td>SpEd</td>
<td>21%</td>
<td>23%</td>
</tr>
<tr>
<td>Asian/PI</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>Black</td>
<td>42%</td>
<td>46%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>Indian/AK Native</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Mixed</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>White</td>
<td>49%</td>
<td>45%</td>
</tr>
<tr>
<td>Grade 3 Math Avg. Score</td>
<td>1279</td>
<td>1280</td>
</tr>
</tbody>
</table>

When considering attrition as a potential source of bias, it is important to note that this study does not attempt to compare two groups of students to each other, as in an experiment requiring group equivalence; rather, this study examines various measurement techniques. As such, what is most important is whether the learning trajectories of students in the subsample are significantly different from students in the full sample. If such differences exist, in order for this to be a problem they must occur in a way that makes the measurement techniques behave differently in the subsample compared to the full sample. There is not strong evidence that this is the case.

Sample Descriptive Statistics

The FCAT is a vertically equated exam, meaning that student’s longitudinal test scores\(^{13}\) are on the same metric over time and across grades. Descriptive statistics for the

\(^{13}\) This is only true of the FCAT Developmental Scale Scores. Students in Florida are also given scores on other metrics. In this study any reference to “scores” or “scale scores” or “test scores” refer to the Developmental Scale Scores.
The study district's average Mathematics FCAT DSS in 3rd grade 2002 was 1279 with a standard deviation of 297. From 2003 through 2005, the average student from the 2002 cohort of 3rd graders scored 1418, 1594, and 1617 respectively. These scores imply average annual gains of approximately 139, 176, and 23 points on the developmental scale. These average gains are fairly similar to statewide aggregated average gains between 2002 and 2005 for 3rd to 6th graders. The statewide average Mathematics FCAT score for 3rd graders in 2002 was 1308, for 4th graders in 2003 was 1446, for 5th graders in 2004 was 1616, and for 6th graders in 2005 was 1653. Statewide average mathematics aggregated gains were approximately 138, 170, and 37. Although somewhat imprecise due to grade retention, grade skipping, and student mobility, these numbers suggest that the average learning trajectories of students in the study district were fairly similar to the average learning trajectory of students in the state as a whole. Notably, typical gains in

---

14 Due to attrition, these “average gains” are not perfectly precise. For example, the actual average gain of the 8,874 students who remained in the sample was 135 between 2002 and 2003.
the study district and throughout the state were significantly larger from 3rd to 5th grade compared to the relatively modest gains observed between 5th and 6th grade, reflecting a curvilinear developmental scale.

Table 3.3 (above) also provides the pass rates among students in the 2002 cohort of 3rd graders from the study district. Fifty-five percent of students passed the 3rd grade exam in 2002. In 2003-2005, the percentage of these students who passed the state-wide exam was 50, 50, and 41, respectively. While the percent proficient decreased over time, this does not suggest that learning did not take place. The percent proficient can decrease over time even when learning occurs, because the proficiency cut scores rise as students move to higher grades. This explains the increasing average test scores (which do imply learning on a vertically equated assessment) and the decreasing percentage of students passing the statewide exam. Most significant is the fact that in 2005 41 percent of the sample was observed scoring above the 6th grade proficiency cut score. This percentage is relevant because the projection models examined in the analyses attempt to project proficiency status in 2005.

Finally, the last column in Table 3.3 represents the percentage of students whose 2003-2005 score exceeded the proficiency cut score for the grade they would have been in had they advanced exactly one grade per year. The percentages are fairly similar to the percentage of students who passed the exam in their actual grade, suggesting that grade retention and grade skipping should not significantly influence the analyses.

This section provided a brief overview of the data used in these analyses. The following section mathematically describes each of the five unique projection models as designed by the states for use in the GMPP.
Methods

The goal of the state projection models used under the GMPP is to give schools credit for students who are on track to become proficient. However, states use different methods to determine whether students are on track to become proficient; that is, they use different methods to project future proficiency. In this section I describe the formal mathematical equations behind each state’s procedure for determining whether each student is on track to become proficient. The analysis techniques used to answer my research questions are described within the next chapter, Chapter 4: Analyses and Results.

States typically allow schools three or four years to bring students up to proficiency. In my analyses I force each model to give students three years to become proficient in order to make consistent comparisons among the various state models. If states are unable to accurately project future proficiency when giving students three years to become proficient, they should not be expected make accurate projections an additional year into the future. The logic of each state’s projection model is applied to Florida’s mathematics developmental scale and to student achievement data from the study district. Limitations of this approach are discussed in Chapter 6: Study Limitations.

The models are described starting with Florida’s simple linear projection model and ending with Tennessee’s complex statistical model.
Florida’s Projection Model

Florida uses a linear projection model. Under Florida’s model a student is labeled on track to become proficient, and the school receives “credit” for her, if her observed achievement score exceeds her growth target. Each student’s growth target is set separately. The growth targets are placed along a linear trajectory from a student’s initial achievement score to the proficiency cut score three years after the student’s grade of first enrollment (or first instate exam). For example, a student’s year two growth target is set using equation (3-1):

\[(3-1) \quad \tilde{y}_{i2} = y_{i1} + \frac{1}{3}(y_{4,\text{cutscore}} - y_{i1})\]

Where:

\(\tilde{y}_{i2}\) = student i’s year two growth target

\(y_{i1}\) = student i’s year one observed developmental scale score

\(y_{4,\text{cutscore}}\) = year four proficiency cut score

In order to meet her year two growth target, a student must make up one third of the distance from her initial achievement score to the proficiency cut score in year four. If a student’s observed year two score \((y_{i2})\) is greater than or equal to the year two growth target \((\tilde{y}_{i2})\), then she is labeled on track to become proficient. This is a two-year projection since based on a student’s year one and year two scores, the state projects whether she is on track to become proficient by year four, i.e. two years into the future.

A student’s year three growth target is set using equation (3-2):
In order to meet her year three growth target, a student must make up two thirds of the distance from her initial achievement score to the proficiency cut score in year four. If a student’s observed year three scale score \((y_{3})\) is greater than or equal to the year three growth target \((\tilde{y}_{3})\) then she is labeled on track to become proficient. This is a one-year projection since based on a student’s year one and year three scores, the state projects whether she is on track to become proficient by year four, i.e. one year into the future.

In year four the projection model is no longer used; schools only receive credit for those students who actually score at or above proficient.

**Arizona’s Projection Model**

Arizona’s projection model involves a two-stage process. In step one linear growth targets are set in a way very similar to Florida’s method. Year two growth targets are set as described in equation (3-1) above. However, rather than compare a student’s *observed* year two score to her year two growth target (as Florida does), Arizona compares the lower bound of a student’s *predicted* year two score to her growth target. Note that Arizona is *not* creating a projected interval for an *unobserved* future score; rather, they are creating an interval for a predicted year two score (even though the observed year two score is known) in an attempt to account for unreliability of gain scores and regression towards the mean. The lower bound of a student’s predicted year
two score is calculated through a school fixed effects regression model, described by equation (3-3):

\[ y_2 = \alpha_s + \beta y_1 + \varepsilon \]  

Where:

\( y_t \) = observed year \( t \) developmental scale score

\( \alpha_s \) = Fixed effect for school \( s \)

\( \beta \) = Regression coefficient for year one developmental scale score

The model described in equation (3-3) regresses students’ current observed scores (\( y_2 \)) on their past scores (\( y_1 \)) and a school fixed effect (\( \alpha_s \)). For each student, a year two score is predicted (\( \hat{y}_2 \)) using the estimated coefficients obtained from the regression model (\( \alpha_s \) and \( \beta \)) and the student’s year one scale score (\( y_1 \)) as described in equation (3-4):

\[ \hat{y}_2 = \alpha_s + \beta y_1 \]

In order to make a conservative estimate of student performance, the lower bound of a 95 percent interval around \( \hat{y}_2 \) is compared to the growth target. The lower bound of a 95 percent interval around \( \hat{y}_2 \) is calculated using equation (3-5):

\[ \hat{y}_{2, \text{Lower}97.5} = \hat{y}_2 - t_{2.5} \cdot SE(\hat{y}_2) \]

The standard error is defined as \( SE(\hat{y}_2) = \sqrt{h_i s^2} \), where \( h_i = x_i (X'X)^{-1} x_i' \); \( X \) is the matrix of regressors, \( x_i \) is the \( i^{th} \) row of \( X \) and \( s^2 \) is the mean squared error.

If the lower bound of a student’s predicted year two scale score (\( \hat{y}_{2, \text{Lower}97.5} \)) is greater than or equal to the year two growth target (\( \bar{y}_{1.2} \) described in equation (3-1)), then
the student is labeled on track to become proficient. This is a two-year projection because based on a student’s year one score and predicted year two score, the state projects whether a student is on track to become proficient by year four, i.e. two years into the future.

Arizona’s year three growth target is also set using a linear model, but instead of drawing a linear trajectory from a student’s observed year one score to the year four proficiency target (like Florida’s model), Arizona’s model draws a straight line from a student’s observed year two score to the year four proficiency target. Thus, a student’s year three growth target is set using equation (3-6):

\[
\tilde{y}_{i3} = y_{i2} + \left(y_{4\text{-cut score}} - y_{i2}\right)/2
\]

Again, rather than compare a student’s observed year three score to her growth target, the lower bound of a student’s predicted year three score is compared to her growth target. The lower bound of a student’s predicted year three score is calculated through a school fixed effects regression model, described in equation (3-7).

\[
y_{3} = \alpha_{s} + f_{2}y_{2} + \varepsilon
\]

The rest of the calculations follow the same pattern as described for the year two growth target.

North Carolina’s Projection Model

North Carolina uses a linear projection model much like Florida’s model; however, North Carolina applies the linear projection model to standardized scores (which they call C-Scale scores) rather than to the developmental scale scores. Standardized scores are calculated separately for each grade and subject. They are
obtained by subtracting the mean developmental scale score (from the standard setting year) from each student’s observed developmental scale score, and then dividing by the standard deviation (from the standard setting year), as described in equation (3-8):

\[
y_{i,C\text{-Scale}} = \frac{y_i - \bar{y}}{s}
\]

Where:

\(y_{i,C\text{-Scale}}\) = Student’s C-Scale score

\(y_i\) = Student’s developmental scale score

\(\bar{y}\) = The mean developmental scale score of all students in the standard setting year

\(s\) = The standard deviation of the developmental scale scores of all students in the standard setting year

North Carolina’s year two and year three growth targets can be described mathematically by applying Florida’s linear projection equations to the C-Scale scores. Year two growth targets can be described using equation (3-9):

\[
\tilde{y}_{i2,C\text{-Scale}} = y_{i1,C\text{-Scale}} + \frac{1}{3}(y_{4,\text{cutcore,C\text{-Scale}}} - y_{i1,C\text{-Scale}})
\]

Where:

\(\tilde{y}_{i2,C\text{-Scale}}\) = student’s year two growth target on the C-Scale

\(y_{i1,C\text{-Scale}}\) = student’s observed year one C-Scale score

\(y_{4,\text{cutcore,C\text{-Scale}}}\) = year four proficiency cut score on the C-Scale

In order to meet her year two growth target, a student must make up one third of the distance from her initial achievement score (on the C-Scale) to the proficiency cut score (on the C-Scale) in year four. If a student’s observed year two C-Scale score
(\(y_{12,C\text{-Scale}}\)) is greater than or equal to the year two C-Scale growth target (\(\tilde{y}_{12,C\text{-Scale}}\)), then she is labeled on track to become proficient. This is a two-year projection since based on a student’s year one and two scores, the state projects whether a student is on track to become proficient by year four, i.e. two years into the future.

Year three growth targets can be described using equation (3-10).

\[
(3-10) \quad \tilde{y}_{13,C\text{-Scale}} = y_{11,C\text{-Scale}} + \frac{2}{3}(y_{4,\text{cutscore,C\text{-Scale}}} - y_{11,C\text{-Scale}})
\]

Arkansas’ Projection Model

Arkansas’ proportional growth model assumes that a student’s growth will mimic the pattern of change in proficiency cut scores over time. Consider the proficiency cut scores from Florida’s mathematics FCAT. The cut scores from 3rd to 6th grade are 1269, 1444, 1632, and 1692 respectively. The total change in the proficiency cut score from 3rd to 6th grade is 423 points. The change from 3rd to 4th grade (175 points) represents 41.37 percent of the total change in proficiency cut score from 3rd to 6th grade. For purposes of comparison in this study, Arkansas’ on track to become proficient growth targets are set such that between 3rd and 4th grade each student must make up 41.37 percent of the gap between her initial achievement score and the 6th grade proficiency cut score.\(^{15}\) The year two growth target can be described by equation (3-11):\(^{16}\)

---

\(^{15}\) In actuality, Arkansas gives students four years to become proficient, not three years. Therefore, 7th grade would be the actual endpoint. However, for the sake of a more fair comparison, I consider three years after grade of first enrollment in my analyses.

\(^{16}\) The value of .4137 is specific to a student whose first mathematics exam was taken in 3rd grade. If a student entered the state school system in 7th grade, this value would change to reflect the changes in the proficiency cut scores at that point in the developmental scale.
(3-11) \[ \tilde{y}_{i2} = y_{i1} + P_{1-2} * (y_{4,\text{cutscore}} - y_{i1}) \]

Where:

\( \tilde{y}_{i2} \) = student i’s year two growth target

\( y_{i1} \) = student i’s observed year one developmental scale score

\( y_{4,\text{cutscore}} \) = year four proficiency cut score

\( P_{1-2} \) = the proportion of the change in proficiency cut score from year 1 to year 4 which occurs between year 1 and year 2.

In order to meet her year two growth target, a student must make up \( P_{1-2} \) percent of the distance from her initial achievement score to the proficiency cut score in year four (\( P_{1-2} = .4137 \) if the student’s grade of first enrollment is 3rd grade). If a student’s observed year two scale score (\( y_{i2} \)) exceeds the year two growth target (\( \tilde{y}_{i2} \)) then the student is deemed on track to become proficient in year two. This is a two-year projection, since based on a student’s year one and year two scores, the state projects that a student is on track to become proficient by year four, two years into the future.

Equation (3-11) described the year two growth target. Through similar logic, the year three growth target can be calculated using equation (3-12):

(3-12) \[ \tilde{y}_{i3} = y_{i1} + P_{1-3} * (y_{4,\text{cutscore}} - y_{i1}) \]

In order to meet her year two growth target, a student must make up \( P_{1-3} \) percent of the distance from her initial achievement score to the proficiency cut score in year four. If a student’s observed year three scale score (\( y_{i3} \)) exceeds the year three growth target (\( \tilde{y}_{i3} \)) then the student is deemed on track to become proficient in year three. This
is a one-year projection, since based on a student’s year one and year three scores, the state projects that a student is on track to become proficient by year four, one year into the future.

**Tennessee’s Projection Model (EVAAS®)**

Tennessee’s projection model is a statistical model. In Tennessee’s proposal (Tennessee Department of Education, 2006) and an article by Wright et al. (2006), they describe EVAAS using equation (3-13):

\[(3-13) \quad \text{Projected Score} = M_y + b_1(X_1 - M_x) + b_2(X_2 - M_x) + ... = M_y + x^T b\]

This equation can be re-written to correspond with the mathematical symbols used in the rest of this chapter as follows (3-14):

\[(3-14) \quad \tilde{y}_4 = \tilde{y}_4 + \beta_1(y_1 - \tilde{y}_1) + \beta_2(y_2 - \tilde{y}_2) + ...\]

Where:

\(\tilde{y}_t = \) The grand mean DSS of students in year \(t\).

EVAAS uses the longitudinal records of the current 6th graders, i.e. their 3rd through 6th grade achievement scores, to calculate the regression parameters (\(\beta\)’s). Once calculated, the parameters are then applied to the current 5th grade students who have \(y_1, y_2,\) and \(y_3\) scores, but no \(y_4\) score, in order to obtain a projected \(y_4\) score (\(\tilde{y}_4\)).

To paraphrase Wright et al (2006), Tennessee obtains the projection parameters using the most recent year’s data. They use students who have a \(y_4\) value in the most recent year and \(y_1, y_2,\) and \(y_3\) values from earlier years to get the projection parameters. Projections are then obtained by applying these parameters to students who have \(y_1, y_2,\) and \(y_3\) values.
(or $y_1$ and $y_2$ values only if the student is in year two) in the current year (and earlier years) but no $y_4$ value (Wright et al., 2006, p. 390). Projected scores are compared to the year-four proficiency cut score in order to determine if a student is on track to become proficient. If a student is projected to score at or above the proficiency cut score then she is labeled on track to become proficient.

In order to mimic the EVAAS approach and to be able to determine the accuracy of the projections would typically require 5 years of data. For example, the parameter estimates could be obtained using year one through year four scores of a cohort of students who began 3rd grade in year one. The parameter estimates could then be applied to the next cohort of students, who begin 3rd grade in year two. Projections, for the second cohort of students, could then be made in year four to project year five scores (typically 6th grade scores); however, in order to determine the accuracy of the projections one would need to know the year five scores. In other words, verification of this model’s accuracy requires four years of data from the same grades for two different cohorts of students, thus requiring a minimum of five years worth of data.

With only four years of data on hand I circumvented this issue by randomly separating the data into two equally sized groups, one “training” group, the other a “validation” group. Using year one through year four data, the training group was used to obtain the parameter estimates in equation (3-14). Then, applying the parameter estimates to the validation groups’ year one, two, and three scores, projections were obtained for year four. Projected scores were then converted into proficient or not proficient based on whether they were above or below the proficiency cut score. If a student’s projected score was above the proficiency cut score, then she was deemed on
track to become proficient. Projections were also used to calculate residuals, the difference between the projected scores in year four and the observed scores in year four.

Since both the training and validation dataset came from the same cohort of students, one might expect individual-level projections to be more accurate compared to a true application of Tennessee’s model, which applies model coefficients from prior cohorts to future cohorts. However since this methodology relies on the intra-student correlations over time, it is not too likely that a one cohort shift will decrease accuracy significantly, i.e. intra-student correlations are unlikely to differ greatly from one cohort to the next.

Although equation (3-14) appears like an ordinary multiple regression, the parameter estimates are obtained in a slightly different way than might be expected. The steps used to obtain regression coefficients are described below based upon the methodology as explained by Wright et al. (2006) and in Tennessee’s growth model proposal (Tennessee Department of Education, 2006):

Step 1: School-mean center the scale scores of the students in the training dataset by subtracting the school mean from the students’ scores in each school.

Step 2: The EM algorithm, as implemented in SAS PROC MI (applied to the school-mean centered scale scores), is used to calculate the pooled-within-school covariance matrix17 for the training dataset.

---

17 Obtaining the covariance matrix through SAS’s proc mi is described in Tennessee’s GMPP proposal: “The covariance matrix of these centered scores is obtained by maximum likelihood (ML) estimation using the EM algorithm implemented in the MI procedure in SAS/STAT” (Tennessee Department of Education, 2006, p. 16).
Step 3: Regression coefficients are estimated from the covariance matrix obtained in step two, using a fixed intercept representing the grand mean.\(^{18}\) Step three was conducted separately for one and two-year projections.

a. **Two-year projections**: Regression parameters for the two-year projections are estimated based off the year one, year two, and year four covariances obtained in step two. The fixed intercept (i.e. the training dataset’s grand mean, \(\bar{y}_4\)) and the parameters obtained from this model (\(b_1\) and \(b_2\)) are then applied to those students in the validation dataset with year one and year two scores, in order to calculate projected year four scores (\(\bar{y}_4\)).\(^{19}\)

b. **One-year projections**: Two separate regression models are run in order to make one-year projections. The separate models are run in order to make projections for students with two unique patterns of data.\(^{20}\) The first model is used for students with complete data and estimates regression parameters off the complete covariance matrix obtained in step two. The fixed intercept (i.e. the training dataset’s grand mean, \(\bar{y}_4\)) and the parameters obtained from this model (\(b_1, b_2,\) and \(b_3\)) are

---

\(^{18}\) The “grand mean” is calculated as the unweighted mean of the school means. This interpretation stems from the statement “Means for an “average school” are obtained by calculating school-mean scores and averaging them over schools” (Wright, Sanders, & McCall, 2006, p.390).

\(^{19}\) The scores from the validation dataset plugged into the regression equation represent grand mean centered scores, as described in equations (3-13) and (3-14).

\(^{20}\) Wright et al. (2007) describe the “covariance matrix of all the predictors plus the response” using the matrix \(C\). They then state that “For any given student, one can use the subset of \(C\) corresponding to that student’s set of scores to obtain the regression coefficients for projecting that student’s Y value” (Wright et al., 2006, p.390).
then applied to those students in the validation dataset with year one, two, and three scores, in order to calculate projected year four scores ($\hat{y}_4$). The second model is used for those students with incomplete data. It estimates regression parameters off the year one, year three, and year four covariances obtained in step two. The fixed intercept (i.e. the training dataset’s grand mean, $\bar{y}_4$) and the parameters obtained from this model ($b_1$ and $b_2$) are then applied to those students in the validation dataset with year one and year three scores, in order to calculate projected year four scores ($\hat{y}_4$). One-year projections are restricted to only those students (7,477) with at least year one, year three and year four scores. Very few students (135) fell into the category of having a year three score, but no year two score. It is for this small subset of students that the greatest uncertainty lies regarding my replication of EVAAS’s methodology.

This chapter provided a detailed description of each of the five unique projection models being implemented under the GMPP as of 2008. In the subsequent chapter each of the models is applied to the study district’s data. The specific analyses and techniques are included within the text of the next chapter.
Chapter 4: Analyses and Results

The objective of this chapter is to empirically investigate each of the four major research questions posed in this paper. The analyses are separated into four sections, each addressing one of the four research questions. Sections one and two focus on the accuracy of the projection models at the individual level and at the school level, respectively. The third section addresses whether the individual-level growth goals of NCLB (status) and the GMPP (projection models) are realistic. The final section considers the similarities and differences among measurements of school performance using a status model, a projection model, and a value-added model.

Research Question 1: How accurate are projection models at the student level?

One of the goals of the federal Growth Model Pilot Program is to give schools credit for those students who are not currently proficient, but appear on track to become proficient in the near future. To measure whether students are on track to become proficient, states have developed different projection models which assess whether students have made sufficient learning gains such that one might expect they will become proficient in the future. Depending upon the state’s proposal, projections are made one, two, and sometimes more than two years into the future. In these analyses I assess the accuracy of projections made one and two years into the future. Depending upon the state’s proposal, projection models are applied either to all students or only to those students who are currently not proficient. In these analyses I apply projection models to all students. Essentially, in this section I look to answer the question: When a state’s model labels a student as on track to become proficient, does she become proficient?
Six of the first seven states piloting projection models under the GMPP require that students become proficient a fixed number of years after their “grade of first enrollment,” where grade of first enrollment is considered to be the grade in which students take their first instate assessment. Florida, for example, requires students to achieve proficiency three years after their grade of first enrollment in the state. Since most students (excluding those who transfer into a Florida public school after 3rd grade) take their first state-wide assessment in 3rd grade, under Florida’s GMPP proposal most students must become proficient by 6th grade. From 6th grade onwards projections are not calculated for students (excluding those who transfer into a Florida public school after 3rd grade), rather students must actually be proficient for a school to receive credit for their performance. By not resetting the target grade at which students must become proficient, states ensure that students are required to achieve proficiency a set number of years after their grades of first enrollment. In contrast, upping the target proficiency grade yearly may lead to some students always being on track to become proficient, but never actually achieving proficiency. Under a system like Florida’s, where students have three years to become proficient after their grade of first enrollment, projection models are most likely to have an impact at the elementary school level. This is the case because 3rd grade is the most common grade of first enrollment in which an exam is given; consequently, for most students projections are calculated in 4th and 5th grade only. It is for this reason that these analyses track the cohort of students who took the 3rd grade assessment in

---

21 Tennessee updates their proficiency target each year.

22 Projections are not made in 3rd grade because testing typically begins in 3rd grade, so “growth” cannot be assessed with only a single time point. North Carolina is the exception. They give a 3rd grade pretest and are thus able to make projections at the end of 3rd grade. In 6th grade students must actually be proficient for a school to receive credit for them, thus projections are not made in 6th grade.
2002, since they are the students most likely to be affected under most states’ proposals for the GMPP.

Although states vary with respect to how many years after grade of first enrollment students have to attain proficiency, in order to compare states’ projection models’ methodologies consistently, I set the proficiency target at three years after grade of first enrollment. As a result, under the GMPP my sample of 10,007 3rd graders in 2002 would be required to become proficient by 2005. Only those students (7,550) with achievement scores in at least the years 2002 and 2005 were included in the analyses. Using this historical data I am able to make projections one and two years into the future and check the accuracy of the projections. That is, using students’ 2002 and 2003 achievement scores I project whether they are on track to be proficient in 2005 (a two-year projection). The projections are checked for accuracy by comparing students’ projected proficiency status to their actual 2005 proficiency status. Similarly, using students 2002, 2003, and 2004 achievement scores I project whether students are on track to be proficient in 2005 (a one-year projection) and then compare the projections to students’ actual proficiency status. Most state models (Arkansas, Florida, and North Carolina) only require the use of 2002 and 2004 data in order to make projections for 2005, so in those cases the 2003 data were ignored. While some states attempt to make projections more than two years into the future, we can be confident that projection accuracy will only degrade as we attempt to project farther into the future.

States also vary with respect to the population to whom they apply their projection model. Some states only apply their projection model to students who are

---

23 When referring to the school year, I will refer to the calendar year in which the student took the exam.
currently *not* proficient. These states give schools credit for students who are either proficient or on track to become proficient. Other states apply projection models to *all* students, only giving schools credit for those students who are on track to be proficient. This difference can be significant because some students who are currently proficient may not be on track to remain proficient. All analyses conducted in this paper utilize the latter approach, applying projection models to all students.

These projections are calculated using each of the five state projection methods described in Chapter 3: Data and Methods. The methods are all applied to Florida’s mathematics FCAT, a vertically equated, reliable assessment.

*Projection Model Accuracy (Student Level)*

Projected proficiency is compared to observed proficiency in order to determine projection model accuracy. Table 4.1 (below) compares the projections to the actual observed results. State models are arranged alphabetically.

In Table 4.1 a “true positive” is defined as those students who were projected to be proficient in 2005 and actually were proficient in 2005. “True negatives” are those students who were projected not to be proficient in 2005 and were not proficient in 2005. Summing the true positives and the true negatives yields the “overall accuracy” of the model, or the percent of the projections that were correct. “False positives” are those students who were projected to be proficient but were observed scoring below proficiency. “False negatives” represent those students who were projected to score below proficiency but actually scored above proficiency.
Several striking findings can be observed in Table 4.1. First, notice that overall accuracy is largely related to the model used to make projections. When making one-year projections Arizona’s and Florida’s models are accurate only 72 and 70 percent of the time respectively, whereas Arkansas’ and Tennessee’s models are accurate 82 percent of the time. Similar results are observed when making two-year projections, where the most accurate model (Tennessee) is correct 76 percent of the time, and the least accurate models (Arizona and Florida) are correct only 61 percent of the time. These large differences in projection model accuracy indicate that states’ decisions regarding which

---

24 As described in Chapter 3: Data and Methods, the results for Tennessee’s model were obtained by randomly separating the data into two equal halves – a training dataset, used to obtain regression parameters, and a validation dataset, used to assess the accuracy of the model. The results presented in Table 4.1 (and the rest of this chapter) represent the application of Tennessee’s model to a unique training and validation dataset. Since the results could be influenced by sampling variability, sensitivity analyses were conducted by recreating Tennessee’s results from Table 4.1 using 1000 randomly created training and validation datasets. The mean overall accuracy, true positives, true negatives, false positives, and false negatives were all within ±1 percent of the values observed in Table 4.1, with standard deviations that were all less than 1.7. Thus, there is no evidence that the results presented are a result of sampling variability.

---
projection model to use has a large influence on the level of accuracy of their individual-level projections.

With regard to the overall accuracy of the model, as expected, single-year projections are more accurate than two-year projections. While this finding is unsurprising, what is of significance is how much less accurate Arizona’s, Arkansas’, Florida’s, and North Carolina’s projections are when making two-year projections rather than one-year projections. These models are only accurate 61 to 65 percent of the time when making projections two years into the future. Given that the 6th grade pass rate among observed students was 41 percent, using no model at all (simply projecting that all students will fail) would be accurate 59 percent of the time. As such, accuracy levels in the low 60’s are relatively poor. In contrast, Tennessee’s two-year projections remain accurate 76 percent of the time, a relatively modest drop in accuracy compared to their one-year projections.

Another way to assess whether the overall accuracy of these models is sufficiently good is to compare their results to a naïve model which does not use individual growth to project future proficiency. That is, we can consider what would happen if we assumed that any student who is currently proficient were labeled on track to become proficient. One would hope that a projection model which uses multiple years of data to project proficiency based on individual growth would, at the very least, be more accurate than simply assuming that all students who are currently proficient will remain proficient in the future. Table 4.2 (below) presents classification rates similar to those in Table 4.1 (above), but rather than using a growth model to project future proficiency, this table
considers the concordance between current status and future proficiency using the naïve model described above.

**Table 4.2 Concordance between Current Status and Future Proficiency (Student Level)**

<table>
<thead>
<tr>
<th></th>
<th>Overall Accuracy</th>
<th>True Positive</th>
<th>True Negative</th>
<th>False Positive</th>
<th>False Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One Year Projections</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Status</td>
<td>82%</td>
<td>37%</td>
<td>45%</td>
<td>13%</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Two Year Projections</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Status</td>
<td>79%</td>
<td>35%</td>
<td>44%</td>
<td>15%</td>
<td>6%</td>
</tr>
</tbody>
</table>

In Table 4.2 the overall accuracy of the one-year projection is 82 percent and the overall accuracy of the two-year projection is 79 percent. These findings are interesting for two reasons. First, this tells us that 82 percent of students remained at the same proficiency status between 2004 and 2005, and 79 percent had the same proficiency status in 2003 and 2005. Second, this suggests that *simply assuming that all students will remain in the same proficiency category is as accurate as the best projection model at making one-year projections, and even more accurate than the best model at making two-year projections.*

Since current status is as good of an indicator of being on track to become proficient as any of the state proposed projection models, there is little evidence for the use of these projection models, unless their accuracy is improved. In order to assess how we can improve these models, it is important to consider where they go wrong.

**Model Bias (Student Level)**

The overall accuracy of the projection models can be broken down into subcategories (true positives, true negatives, etc.) in order to gain insight regarding
potential sources of bias in the models’ projections. Model bias refers to the systematic overestimation or underestimation of projections. *Model bias is a concern because biased models are less likely to make accurate projections and may lead to misguided conclusions.*

The clearest example to illustrate model bias is Florida’s single-year projection model. Under Florida’s linear assumption, the model makes accurate projections 70 percent of the time. As such, 30 percent of projections are incorrect: 27 percent false positives and 3 percent false negatives. This means that 9 out of every 10 incorrect projections involves falsely projecting that students are on track to become proficient. The reason Florida’s projection model has a propensity toward false positives is that typical growth patterns on Florida’s mathematics FCAT show diminishing gains over time, while Florida’s model assumes constant gains over time. Since students’ growth trajectories tend to be curvilinear, Florida’s underspecified linear model is likely to overestimate the number of students who are on track to become proficient, resulting in disproportionately large numbers of false positives and far fewer false negatives.

Similarly, it appears that Tennessee’s model may be biased in the other direction. The majority of Tennessee’s incorrect projections occur when students who were not projected to become proficient end up scoring proficient. Tennessee’s model rarely projects proficiency for students who do not achieve proficiency. It appears that the model is making projections that are generally too low; however, given that the majority of students in 6th grade are not proficient (41 percent of this sample), it may not be unreasonable to observe more false negatives than false positives, even if the model is unbiased.
Analyzing Model Bias (Student Level)

Thus far the accuracy of each state projection models has been analyzed based on binary outcomes – does the model project that a student will become proficient? Did she actually become proficient? While this is a valuable way to assess empirically the accuracy of the models, relatively simple statistical and mathematical theory can be used to better appreciate when, why, and by how much these models make (in)accurate projections.

Underlying each state projection model is a fundamental assumption regarding the shape of student growth trajectories. Florida, Arizona, and North Carolina all have models that assume linear growth. These three models differ in that Arizona attempts to account for regression toward the mean and North Carolina uses a standardized scale, whereas Florida includes neither of these adjustments. Arkansas' model assumes growth that mimics the shape of changes in proficiency cut scores, whereas Tennessee's model assumes that students' growth trajectories will mimic the shape of the previous cohort of students' growth trajectories.

These underlying assumptions dictate where the 4th and 5th grade growth targets are set in order to determine whether a student is on track to become proficient, as described in Chapter 3: Data and Methods. Although most of the states’ models do not explicitly project future scale scores (only Tennessee does this explicitly), the implicit assumptions underlying each state’s model can be applied to students’ observed scale scores (e.g., 3rd and 4th grade scale scores) in order to project future scale scores (e.g., 6th

---

25 Tennessee’s model does not set growth targets. It projects a future scale score.
grade scale scores), not just future proficiency status. By comparing a student’s projected 6th grade scale score to the known 6th grade proficiency cut score, one can obtain proficiency projections (i.e. is a student on track to become proficient?). These proficiency projections are identical to those obtained using the growth targets, since both rely on precisely the same primary assumptions. However, applying the underlying assumptions of state models in order to obtain projected scale scores allows us to gain a more detailed understanding of when and why state models do and do not work. The greater detail is afforded because rather than analyzing the blunt binary outcome obtained when applying each state’s model as explicitly described, obtaining projections on a continuous metric provides a richer depiction of the consequences of each state’s model’s underlying assumptions. Through careful examination of the differences between projected 6th grade scale scores and observed 6th grade scale scores, a deeper understanding of these models can be gained. Appendix A: Analyzing Projection Models formally (mathematically) describes how the logic and fundamental assumption of each state’s model can be used to calculate 6th grade projected scores. The additional analyses of the student-level projections presented from this point forward rely on the use of projected 6th grade scale scores. The school-level results presented afterward revert back to using the projection models as described by the states themselves.

**Absolute Bias (Student Level)**

In assessing model absolute bias one can examine whether projected 6th grade scores tend to overestimate or underestimate observed 6th grade scale scores. Letting \( \hat{y} \) refer to the projected score, and \( y \) refer to the observed score, I employ the term
"absolute bias" to refer to the difference between the mean $E(\hat{y} - y)$ of the distribution of $\hat{y} - y$ and 0 (Levy & Lemeshow, 1999, p. 33). Therefore, the absolute bias is the average residual. Simply stated, the absolute bias is the average distance between students' projected scores and their observed scores.

Figure 4.1 plots the residuals from each state’s one-year projection model. The residuals, plotted on the y-axis, are equal to the projected 2005 score minus the observed 2005 score. A residual equal to zero indicates that the model’s projection was exactly correct, e.g. the model projected a student would score 1703 and she actually scored 1703. Positive residuals reflect projections that were greater than the observed score and negative residuals reflect projections that were less than the observed score. The more observations with residuals close to zero the more likely the model is to be accurate. The further away the model’s average residual is from zero, the more absolute bias the model demonstrates. To put the size of the residuals into context, the observed standard deviation of the mathematics 2005 exam was 250.\textsuperscript{26}

\textsuperscript{26} This is the standard deviation calculated for those students who entered 3\textsuperscript{rd} grade in 2002 and had exam scores in both 2002 and 2005 (n=7,550).
Figure 4.1 shows that over 75 percent of Florida’s *one-year projections* are too high (positive residuals). Florida’s linear projection consistently overestimates how well students will score in 6th grade, demonstrating that Florida’s model is highly biased. This occurs because of the faulty linear assumption. Arizona’s model also tends to overestimate students’ scores, although with less absolute bias than Florida’s model. Arizona’s model is slightly less biased than Florida’s because it uses the conservative lower bound of the 95 percent confidence interval of each student’s predicted time two score. This conservative approach still does not overcome the bias associated with setting linear growth targets on a developmental scale where student growth is curvilinear. In contrast, Arkansas’ and North Carolina’s models are far less biased, with average residuals closer to zero. Tennessee’s model is the only model that is prone to underestimate student achievement, suggesting some degree of model bias. However, Tennessee’s relatively small absolute bias may be mitigated by its superior precision, i.e. the spread of Tennessee’s residuals is much smaller than that of the other states.
Numerically, model precision can be measured by the standard deviation of the residuals, which are 184, 170, 214, 218, and 140 for Arizona, Arkansas, Florida, North Carolina and Tennessee respectively.

Figure 4.2 plots the residuals from each of the states’ two-year projection models. Note that the y-axis now ranges from -3500 to 3500, reflecting projections that are frequently far from the observed scale scores. While evidence of absolute bias is less apparent in this figure (except for Tennessee’s model), what is remarkable is the dramatic imprecision of these models’ projections (except for Tennessee’s model). The standard deviation of the residuals for Arizona, Arkansas, Florida, and North Carolina are 258, 380, 484, and 503 respectively. Consider this: using a projection model that projects that all students will earn exactly the same score (it does not matter which score), the standard deviation of the residuals would be only 250. That is, these four models are actually less precise than a null model. In stark contrast, the standard deviation of the residuals for Tennessee is only 159, indicating that Tennessee’s model is by far the most precise. However, Tennessee’s average two-year projected score is 55 points below the average observed score, or .22 (55/250) standard deviations below the average observed score. This is evidence of significant absolute bias. Nonetheless, Tennessee’s model’s superior precision leads it to be the most accurate.
Relative Bias (Student Level)

A projection model with no absolute bias does not, on average, overestimate or underestimate students' observed scores. In a projection model with no "relative bias," students' projected scores do not provide any indication of the relative magnitude or direction of the residuals. In a model with high absolute bias, like Florida's one-year projection model, on average projections would be more accurate if we subtracted 100 points from them all. In a model with high relative bias, projected scores of 500 may be systematically too low, whereas projected scores of 2500 may systematically be too high. Model relative bias can be assessed by plotting the relationship between projected scores and their residuals.

In order to assess model relative bias, Figure B.1 through Figure B.10 in Appendix B: Residuals plot projected scores vs. residuals for each of the five models' one-year and two-year projections. In a well specified model (one without relative bias) these data should appear random and without any systematic linear or curvilinear trend. As an
example, shown below in Figure 4.3 is a plot of Arkansas’ two-year projected scores (x-axis) vs. residuals (y-axis). A clear pattern exists in these data (\( \rho = .82 \)), an upward linear trend suggesting high relative bias and therefore a misspecified model. As students’ projected scores increase, so do their residuals. Under Arkansas’ model, the lower students’ projected scores are, the more likely that their projected scores are too low. Arkansas’ model has high relative bias.

**Figure 4.3 Plot of Residuals vs. Projected Values for Arkansas’ Two-year Projections (Student Level)**

Residual plots for each state in this study are presented in Appendix B: Residuals. Similar to the plot for Arkansas, there is high relative bias observed in the plots for Arizona’s, Florida’s, and North Carolina’s one-year and two-year projection models, where a clear positive relationship exists between projected scores and residuals. In contrast, the plot of Tennessee’s model’s projected scores and residuals shows little evidence of a relationship between these two variables. Tennessee’s model does not
show strong signs of relative bias because its projections are based on the typical growth trajectories of other students.

**Summary (Student Level)**

The first research question asked “How accurate are projection models at the student level?” As applied to the study district, the best proposed models accurately project proficiency one year into the future 82 percent of the time and the worst proposed model accurately projects proficiency 70 percent of the time. The best proposed model accurately projects proficiency two years into the future 76 percent of the time, and the worst model accurately projects proficiency 61 percent of the time. That is, the five unique models examined behave differently. Models vary in terms of their overall accuracy as well as in terms of the types of errors they tend to make. Florida’s model tends to project that students will perform better than they actually do; in contrast, Tennessee’s model tends to project students will do worse than they actually do. Although Tennessee’s model makes somewhat biased projections, it is the most accurate overall because its projections are the most precise compared to the other models. Tennessee’s model shows significantly less variation in how far off the projected scores are from the actual scores.

However, regardless of which model one chooses, simply projecting that students will remain at the same proficiency status is more accurate than any of the state proposed models, suggesting the models are not working very well at projecting future proficiency status. The implications of these findings will be discussed in the next chapter.
Research Question 2: How accurate are projection models at the school level?

Research question one focused on the accuracy of individual-level projections. Research question two naturally extends question one by asking “what happens when the results of projection models are aggregated to the school level?” This extension of the analysis is critical given that under the GMPP projection models are used to calculate the percentage of students within a given school that are on track to become proficient. This percentage is used to determine whether a school is making adequate yearly progress (AYP).

Although proficiency projections are made at the individual-level, for accountability purposes results are aggregated to the school level to determine the percentage of students within a given school who are on track to become proficient. In the previous section, one measure of the accuracy of the individual-level projections was to compare projected scores with observed actual scores. Similarly, in this section model accuracy is assessed by comparing the percentage of students in each school who were on track to become proficient to the percentage of students who actually became proficient. However, unlike individual-level accuracy, school-level accuracy is not assessed by claiming that a school-level projection is right or wrong. If a model projected that 70 percent of students in a school were on track to become proficient by 6th grade and 69 percent actually became proficient, this could be viewed as a fairly accurate projection, not as wrong because it was off by 1 percent. As a result, instead of viewing the projected percent proficient as either right or wrong, accuracy is assessed by subtracting

---

27 Some states consider the percentage of students within a given school who are either proficient or are on track to become proficient. In these analyses I focus on trying to project the percentage of students in each school who, based on the projection models, are on track to become proficient.

87
the observed percent proficient from the projected percent proficient. The difference is represented in Figure 4.4 as the school-level residuals, plotted for each state’s model’s \textit{one-year} projections. If a model’s school-level residuals are biased and/or imprecise the projections can be seen as inaccurate.

Only those schools with projected scores for 30 or more students’ were included in these analyses. A total of 96 schools were included for Arizona’s, Arkansas’, Florida’s, and North Carolina’s model. A total of 53 schools were included for Tennessee’s model. For an explanation as to why Tennessee’s sample size is smaller than the other states’ sample sizes, see Chapter 3: Data and Methods.

\textbf{Figure 4.4 Box Plot of Residuals from One-year Projections (School Level)}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.4}
\end{figure}

\footnotesize
28 A minimum n-size of 30 was chosen since this is a fairly common minimum n-size under NCLB.

29 Sensitivity analyses were conducted to ensure that results were not influenced by sample size. Findings remained the same when analyses were calculated using Tennessee’s subsample of students only.
In Figure 4.4 (above) the y-axis represents the school-level residuals, where .25 signifies that the projected percent proficient was 25 percent greater than the observed percent proficient. The “+” in the boxplot represents the mean residual, the horizontal bar in the middle of the box, the median. The bottom and top of the box represent the lower and upper quartile respectively. Ideally, residuals are close to or equal to zero, denoting that the projected percent proficient is close to or equal to the percentage of students who became proficient. The farther the mean and median are from zero, the more absolute bias the model demonstrates – suggesting the model systematically over or underestimates the projected percent proficient.

**Absolute Bias (School Level)**

Arizona’s, Florida’s, and Tennessee’s models show significant absolute bias at the school level, with Arizona, and Florida typically overestimating the percent proficient and Tennessee typically underestimating the percent proficient. Florida’s model is so dramatically biased that in 95 out of 96 schools it overestimates the percentage of students who will achieve proficiency. North Carolina’s one-year projections do not appear to be systematically overestimated or underestimated. Arkansas school-level one-year projections are slightly overestimated.

Figure 4.5 (below) depicts school-level residuals for the two-year projections. Much like the individual-level projections, the aggregated results show that the spread of the residuals increases when predictions are made two years into the future rather than one year into the future. All five models demonstrate some degree of absolute bias when making two-year projections. Arkansas’, Florida’s, and North Carolina’s models tend to
overestimate the percentage of proficient students, whereas Tennessee's and now Arizona’s models tend to underestimate the percentage of proficient students. The mechanism behind the opposing results for Arizona's one-year and two-year projections is described in detail in Appendix C: Arizona’s Special Case.

**Figure 4.5 Box Plot of Residuals from Two-year Projections (School Level)**

Table 4.3 provides basic descriptive statistics of the models’ school-level residuals.
Table 4.3 Descriptive Statistics for Projection Model Residuals (School Level)

<table>
<thead>
<tr>
<th>State</th>
<th>N⁰</th>
<th>Mean Deviation</th>
<th>5th Percentile</th>
<th>Median</th>
<th>95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona</td>
<td>96</td>
<td>11%</td>
<td>-29%</td>
<td>13%</td>
<td>48%</td>
</tr>
<tr>
<td>Arkansas</td>
<td>96</td>
<td>5%</td>
<td>-12%</td>
<td>5%</td>
<td>21%</td>
</tr>
<tr>
<td>Florida</td>
<td>96</td>
<td>24%</td>
<td>6%</td>
<td>23%</td>
<td>47%</td>
</tr>
<tr>
<td>North Carolina</td>
<td>96</td>
<td>2%</td>
<td>-15%</td>
<td>0%</td>
<td>23%</td>
</tr>
<tr>
<td>Tennessee</td>
<td>53</td>
<td>-12%</td>
<td>-24%</td>
<td>-11%</td>
<td>-2%</td>
</tr>
</tbody>
</table>

One Year Projections

<table>
<thead>
<tr>
<th>State</th>
<th>N⁰</th>
<th>Mean Deviation</th>
<th>5th Percentile</th>
<th>Median</th>
<th>95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona</td>
<td>99</td>
<td>-23%</td>
<td>-59%</td>
<td>-33%</td>
<td>61%</td>
</tr>
<tr>
<td>Arkansas</td>
<td>99</td>
<td>6%</td>
<td>-18%</td>
<td>5%</td>
<td>32%</td>
</tr>
<tr>
<td>Florida</td>
<td>99</td>
<td>15%</td>
<td>-10%</td>
<td>14%</td>
<td>42%</td>
</tr>
<tr>
<td>North Carolina</td>
<td>99</td>
<td>9%</td>
<td>-16%</td>
<td>8%</td>
<td>37%</td>
</tr>
<tr>
<td>Tennessee</td>
<td>57</td>
<td>-20%</td>
<td>-35%</td>
<td>-19%</td>
<td>-6%</td>
</tr>
</tbody>
</table>

Two Year Projections

N⁰: number of schools where the number of individuals is > 30.

The mean and median provide a measure of the central tendency of the residuals. They are generally very similar in value because the distribution of the residuals tends to be symmetric (Arizona’s two-year projection is the exception and is explained in detail in Appendix C: Arizona’s Special Case). Notably, the mean/median values are typically not close to zero (with the exception of North Carolina’s one-year projection). Consider the meaning of the fact that Tennessee’s one-year projection model has a school-level median residual of -11 percent: this implies that over half of the projected percent proficient estimates were at least 11 percentage points too low. Using Florida’s model, over half of the projected percent proficient estimates were at least 23 percentage points too high. These findings stem from the models’ biases, which, in part, lead to their inaccuracy.

Another way to examine the school-level projections is to consider their precision as measured by the standard deviation of the residuals. The residuals are generally
normally distributed and, as such, approximately 95 percent of the observed residuals lie within 1.96 standard deviations of the mean. The larger the standard deviation of the residuals is, the less precise the model. Comparing the standard deviation of the residuals for one-year and two-year projections. Table 4.3 demonstrates that two-year projections are less precise than one-year projections. In the best case scenario (Tennessee), 95 percent of residuals lie within ± 14 percent of the mean for one-year projections, and within ± 18 percent of the mean for two-year projections. Under Florida’s model, 95 percent of residuals lie within ± 22 percent of the mean for one-year projections, and within ± 32 percent of the mean for two-year projections. Such high levels of model imprecision suggest that even if the models did not show high levels of absolute bias, they still would not be precise enough to make meaningful projections.

Relative Bias (School Level)

Similar to the analysis conducted at the individual level, exploring the relationship between the projected percent proficient and the school-level residuals can inform us regarding where the models tend to be inaccurate. Plots of one and two-year projected percent proficient vs. school-level residual percent proficient can be found in Appendix B: Residuals, Figure B.10 through Figure B.20. Below, Figure 4.6 is an example plot for Arkansas’ one-year projections.
These plots are quite similar for Arkansas’, Florida’s and North Carolina’s one-year and two-year projections, showing a moderate correlation (ranging from .33 to .39) between the residuals and the projected percent proficient. For these three states’ models, knowing what percentage of students were deemed on track to become proficient provides information regarding either the magnitude of the residuals, or whether the projection is likely too high or too low. This indicates that these models show relative bias; therefore, knowing the projected percent proficient, one could systematically adjust this projection in a linear fashion in order to increase accuracy. Systematic error in the residuals implies that these models are misspecified. In contrast, Tennessee’s model residuals are virtually uncorrelated with the projected percent proficient. Although Tennessee’s projected percent proficient estimates are systematically too low, knowing
the projected percent proficient does not provide any additional information regarding the magnitude of the residuals. Tennessee’s model shows absolute bias, but not relative bias.

Plots of the projected percent proficient vs. the residuals are quite unique for Arizona’s model. One-year projections show a strong relationship \((r = .88)\) between these two variables, indicating high relative bias. Schools where the projected percent proficient is high tend to have highly overestimated scores. Schools where the projected percent proficient are low tend to have highly underestimated scores. This model misspecification could partially explain inaccuracies in the projections.

The most surprising scatter plot of the projected percent proficient vs. residuals is for Arizona’s two-year project model, shown in Figure 4.7.

**Figure 4.7 Plot of Residuals vs. Projected Values for Arizona’s Two-year Projection Model (School Level)**

![Plot of Residuals vs. Projected Values for Arizona's Two-year Projection Model (School Level)](image)

At first glance, this plot may appear to be the result of an error, but it is not. In this particular instance of Arizona’s two-year projection model, all schools were
projected to either have 0 percent proficient or 100 percent proficient. While unlikely, this is an artifact of Arizona’s projection methodology. A description of why this occurs is provided in Appendix C: Arizona’s Special Case. Not only does this plot indicate that Arizona’s model is biased, but this finding raises serious concerns regarding Arizona’s methodology.

**Summary (School Level)**

The second research question in this paper asked “How accurate are projection models at the school level?” The simple answer is that they are not very accurate. It is difficult to say which model is the “best” at projecting school-level percent proficient, because this is in part dependent on the type of error one prefers to make. What is clear is that regardless of model selection, even for one-year projections, more than a quarter of all schools’ projected percent proficient were off by 10 percent or more. This occurs largely because the models are biased and imprecise. Even the most precise model (Tennessee’s one-year projection model) makes school-level projections that are off by at least 15 percent a quarter of the time. The implications of these findings will be discussed in the next chapter.

**Research Question 3: Are growth expectations realistic?**

One of the potential benefits of the GMPP is that it allows schools several years to bring students up to proficiency. In contrast, originally NCLB only recognized actual proficiency, regardless of an individual’s progress toward proficiency. This new system may be an improvement over the old one because it may create more realistic growth
expectations for students. Consider a student who scores an entire year below grade level in 3rd grade. Under the NCLB status model a school only receives credit for this student in the following year if she gained enough to achieve proficiency on the 4th grade test. In contrast, under the GMPP a school has three (or four) years to bring this student up to proficiency. As such, instead of having to make two years worth of gains in a single year, now this student is required to make approximately one and a third years worth of gains for three years straight. Where it might not have been feasible for this student to make the extraordinary gains necessary to achieve proficiency by 4th grade, it may be reasonable to expect her to make the gains required to become proficient by 6th grade.

One suggestion for assessing the realism of expectations is what Linn (Linn, 2005) calls an “existence proof.” Linn proposes that we should at least be able to find evidence that the goals have been achieved by someone, somewhere in the past. Linn applies this existence proof criterion to school-level data in order to demonstrate that 100 percent proficiency by 2013-2014 is not realistic under NCLB’s status model. Here, I look to apply a somewhat similar existence proof to the gains required for individual students to become proficient under NCLB and the gains required to become proficient under the GMPP.

I begin by separating students into deciles based on their initial achievement levels (i.e. 2002, 3rd grade achievement levels). Table 4.4 below provides descriptive information for each of the ten deciles.
Table 4.4 Descriptive Statistics by 3rd Grade Decile

<table>
<thead>
<tr>
<th>Decile</th>
<th>n^a</th>
<th>3rd Grade Developmental Scale Score</th>
<th>Median Gains Required to Become Proficient by:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>Median</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>375</td>
<td>768</td>
</tr>
<tr>
<td>2</td>
<td>1012</td>
<td>907</td>
<td>986</td>
</tr>
<tr>
<td>3</td>
<td>1036</td>
<td>1050</td>
<td>1101</td>
</tr>
<tr>
<td>4</td>
<td>992</td>
<td>1157</td>
<td>1196</td>
</tr>
<tr>
<td>5</td>
<td>971</td>
<td>1235</td>
<td>1268</td>
</tr>
<tr>
<td>6</td>
<td>1055</td>
<td>1305</td>
<td>1337</td>
</tr>
<tr>
<td>7</td>
<td>946</td>
<td>1374</td>
<td>1404</td>
</tr>
<tr>
<td>8</td>
<td>1009</td>
<td>1439</td>
<td>1476</td>
</tr>
<tr>
<td>9</td>
<td>1005</td>
<td>1522</td>
<td>1568</td>
</tr>
<tr>
<td>10</td>
<td>981</td>
<td>1633</td>
<td>1726</td>
</tr>
</tbody>
</table>

*Decile sizes are unequal because deciles were created such that students with the same scale score were placed in the same decile (always the lower decile).

Note that the 4th and 6th grade proficiency cut scores are 1444 and 1692 respectively.

Students with low 3rd grade achievement scores will need to make significantly larger gains in order to become proficient by 4th grade compared to their initially high achieving counterparts. For example, students who scored in the bottom 10th percentile (decile 1) on the 3rd grade achievement exam must make median gains of 676 points in order to become proficient by 4th grade. In contrast, students who scored in the top three deciles on the 3rd grade exam do not need make any gains at all in order to be proficient in 4th grade.

Once students have been separated into deciles, I next calculate the percentage of students in each decile who were able score above the 4th grade proficiency cut score the following year (2003). In this way I am able untangle whether any students whose initial achievement levels were very low managed to make sufficient learning gains to

---

30 This percentage is not exactly the same as the percentage of students who were actually proficient in 2003 because some students were retained in grade, some students moved up a single grade, and some students skipped a grade. The overall conclusions are unaffected by this distinction. For simplicity, I will use looser language from this point forward. Numbers in Figure 4.8 and Figure 4.9 represent the percentage of students who scored above the 4th and 6th grade cut scores regardless of students' actual grades.
achieve proficiency under the requirements of NCLB. That is, does any evidence exist that the required learning gains under NCLB’s status model have been achieved by initially low performing students?

In Figure 4.8 (below) the x-axis represents the decile in which students scored on the 3rd grade mathematics exam in 2002. The y-axis depicts the percentage of students who scored at or above the 4th grade proficiency cut score in 2003. The data point farthest to the left tells us that two percent of those students who scored in the bottom 10th percentile (decile 1) in 3rd grade managed to make sufficient gains in order to become proficient in 4th grade. Similarly, the second data point tells us that five percent of those students who scored in the 11th to 20th (decile 2) percentile in 3rd grade were able to achieve proficiency by 4th grade. While these numbers do not suggest that NCLB’s status measure completely fails the existence proof, they demonstrate that students who score in the bottom of the distribution (in this study district) in 3rd grade are highly unlikely to become proficient in a single year. For these students the gains required to become proficient in a single year do not appear to be realistic. In contrast, students who score above the 80th percentile (deciles 9 and 10) in 3rd grade are more than 90 percent likely to achieve proficiency in 4th grade. Clearly (and unsurprisingly) there is a strong relationship between students’ 3rd grade achievement score and their likelihood of achieving a proficient score on the 4th grade exam. A school with many initially low performing students is extremely unlikely to be able to bring them up to proficiency. For example, if teachers/schools were all equally effective, then a school with just three students in the bottom decile would only have approximately a one in one hundred thousand (.02\(^3\)) chance of bringing all three of these students up to proficiency.
Under the GMPP, students have until 6th grade to become proficient. The hope is that by giving students a few extra years to become proficient, the goal may now be more realistic. This may be especially true for those students whose initial achievement levels were particularly low, because now they have several years to make up for their initial deficit. Still, as Table 4.4 (above) demonstrates, students with low 3rd grade achievement scores will need to make significantly larger gains in order to become proficient by 6th grade compared to their initially high achieving counterparts.

Figure 4.9 (below) plots the percentage of students, by 3rd grade decile, who become proficient by 6th grade. The point to the far left tells us that three percent of those students who scored in the bottom 10th percentile (decile 1) in 3rd grade managed to make sufficient gains in order to become proficient by 6th grade. The second data point tells us that five percent of those students who were in the 11th to 20th percentile (decile 2) in 3rd grade were able to achieve proficiency by 6th grade. This suggests that, much like the gains required under NCLB, the gains required by the GMPP are achieved by a very
small proportion of initially low performing students and therefore may not be realistic. Given the strength of the association between a student’s likelihood of passing the state exam and her initial achievement level, it is clear that schools with initially low performing students are at a huge disadvantage under a status model or a projection model.

**Figure 4.9 Percent Proficient by 6th Grade (Required under GMPP)**

For a school to receive credit for initially low performing students in Florida, NCLB requires a single year of extremely large gains whereas the GMPP requires three years of fairly large gains. These analyses demonstrate that both of these required levels of growth are about equally as uncommon on Florida’s mathematics assessment. However, when interpreting Figure 4.8 and Figure 4.9 it should be noted that 49 percent of students in the sample scored above the 4th grade proficiency cut score, whereas only 41 percent of students in the sample scored above the 6th grade proficiency cut score. As such, the percent proficient in Figure 4.9 tends to be eight points lower than the percent proficient in Figure 4.8. It is therefore somewhat reassuring that the first four deciles all
show gaps of less than eight percentage points between the 4th and 6th grade observed proficiency. This suggests that giving schools two extra years to bring students up to proficiency slightly helps reduce dependence on students’ initial achievement levels. Still, since the shapes of the curves in these two figures are fairly similar, giving schools two additional years to bring their students up to proficiency is unlikely to make the learning gain goals much more realistic.

Of considerable note is that the above findings are largely influenced by the proficiency cut scores on the 4th and 6th grade exams. If the cut scores were lowered or raised (i.e. if the state exam was easier or harder to pass) one would expect the entire curves in Figure 4.8 and Figure 4.9 to shift up or down. Although Florida’s growth goals appear unrealistic for initially low performing students, this may not be the case in other states with lower proficiency cut scores. However, in all states, a student’s likelihood of passing the state exam is going to be strongly associated with her initial achievement level.

Summary – Realistic Goals

In Florida, how realistic the learning gain goal is for any individual student is highly dependent upon their initial achievement level under both NCLB and the GMPP. This is because at their core, both NCLB and the GMPP are “variable gains models” – that is, models that require students to achieve different amounts of learning gains depending upon their initial achievement levels (i.e., lower initial performance requires larger gains). Consequently, it should not be surprising that if the goals of NCLB are unrealistic, then the goals of the GMPP are likely to be unrealistic as well.
For initially low performing students, the gains required to attain proficiency by 4\textsuperscript{th} grade are infrequently attained. Similarly, for initially low performing students, the gains required to attain proficiency by 6\textsuperscript{th} grade are infrequently attained. The attainability of the goals of NCLB and the GMPP are both highly dependent upon students’ initial achievement levels, something which may be beyond the control of the school.

Research Question 4: How do status, projection, and value-added measures of school performance compare to each other?

Strong criticism of NCLB’s aggregated school-level status and safe harbor measures of AYP led to the creation of the GMPP. It may be the case that in order to improve measures of school performance it is necessary to track individual students over time, using student-level growth models rather than simpler school-level status models or school-level change models. However, not all student-level growth models are necessarily the same. The projection models used as part of the GMPP require students to make sufficient learning gains such that they are on track to become proficient in the near future. Consequently, these models require different levels of individual gains depending upon students’ initial achievement levels. Using projection models schools are judged, in part, based on the initial status of their students. In contrast, value-added models compare the average gains of students in a school to the average gains of students in other schools, controlling for students’ initial achievement levels.\textsuperscript{31} Using value-added models, schools are not judged based on the initial achievement levels of their students;

\textsuperscript{31} Although this is not precisely the case, average individual gains scores have a .98 correlation with the value-added scores in this dataset, so they can be thought of as essentially the same.
however, these models do not require students to reach a proficiency standard. In this section I compare these three approaches to assessing school performance: status (NCLB), projection models (GMPP), and value-added models (used in some state accountability systems).

One way of comparing these three approaches to measuring school performance is to consider schools’ relative rankings under each method. In this way it can be assessed whether the three different methods are providing unique or redundant information.

Schools’ performances using each method were computed as follows:

1. Percent proficient (NLCB) – Under NCLB’s status model schools are judged (in part) based on the percent of students who are proficient on the state exam. In these analyses each school’s percent proficient measure represents the percentage of 5th grade students who passed the state exam in 2004.

2. Percent on track to become proficient (GMPP) - Under some state’s GMPP proposals, schools are judged based on the percentage of students who are on track to become proficient on the state exam. Since one-year projections are more accurate than two-year projections, I look at the percentage of students in each school who, in 5th grade 2004, were on track to become proficient by 6th grade 2005.33

32 5th grade was selected for two reasons: first, in previous analyses the author established the accuracy of different state projection models in 5th grade. Second, the value-added model used in this analysis benefits from the use of more years of longitudinal data, so it is advantageous to use at least three years of available data.

33 Results are presented using Arkansas’ model, since its one-year projection model is fairly accurate and unbiased at the school level. Florida’s and North Carolina’s models were used in sensitivity analyses.
3. *Value-added score* – School value-added scores are computed using the three-year historical records of 5th grade students in 2004. The value-added methodology is described in Chapter 2: Background and Literature by Equation (2-2). In this application the $\theta$’s represent school random effects rather than teacher random effects, and schools’ performance levels are indicated by their standardized random effects. For simplicity, schools’ value-added scores can be thought of as approximately equal to the average individual scale score gains from 2003 to 2004, just on a different scale.

Table 4.5 below shows the correlations between schools’ percent proficient (in 5th grade), percent on track to become proficient (in 5th grade), and value-added scores (for the 5th grade).

<table>
<thead>
<tr>
<th>Percent Proficient (NCLB)</th>
<th>Percent on Track to become Proficient (GMPP)</th>
<th>Value-Added Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Proficient (NCLB)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Percent on Track to become Proficient (GMPP)</td>
<td>0.99</td>
<td>-</td>
</tr>
<tr>
<td>Value-Added Score</td>
<td>0.47</td>
<td>0.54</td>
</tr>
</tbody>
</table>

All correlations significant at the $p < .0001$ level.
* Using Arkansas' projection model

The correlations reveal considerable similarities among measures (evidenced by significant positive correlations). Most notably, the correlation between the school-level percent proficient and percent on track to become proficient is extremely high ($r = .99$).
This suggests that schools relative rankings under a status model (NCLB) and a projection model (GMPP) are nearly identical and these models are providing essentially redundant information. Since correlations are scale-invariant it is possible that some schools could make AYP based on a projection model that otherwise would not have made AYP based on a status model. However, such a high correlation implies that differences between schools making AYP under a status model vs. a projection model are likely a function of the difficulty of being deemed proficient in the status year vs. the projection year (i.e., the difficulty of the 5th grade exam compared to the 6th grade exam).

The above analyses may be affected by the particular inaccuracies of the selected projection model. That is, school-level percent on track to become proficient are shown using Arkansas’ projection model (due to its relatively high accuracy and low bias). Sensitivity analysis show that, using Florida’s or North Carolina’s projection model, this correlation does not drop below .90, suggesting that this finding is fairly robust to model selection.

Another way to consider the similarity between the school-level percent proficient and the school-level percent on track to become proficient, independent of which projection model is selected, is to compare the percentage of proficient 5th graders in each school to the percentage of those same students who became/remained proficient in 6th grade. In other words, if a projection model was perfectly accurate at projecting future proficiency, then how would the school-level percent proficient compare to the percent on track to become proficient? The correlation between these two measures is .91.34 This suggests that if we were able to create a projection model that was 100 percent

34 The correlation between school-level percent proficient in 4th grade and the percentage of students who became proficient by 6th grade was .90.
accurate we would find schools’ relative rankings to be very highly correlated to their rankings obtained under the old status model. Model selection does not appear to be influencing these results; rather, status and projection are mostly redundant measures.

While school-level status and projection models yield extremely similar results, also of considerable note in Table 4.5 (above) is that the correlations between a) value-added scores and percent proficient and b) value-added scores and percent on track to become proficient, are both only moderate. The modest correlation between percent proficient and value-added scores ($r = .47$) is not surprising and suggests a significant difference in the information conveyed by these metrics. The modest correlation between the percent on track to become proficient and value-added scores ($r = .54$) may come as a surprise to those who think all growth models are the same. It should be clear that projection models and value-added models are not only theoretically different, but actually result in significant differences in their assessments of schools’ relative performances. Whereas projection models yield nearly identical assessments of schools’ relative performance compared to a simple status model, value-added models provide different information from NCLB’s status model and projection models.

**Summary – Comparing Measures of School Performance**

Percent proficient (NCLB’s status) and the percent on track to become proficient (GMPP’s projection models) measure extremely similar constructs. Value-added scores are substantively different from both status and projection models. Interpretation of these results will be offered in the next chapter.
This chapter primarily provided a numerical description of the calculations and results used to address the research questions. In the subsequent chapter of this paper, interpretation of the practical significance of these findings is offered, as well as suggestions for improving the state projection models.
Chapter 5: Discussion

Projection Model Accuracy

The first and second research questions addressed in this work asked about the accuracy of the state proposed projection models. Applying the method behind each state’s model to extant data from the study district, the accuracy of projections at the individual level was measured. Analyses demonstrated that accuracy levels varied greatly depending upon the projection method, and that making projections two years into the future was significantly more challenging than making projections a single year into the future.

For school accountability purposes, the accuracy of projections at the student level is essentially irrelevant since states only consider the aggregated results of these models when determining whether schools make AYP. However, inaccuracy at the student level is still important for two reasons. First, several states (Arizona, Arkansas, Florida, Ohio, Tennessee) plan to report the results of individual projections to students, parents, teachers, and/or schools (Arizona Department of Education, 2007b; Arkansas Department of Education, 2006; Florida Department Of Education, 2006; Ohio Department of Education, 2006; Tennessee Department of Education, 2006). Second, understanding the way in which state models function at the individual level provides useful information to help understand why the models work the way they do at the aggregated school level.

When making individual-level projections one year into the future, the most accurate models (Tennessee and Arkansas) were accurate 82 percent of the time, the least accurate model (Florida) was accurate only 70 percent of the time. Under Florida’s
model more than one in four students who would have been labeled on track to become proficient did not become proficient. In states (like Florida) which intend to report results at the individual-level, this degree of inaccuracy is unacceptable. Such inaccurate individual-level reporting will result in false expectations for students, parents, and teachers. Under a model like Florida’s, too many children who are on the path towards being “left behind,” will be misidentified as “on track.” Such high levels of inaccuracy have the potential to result in an inefficient distribution of resources and misplaced efforts to improve student achievement. In addition, such frequent misinformation undermines the credibility of the entire accountability system. Finally, misplaced blame will likely be placed on the final grade (6th or 7th) when students are finally deemed not proficient after years of being on track to become proficient.

While individual-level reporting may be a bad idea using the least accurate model (Florida), one might speculate that it is more reasonable to use one of the more accurate models (Arkansas or Tennessee)? Using the more accurate models, on this dataset, provided no evidence that projections are any more accurate than simply assuming all students will remain at their current proficiency status. Given this fact, it is difficult to justify using these models for individual-level reporting, unless the models are improved or shown to perform better when applied in other states.

At the school level, the models do not fare any better. All models demonstrate bias, systematically overestimating or underestimating the percentage of students who are on track to become proficient. Perhaps of greater concern is that even if we ignore the bias of these models, there is still large variation in the accuracy of school-level projections. If we were to attempt to create 95 percent confidence intervals around the
projected percent proficient, they would have to span at least 20 percentage points using the most precise model, even after adjusting for model bias and making projections only a single year into the future. The fact that the best we can do when making school-level projections is to claim that, for example, between 30 and 50 percent of students in a school are on track to become proficient, is unacceptable. If we do not use confidence intervals and simply rely on point estimates, then large numbers of schools will be rewarded or sanctioned based on measurement error.

Why are the projection models not very accurate? How can we do better? In the subsequent paragraphs I briefly review the benefits and drawbacks of each state model. Then, suggestions for improving the models are offered. I begin with Florida’s simple linear model and finish with Tennessee’s complex statistical model.

**Florida**

Florida’s projection model relies on the underlying assumption of linear growth. The main benefit of this approach is its simplicity and transparency for students, parents, and school staff. However, this benefit is offset by the fact that projections under this model are highly inaccurate. This inaccuracy is particularly apparent on Florida’s mathematics assessment, where growth is typically curvilinear. The consequence of using a linear model when most students’ growth trajectories are curvilinear is that the model often projects students are on track to become proficient when they are not. As a result, under Florida’s linear model many schools that under serve their students (by NCLB standards) will not be identified as in need of improvement.
Florida’s model’s inaccuracy is not due entirely to the linearity assumption. In addition to the faulty linear assumption, Florida’s model does not account for measurement error in student test scores, which results in distorted gain scores. Because of measurement error, those students whose observed time one scores are at the extreme ends of the scoring distribution have true scores that, on average, are closer to the mean. As a result, gains calculated based on extreme scores are likely to be distorted. Consequently, those students with extremely low scores at time one are more likely to have overestimated gains; students with extremely high time one scores are more likely to have underestimated gains. Ignoring this fact leads to less accurate projections of students’ future performances. This can loosely be thought of as a problem of regression towards the mean. This is not precisely the case because if the variance in students’ test scores increases over time (i.e., scores fan outward), then extreme scores may not regress towards the mean, although they will still be distorted.

To improve their growth measure under the GMPP, Florida should consider revising the way which they set growth targets in order to reflect the shape of typical growth patterns on their state exams. In addition, Florida should use a measure that accounts for measurement error and the associated distorted gain scores.

Arizona

Arizona’s projection model also relies on the underlying assumption of linear growth; however, unlike Florida’s model, their model attempts to account for regression toward the mean. Nonetheless, Arizona’s model was still highly inaccurate when applied to Florida’s mathematics assessment. However, there is evidence that their model
performs better when applied to Arizona’s own state exam. Arizona was the only state whose GMPP proposal demonstrated that they tested the accuracy of their model on their own historical student achievement data. According to their state proposal, one-year projections were accurate 80 percent of the time using their model (compared to 72 percent accuracy when applying their state’s method to data from the study district in Florida) (Arizona Department of Education, 2007a). This may partly reflect the fact that Arizona’s exams had an average pass rate of approximately 70 percent under their calculations. Consequently, projected proficiency is easier to ascertain on Arizona’s exams compared to on the study district’s 6th grade exam which had a 41 percent pass rate.35 While these results are somewhat reassuring, unfortunately Arizona did not report the accuracy of their two-year projections, or the accuracy of their aggregated school-level results.

Arizona’s model still has several fundamental problems. Like Florida, Arizona sets linear growth targets. If typical student growth trajectories on Arizona’s state exam are linear, then the linear assumption may not be a problem. However, based on the shape of the mathematics and reading proficiency cut scores on Arizona’s state exam (see Figure 5.1 below), the assumption of linearity does not appear to be tenable. On their state assessment changes in proficiency cut scores are constant between 3rd and 5th grade, but they drop in 6th grade and then remain constant from 6th to 8th grade (in both mathematics and reading) (Arizona Department of Education, 2007b). While proficiency cut scores do not necessarily align with typical student growth trajectories, this provides evidence that Arizona’s linear assumption is likely faulty.

35 Projecting all students to become proficient would yield 70 percent accuracy in Arizona, whereas projecting all students to fail in the study district would only have been 59 percent accurate.
Arizona’s model attempts to adjust for distortion of gain scores due to measurement error or “regression to the mean.” This is a laudable goal, given that this known phenomenon is likely to bias projections. However, Arizona’s method for accounting for regression toward the mean may be flawed. Recall from Chapter 3: Data and Methods, rather than comparing observed year two scores to the growth target, Arizona compares the lower bound of predicted year two scores (from a regression model) to the growth target. The exact logic behind this approach is not clearly articulated in the state proposal, but there are a few concerns with this approach.

Regression towards the mean does not imply that the best guess of a student’s true gain score is equal to the population mean gain score. The special case where this would be true is if gain scores were 100 percent measurement error. Although Arizona’s model does not precisely assume a student will make gains equal to the mean gain

---

36 Although often referred to as “regression to the mean,” I choose to use the more accurate “regression towards the mean” to highlight this mistaken interpretation.
score, their model essentially ignores students’ time two scores (and thus their actual gains) in favor of calculating gains based on the regression predicted time-two scores. Under Arizona’s model, two students in the same school with the same time-one score and vastly different time-two scores will both have identical predicted time-two score. Thus, they will both be labeled the same with respect to being on track to become proficient. That is, two students in the same school with identical time-one score will both be labeled the same regardless of how disparate their observed gain scores are.

Given that students’ observed time-two scores are known, a more precise estimate of students’ “true” gains probably lies somewhere between their observed gains (what Florida’s model uses) and gains calculated by subtracting predicted time-two scores from observed time-one scores (what Arizona uses).

Another concern with Arizona’s approach is that two students in different schools with precisely the same time-one and time-two scores could receive different labels with respect to being on track to become proficient. This occurs because predicted time-two scores are obtained from a regression model that contains school fixed effects. While accounting for differences among schools may likely lead to more precise projections, it is unclear whether the GMPP intended to allow a growth model where two schools receive different amounts of credit for students with identical test scores. This is not necessarily a problem, but it is certainly something to consider.

In summary, Arizona should consider adjusting the way they set their growth targets (linearly) in favor of a method that takes into account the shape of typical student growth trajectories. Arizona should also consider a different approach to adjusting for

---

37 There model includes school fixed effects and a predictor variable, pretest score, which influence estimated gains.
the distortion of gain scores resulting from measurement error. Since gain scores tend to be fairly volatile, their approach is likely better than making no adjustment at all (e.g., Florida), but a better method would take more consideration of a student’s time two score. Finally, Arizona should consider the ramifications of using a school fixed effects regression model. Will they be able to justify to the public why two students with the exact same test scores receive different “growth” labels?

**North Carolina**

Like Florida’s and Arizona’s models, North Carolina’s model also assumes linear growth. However, North Carolina first converts all scores to a standardized metric. The standardized metric scores are like z-scores, where the means and standard deviations used to convert scale scores onto the common metric (called a C-scale) come from the standard setting year. This metric can be thought of as a “time-locked modified z-scale” (North Carolina Department of Education, 2006).

North Carolina’s method is versatile because it does not rely on the use of data from a vertically equated exam. However, North Carolina’s model still makes fundamental faulty assumptions. North Carolina’s model assumes linear growth on the C-scale. The model implicitly assumes that if a student scores 1.0 standard deviation below the mean in 3rd grade and .5 standard deviations below the mean in 4th grade, then she will likely score near the mean in 5th grade. Does this growth trajectory accurately reflect reality? Probably not. If this assumption were correct, there would be a high correlation between students’ year one to year two gains and their year two to year three gains. Using the data from the study district, there was a moderate negative correlation \( r \)
Students whose C-scale gains were positive between year one and year two tended to make negative C-scale gains between years two and year three. Students whose C-scale gains were negative between year one and year two tended to make positive C-Scale gains between year two and year three. This may reflect the fact that differences among students' C-scale gain scores are largely a result of measurement error and/or other one-time, non-persistent factors. While converting students' scores onto the C-scale may help alleviate the problems associated a curvilinear developmental scale, the assumption of linear growth on the C-scale does not accurately reflect typical learning trajectories of students.

Using extant data, North Carolina should test their model's underlying assumption of linear growth on the C-scale on their own state exam. While it is possible they will find the linear assumption to be more realistic on their own state assessment, it is doubtful due to the unreliability of gain scores as calculated using the C-scale.

**Arkansas**

Unlike Florida's, Arizona's and North Carolina's models, Arkansas' model does not assume linear growth. Instead, Arkansas' model assumes that students' learning trajectories will follow the shape of the proficiency cut scores. On a curvilinear developmental scale this assumption is likely to be more accurate than the assumption of linear growth. Still, proficiency cut scores do not necessarily reflect students' typical gain scores.

---

38 If all students' "true" gain scores were equal and observed differences among gain scores arose only because of measurement error, then the correlation between students' year one to year two gains and their year two to year three gains would be equal to -.50.
learning trajectories. For example, Table 5.1 below compares percent changes in proficiency cut scores to the average observed percent changes in scale scores on Florida’s Mathematics FCAT. Of the total change in proficiency cut scores between 3rd and 6th grade, 41 percent occurs between 3rd and 4th grade, 44 percent between 4th and 5th grade, and 14 percent is between 5th and 6th grade. However, as observed in the study district, of the total scale score gains made by actual students between 3rd and 6th grade, on average 32 percent occurred between 3rd and 4th grade, 64 percent between 4th and 5th grade, and 5 percent between 5th and 6th grade.39

<table>
<thead>
<tr>
<th>Change in Percent Change in Average Observed</th>
<th>Proficiency Cut</th>
<th>Proficiency Cut</th>
<th>Proficiency Cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>Score</td>
<td>Score</td>
<td>Score</td>
</tr>
<tr>
<td>3rd to 4th Grade</td>
<td>175</td>
<td>41</td>
<td>32</td>
</tr>
<tr>
<td>4th to 5th Grade</td>
<td>188</td>
<td>44</td>
<td>64</td>
</tr>
<tr>
<td>5th to 6th Grade</td>
<td>60</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>3rd to 6th Grade</td>
<td>423</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

On Florida’s mathematics FCAT changes in proficiency cut scores do not mimic the shape of typical student growth trajectories (in the study district), a fact that likely leads to less accurate projections. It is possible that on Arkansas’ own state exam proficiency cut scores align well with students’ typical growth trajectories (an empirical question), in which case Arkansas’ model may be more accurate on their state exam compared to its performance on Florida’s exam. That said, if the goal of an accountability system is to improve student performance, it is hoped that student growth trajectories change over

39 Note that these calculations reflect the average of the percent change, e.g. \( \frac{1}{n} \sum_{i=1}^{n} \frac{y_{1i} - y_{2i}}{y_{1i}} \), not the proportion of the average gains, e.g. \( \frac{1}{n} \sum_{i=1}^{n} \frac{y_{1i} - y_{3i}}{y_{1i}} \).
time. Consequently, even if student growth trajectories mimic the proficiency cut scores in the standard setting year, as time passes this may change.

Like other models discussed, Arkansas' model does not account for distortion of gain scores due to measurement error. Students whose gains are particularly large (or small) between years one and two are less likely to maintain their same rate of growth, even if that rate is not assumed to be linear.

Arkansas should provide evidence that the changes in their proficiency cut scores reflect typical patterns of student growth in their state. They should also consider an attempt to account for the distortion of gain scores.

**Tennessee**

Tennessee's model is the most sophisticated and versatile of the state proposed projection models. Taking a statistical approach which bases projections on observed historical patterns of student achievement data is more likely to produce accurate projections than the other approaches. Tennessee's regression based approach should account for the distortions of gain scores due to measurement error. Their method does not make any assumption about a pre-specified shape of growth trajectories, only that growth trajectories will be similar to what has been observed in the past.

One drawback of Tennessee's approach is its lack of transparency for the public. Tennessee's projection methodology can be described in general terms for public consumption; however, the details are complicated enough that most consumers will not fully understand how projections are calculated. This is not necessarily a major problem; as Sanders points out "Most everyone can use a cellular telephone, but virtually no one
knows, or needs to know, how to build the phone” (Sanders, 2000, p. 336). However, when the details necessary to replicate the methodology are not available, its rigor becomes largely untestable. As noted earlier in this dissertation, contact with the Tennessee Department of Education has not led to sufficient information necessary to replicate Tennessee’s approach with certainty.

In addition to its lack of transparency, my best attempt to replicate Tennessee’s projection model did not yield particularly impressive results in terms of the model’s accuracy at either the individual-level or the school level. The model’s mediocre performance may be a result of school-mean centering the training dataset. Whatever the reason, Tennessee’s model did not make projections which were any better than simply assuming that students would remain at the same proficiency status, ignoring growth all together.

Applying the five state proposed projection model methodologies to historical data from the study district helped illuminate many of the challenges associated with projecting students’ future proficiency status. While none of the models made accurate enough projections to instill confidence in their use as part of a high stakes accountability system, all models can potentially be improved upon in order to increase accuracy.

**How Can We Do Better?**

Two themes emerged in analyzing the assumptions of state proposed projection models. The first is the need for states to use projection models that match typical student growth patterns on their state exams. If learning trajectories on the state exam are curvilinear, then using a model which assumes linear growth is going to overestimate the
number of students who are making sufficient growth. Using historical data, each state can and should be required to test their model’s assumed growth trajectory.

The second emergent theme is that of distortion of gain scores. Anytime we take imperfect measures of a variable (i.e. any variable with measurement error / sampling variability) gain scores will be distorted. When measuring a hypothetical construct like “academic achievement,” measurement error is always present; consequently, students with extreme scores at one time point are more likely have distorted gain scores. Appropriate growth models should take this into account.

Tennessee’s model seems to deal fairly well with the above critiques, yet its projections are still not as accurate as desired. One way to improve Tennessee’s model is to change the goal of the model. Rather than using projection models to determine the black and white binary outcome “is a student on track to become proficient?” models could be used to estimate the probability that a student will become proficient.

As applied in the analyses in this paper, Tennessee’s model produced projected 6th grade scale scores. These projected scale scores were then compared to Florida’s mathematics 6th grade proficiency cut score, 1692. If the projected score met or exceeded the proficiency cut score, then students were deemed on track to become proficient. This approach does not distinguish a projected 6th grade score of 1900 from a projected 6th grade score of 1695. A student whose projected 6th grade score is 1900 may be 99 percent likely to become proficient by 6th grade, whereas a student whose projected 6th grade score is 1695 may only be 52 percent likely to become proficient by 6th grade. Given a student’s projected score and the standard error of the projection, one could
calculate the likelihood that a student will score at or above the proficiency cut score. Calculating these probabilities could serve two purposes.

First, for those states that desire to report projection results at the individual-level, these probabilities could be used rather than the “on track” or “not on track.” Reports could present the estimated likelihood that a student will become proficient given her learning trajectory. This way a student who is told that she is 52 percent likely to become proficient by 6th grade will not be surprised if she does not reach this goal. Distinguishing among students who are “on track” could reduce some of the negative impacts of false positives, since students, parents, and teachers would be aware of the fact that there is a chance a student will not become proficient or vice versa.

Second, this approach could help improve school-level projected percent proficient. Under the current model in order to determine the percentage of students who are on track to become proficient, states essentially add up the blunt raw number of students who are on track to become proficient and divide by the total number of students. This approach does not utilize all known information – that some students are projected to score well above proficiency, whereas others are on the borderline. A potentially more precise approach would be to sum up the estimated likelihood that each student will become proficient and then divide by the total number of students. This would distinguish among students who are almost certainly on track, those students who are on the borderline, and those students who have only a remote chance of becoming proficient (May & Baldwin, 2006). It is possible that such an approach would improve school-level projections that currently do not maximize the information provided by the individual-level projections.
Another option is to try other types of growth models. Tennessee's model might be more accurate if it did not school mean center the training data set. Alternatively, a mixed effects growth curve model might be more accurate. Using a randomly selected subset of students from the study district, a growth model with fixed effects for time and time squared, random effects for intercept, time and time squared, and an unstructured correlation matrix produced more accurate individual-level one-year projections than any of the state proposed models (projections were accurate 84.6 percent of the time). Experimenting with other statistical models will likely yield projections that are more accurate and less bias at both the individual and school levels. Still, it is uncertain whether model improvements will lead to school-level projections that are accurate enough to be useful in a high stakes accountability system. This is an area of needed future research.

Summary

As applied to historical data from the study district, the projection models currently used under the GMPP do not demonstrate impressive levels of accuracy at the individual-level or at the school level. Many state models rely on faulty assumptions when making their projections, so inaccuracy should not be surprising. Even using the more promising methods, which do not rely on as many strict assumptions, the measurement challenges associated with projecting students' future proficiency are fairly

---

40 This model was run in a fashion similar to Tennessee's model, using a training data set and a validation data set. Approximately 1,800 students were included in each data set. The statistical model was run 500 times using randomly selected training and validation data sets each time. The 84.6 percent accuracy represents the average overall accuracy (true positives plus true negatives) across all 500 runs. The model crashed when it was attempted using training and validation data sets of size 3,600.
significant. At a bare minimum, for projection models to be deemed useful they should be more accurate than assuming that all students will remain at the same proficiency status. Several of the models used in this paper, which demonstrated disproportionately high numbers of false positives, are simply letting schools off the hook for a few years, at which point in time those students who were supposedly on track to become proficient will likely fail.

Realistic Goals

The third research question assesses whether the individual-level gains goals of NCLB and the GMPP are realistic. Findings indicate that in the study district, the goals of both NCLB and the GMPP do not appear to be realistic for initially low performing students.

Previous work by Linn (2005) demonstrated that the NCLB goal of universal proficiency by 2013-2014 can not reasonably be expected to be attained at the school level. The research in this paper breaks the goal of attaining proficiency down to the individual-level, assessing the individual-level learning gains required under a status model. Although NCLB’s status measure does not explicitly calculate individual-level learning gains, a status model can be thought of as a variable gains model, whether or not gains are actually calculated. Under a status model, each student’s required learning gains are equal to the amount of learning the student must make in order to become proficient by the end of the year. As such, the magnitudes of required learning gains are completely dependent upon each student’s initial achievement level.
The analysis of NCLB’s goals demonstrated two facts. First, the probability of a student becoming proficient by 4\textsuperscript{th} grade is highly associated with the student’s 3\textsuperscript{rd} grade achievement level. Second, in the study district the probability of an initially low performing student becoming proficient by 4\textsuperscript{th} grade is extremely low. The first fact suggests that under a status model schools are largely being judged based on that which may be out of their control – students’ initial achievement levels. The second fact suggests that, for initially low performing students (in the study district), the goal of becoming proficient in a single year as required under NCLB’s status model is not realistic. While considering the status goal of NCLB from a variable gains perspective may be somewhat novel, these general findings are fairly unsurprising. However, the exercise is useful because thinking of status as a variable gains model provides insight into the results regarding the GMPP.

Just like NCLB’s status model, the GMPP’s projection models are also variable gains models. Under projection models each student’s required learning gains are equal to the amount of learning that student must make in order to become proficient several years into the future. Just like under NCLB’s status model, under the GMPP’s projection models the magnitudes of required learning gains are completely dependent upon each student’s initial achievement level. Given these facts, perhaps it should be unsurprising that analysis of the GMPP’s goals demonstrated the same two facts as the analysis of NCLB’s goals. First, the probability of a student becoming proficient by 6\textsuperscript{th} grade is highly associated with the student’s 3\textsuperscript{rd} grade achievement level. Second, in the study district, the probability of an initially low performing student becoming proficient by 6\textsuperscript{th} grade is extremely low.
Both status and projection models require “growth toward a standard.” Status requires students to become proficient immediately; projection allows students several years to become proficient. In these analyses both models do not appear to have realistic gains expectations for initially low performing students. Of students in the bottom 20 percentile in 3rd grade, fewer than five percent were able to reach proficiency by 4th grade. The same was true regarding becoming proficient by 6th grade. Simply moving from a status model to a projection model does not necessarily lead to an accountability system that has more realistic goals.

It should be noted that findings regarding whether the goals of NCLB and the GMPP are realistic are highly dependent upon state’s defined proficiency cut scores. In a state with lower proficiency standards than Florida’s proficiency standards the goals of both NCLB and the GMPP could be realistic for all students.

What may be of greater interest from the analysis of the gains goals is that students’ likelihood to become proficient under a status model and a projection model are both highly associated with students’ initial achievement levels. This result is crucial because some people look to the GMPP as a major change in accountability policy (Hershberg, 2005), one that moves away from holding schools accountable, in large part, for students’ initial achievement levels. These analyses unequivocally demonstrate that this is not the case. Under the GMPP schools are still judged by that which is largely out of their control – the initial achievement levels of their students. While this is unfair to schools, it is a necessity in order to uphold the core principles of NCLB -- bringing all students up to proficiency regardless of their initial achievement levels. If researchers, educators, and policymakers come to fully understand the implementation of growth
models under the GMPP, then we should expect that the exact same criticisms of NCLB’s status model will be invoked regarding the GMPP’s “growth towards a standard” model.

**Comparing Measures of School Performance**

The final research question looks to assess the similarities and differences among measures of school performance. Before examining the results, it is important to keep in mind that one critical reason for interest in using individual-level growth models to measure school performance is because of the belief that there might be low status high growth schools and high status low growth schools, both of which go unrecognized under the original NCLB accountability system (Hershberg, 2005). That is, in some schools students may make relatively large learning gains (i.e., high growth), yet few students may pass the year-end proficiency exam (i.e., low status) simply because students’ initial achievement levels were very low. These schools may be relatively effective (compared to other schools), even though they are not as effective at bringing students up to proficiency. Likewise, in some schools students may make relatively small learning gains (i.e., low growth), yet most students pass the year-end proficiency exam (i.e., high status) simply because students’ initial achievement levels were very high. These schools may be relatively ineffective (compared to other schools), even though they are effective at bringing students up to proficiency. Figure 5.2 provides a visual depiction of how an accountability system might consider both status and growth when assessing schools’ performances.
With Figure 5.2 in mind let us examine the relationships among the three measures of school performance.

Figure 5.3 (below) plots each of the study school’s status on the x-axis and their “growth” on the y-axis. In this figure Status is measured by the percentage of students in each school who were proficient in 5\textsuperscript{th} grade. “Growth” is measured for these same 5\textsuperscript{th} graders as under a projection model - the percentage of 5\textsuperscript{th} grade students who became proficient by 6\textsuperscript{th} grade (i.e. - the results of a projection model that was 100 percent accurate). The correlation between these two measures is .91.\textsuperscript{41} This implies that the status and projection measures of school performance are highly similar and that this

\textsuperscript{41}The displayed analysis shows results under a projection model that is perfectly accurate. Sensitivity analyses show that the correlation between status and projection using Arkansas’, Florida’s, or North Carolina’s projection model are also all above .90.
result is likely to hold regardless of the type of projection model used, so long as the model is fairly accurate.

Figure 5.3 - Percent Proficient vs. Percent Who Become Proficient (School Level)

This finding is clear and potentially surprising to those who believe that the GMPP’s “growth models” are an important improvement to NCLB’s original measures of school performance. Measuring schools’ relative performances under a status model or a projection model yields nearly redundant information. The reason for this has to do with the way that the GMPP’s projection models measure growth. Rather than requiring students to exceed a fixed amount of growth, the required learning gain each student must make depends upon their initial achievement level. Therefore, just like under a status model, schools are faced with relatively easier or more challenging tasks depending upon the achievement levels of their students when they first enter their school. Consequently, projection models are unlikely to have a large impact on how we measure schools’ performances, because they are not very different from the status model.
In light of Figure 5.2 (above), does Figure 5.3 (above) imply that there are virtually no low status high growth schools and very few high status low growth schools? Was the call by researchers to track individual students over time misguided, since it does not appear provide much new information? Figure 5.4 would suggest not. Like Figure 5.3 (above), this display plots each school’s status on the x-axis and their “growth” on the y-axis. As in Figure 5.3, status is measured by the percentage of students in each school who were proficient in 5th grade. “Growth” is measured by the schools 5th grade value-added score. Although the value-added model is statistically complex, in this case it is a reasonable approximation to think of this measure as representing the average individual gains students in a particular school made between 4th and 5th grade, just on a different metric.

**Figure 5.4 - Percent Proficient vs. Value-Added Score (School Level)**

![Graph showing the correlation between percent proficient and value-added score.](image)

The correlation between these two measures is .47 (as previously noted in Table 4.5). While status and value-added are moderately positively related, Figure 5.4 shows that there are many low status high growth schools, and several high status lower growth
schools. The schools in the upper left quadrant represent schools where students learned a lot relative to students in other schools, but because these students began the 5th grade school year at low achievement levels, their progress goes unrecognized by the status model or the GMPP's projection models. While it is desirable for the students in these schools to make even more dramatic learning gains such that they will reach proficiency, a nuanced accountability system would treat these schools differently than the low status low growth schools (lower left quadrant) that are more clearly underperforming. For example, consider the school with approximately 30 percent of its students passing the state exam and a value-added score above five. While the students in this school are making gains that are likely the second largest among all 96 schools in the study district, under NCLB this school will be labeled as failing. Students attending this school could be given the option to transfer, and if they choose to transfer they will likely attend a school where students are making much smaller learning gains. If this school consistently performed the same way, it could be restructured even though it is likely one of the most effective schools in the district.

In the previous section the status model and the projection model were both described as variable gains models. It was noted that under both models the likelihood that a school will receive “credit” for a student is highly dependent upon the student’s initial achievement level. Given this information, it should not be surprising that schools’ rankings tend to be nearly identical under the status model and the projection model. This finding validates experts intuitive concerns regarding whether projection models will “deliver results much different from those of the current accountability system” (Hoff, 2007) and confirms initial cross state findings of the Editorial Projects in
Education Research Center that show that the GMPP models “don’t appear to be making a big difference in the proportion of schools meeting annual goals under the federal law” (Klein, 2007, p. 24). This work not only demonstrates that the use of projection models is unlikely to have large impacts on accountability, it also explains the reason why: projection models are extremely similar to status models - both can be thought of as variable gains models that require differential learning gains based on students’ initial achievement levels.

The rules guiding the federal GMPP adhere to the core principle of NCLB – that all students must reach proficiency regardless of their initial achievement levels. As a result, projection models have become the centerpiece of the GMPP. The projection models used under the GMPP track individual students longitudinally, so they may sound similar to value-added models. Value-added models attempt to compare schools’ relative performances “fairly,” i.e., after accounting for students’ initial achievement levels. Value-added models attempt to judge schools based solely upon that which is within their control. Given the fact that projection models and value-added models sound similar, one of the major empirical findings in this research is that they yield very different assessments of schools’ performances. Not all growth models are alike.

It is well known that NCLB’s status model of old and value-added models represent two fairly different approaches to measuring school performance. Projection models seem like an interesting middle ground – they still hold all students up to the proficiency standards (like a status model, but unlike a value-added model), yet they utilize longitudinal individual-level data to measure growth (like a value-added model, but unlike a status model). *Perhaps the most important empirical finding from this*
research is that, although projection models and value-added models both utilize longitudinal student-level data, and they both measure “growth,” projection models are far more similar to the old NCLB status measure than they are to value-added models.
Chapter 6: Study Limitations

This study provides useful information regarding the accuracy of state proposed projection models; it informs about how realistic the status and projection goals of NCLB and the GMPP are; and it compares status, projection and value-added models as measures of school performance. However, there are limitations to the conclusions which can be drawn from this research.

Projection Model Accuracy

With respect to projection model accuracy, there are several plausible concerns regarding the generalizability of this study, some minor, others more significant. The potential limitations to the assessment of model accuracy are: 1) The study district data may not generalize to the whole state of Florida; 2) applying other state’s projection models to data from Florida ignores the fact that states’ models may have been designed for use on their specific state exam; and 3) the models may be more / less accurate depending upon the difficulty of states’ assessments.

The concern that the study district’s results may not generalize to the whole state of Florida is minor. While the study district is not demographically similar to the state as a whole, as noted in Chapter 3: Data, the study district is fairly similar to the state as a whole with respect to achievement trajectories. The average scale scores in the study district from 2002 through 2005 were consistently slightly below the statewide average.\(^{42}\) The standard deviations in the study district were also quite similar to the state as a

---

\(^{42}\) The study district’s average scale scores were between .10 and .15 standard deviations below the state wide average from 2002-2005.
whole.\(^4\) In an experimental study these numbers would not indicate equivalent groups; however, this is not an experiment where group equivalence is essential. In assessing the generalizability of the study districts results to the state as a whole, what is most important is whether the learning trajectories of students in the study district are significantly different from students in the state as a whole and, if such differences existed, that they occur in a way which would make the projection models behave differently in the study district compared to the rest of the state. There is not much reason to believe that this is the case.

The second concern with respect to model accuracy is that applying another state’s model to data from Florida ignores the fact that each state’s model may have been designed with the properties of their own state exam in mind. This concern is of moderate significance. In these analyses the underlying assumptions of states’ models were applied to Florida’s assessment and data. Notably, while accuracy of projections was influenced by model selection, results indicate that Florida’s model was the least accurate. All other states’ models outperformed Florida’s model, even though Florida’s model was applied to their own state assessment. In the case of Florida’s projection model, applying it to their state exam was of no advantage. This occurred because Florida’s model sets linear growth targets even though their developmental scale is curvilinear. Florida’s model would likely demonstrate the same bias if applied to Arizona’s or Arkansas’ state exams, because their exams’ proficiency cut scores indicate that their developmental scales are also curvilinear. More generally, the reason that Florida’s model did not perform well was because their model’s assumptions are

\(^4\) The study district’s standard deviation was around 10 points lower (around three to four percent) than the state average from 2002-2005.
unrealistic. A more detailed discussion of the underlying assumptions of each model can be found in Chapter 5: Discussion. What is most significant is that all models (except Tennessee’s model) make assumptions which are unlikely to reflect reality regardless of the exam upon which they are applied. Still, it is possible that the degree to which the state models’ assumptions are faulty depends upon which exam the models are applied to, and as a result the findings from this study may not generalize. For this reason, it is strongly recommended that each state assess their own model’s accuracy (and alternative approaches) on their own state data.

The third concern regarding projection model accuracy is that models may be more or less accurate depending upon the difficulty of the state exam. This is extremely likely to impact the absolute accuracy of each state’s model. Based on the difficulty of each state’s exam, making accurate projections can be relatively easier or more difficult than making projections in Florida. Why? Consider a state like North Carolina where, on the 2003 reading assessment, 81 percent of 4th grade students were proficient (Stullich, Eisner, McCrary, & Roney, 2006). In a state where proficiency levels are far from 50 percent (like North Carolina), making accurate projections is significantly easier. A fully naïve model that assumes all students are on track to become proficient would be accurate 81 percent of the time, if a state’s pass rate is 81 percent. In contrast, in the study district 41 percent of the observed students were proficient in 2005; therefore a naïve model (which assumes all students are not on track to become proficient) would only be accurate 59 percent of the time.

Generally, making accurate projections will be more challenging the closer the pass rate is to 50 percent; this occurs because the closer the pass rate is to 50 percent the
more students are on the “bubble,” and those are the students for whom it is most difficult to make projections. Since pass rates vary significantly from state to state (Stullich et al., 2006), the accuracy of any projection model is likely to vary depending upon the state’s data to which it is applied. *It is for this reason that one cannot reasonably compare the accuracy of North Carolina’s model applied to North Carolina’s state data to the accuracy of Florida’s model applied to Florida’s state data.*

This is also the reason why I do not recommend setting a fixed accuracy level above which a model is deemed acceptably accurate. An acceptable accuracy level will depend upon the pass rate on the state exam. Applying North Carolina’s model to their state data would very likely yield a higher overall level of accuracy than what was observed applying their model to Florida’s data. However, this would not necessarily imply that North Carolina’s model is “acceptably” accurate. As recommended in Chapter 4: Analyses and Results, at a minimum, any acceptable model should be more accurate than a simple model which assumes that all students will remain at their same proficiency status. It would be expected that in a state like North Carolina, where proficiency rates are very high, relatively few students switch proficiency status from year to year. Consequently, the minimum acceptable accuracy would be higher in North Carolina than in Florida.

While proficiency rates do affect the overall accuracy of the projection models, proficiency rates are not likely to impact the *relative* accuracy of state models. The relative accuracy of the various state models is most likely to be influenced by how well the models’ assumptions hold up to reality on the state exam to which they are applied.

---

44 Assuming test scores are normally distributed.
**Realistic Goals**

The third set of analyses looked to examine whether the individual level gain goals of a status model and a projection model are realistic. *The results from these analyses are highly dependent upon the difficulty of the state exam.* These analyses demonstrate that for initially low performing students, in the study district, both the status and projection goals are not realistic; however, these findings would be very different in a state with a lower proficiency threshold. That said, what is generalizable is the fact that a students’ likelihood of becoming proficient under a status or a projection model is going to be highly dependent upon a students’ initial achievement level. It can be stated with certainty that both status and projection models judge schools, in part, based on the initial achievement levels of their students.

**Comparing Measures of School Performance**

The fourth set of analyses compares status, projection, and value-added models as measures of school performance. Findings suggest that the relative rankings of schools are extremely similar under a status and a projection model. In contrast, schools’ relative rankings under a value-added model are different than under a status or projection model. There is little reason to believe that these results are unique to this dataset, rather they likely reflect what is being measured by each indicator of school performance.
Chapter 7: Conclusions

Since the passage of NCLB there has been a great deal of discussion regarding the measurement of school performance for accountability purposes. NCLB’s measures of school performance have been criticized in large part because they do not take into account students’ initial achievement levels. Since students enter schools at varying initial achievement levels, the challenge of bringing students up to proficiency is different from school to school. Schools with initially low performing students have a significantly greater challenge than schools with initially high performing students, yet under NCLB all schools are expected to reach the same fixed end of year goals. Many view these requirements (and their associated rewards, sanction, and assistance) as unfair because schools are judged by factors that are often beyond their control, i.e., students’ initial achievement levels. While the end goal is the same for every student and every school, the amount of achievement gains required are highly variable from student to student and from school to school.

An alternative conception for measuring school performance is to devise an accountability system that attempts to compare schools to each other based on each school’s relative contributions to the learning gains of their students. Value-added models are a class of statistical models that attempt to accomplish this, though they do so imperfectly. While some of the imperfections in value-added analysis may be kinks in the specification of the best statistical model, others are more fundamental to what is trying to be estimated. The former concerns are quite serious, but it is the latter concerns that imply that the endeavor may be somewhat of a chase for the Holy Grail. Nonetheless, even if value-added analyses cannot provide perfect measures of schools’
relative effectiveness, they still may be a useful tool in an accountability system (Rubin et al., 2004). Most for-profit businesses hire, fire, and promote individuals based on imperfect measures of performance, yet it is still largely believed that these imperfectly informed decisions are better than the next best alternative. That said, even if value-added models captured precisely what they intend to measure, using them in an accountability system would not require all students to reach proficiency. A school that is highly effective relative to other schools may have students who started at such low initial achievement levels that they never attain proficiency. Typical value-added models should rate such a school highly, even though some students will never become proficient.

The GMPP attempts to balance these two seemingly opposing ideals: One is to avoid judging schools based on that which is largely out of their control (past performance); the other is to require that all students reach a high, fixed level of academic achievement (proficiency). However, because students enter schools at varying achievement levels, balancing these two goals is extremely difficult.

Prior to the GMPP, the balance was steeply tipped towards requiring that all students reach proficiency, ignoring the relative difficulty of this task. A shift to an accountability system based solely on value-added models would completely reverse the balance in favor of attempting to avoid judging schools based on that which is beyond their control. But such a system would not require all students to reach proficiency.

Some may initially see the GMPP as a happy middle ground. However, the projection models used under the GMPP are actually a minor change from the NCLB status measure of school performance because both require schools to bring all students
up to proficiency, regardless of their initial achievement levels. Although projection models track individual students longitudinally and include a “growth” component, they should not be confused with value-added models. Projection models do not attempt to compare schools’ relative effectiveness. Projection models continue to require different gains standards for different students, resulting in a system where schools with initially low-performing students are required to make up significantly more ground than schools with initially high performing students.

At its core, the GMPP’s projection models preserve the core principle of NCLB - that schools be responsible for getting all children up to proficiency. Simultaneously, the models track individual students over time, measuring academic growth. However, this research demonstrates that projection models are not a middle ground between status and value-added. They lean heavily towards the status model. This is not surprising given that NCLB’s chief concern is with raising the achievement levels of all students to proficiency. Rather than holding schools accountable to the same learning gain standards, NCLB and the GMPP both focus on requiring schools to get all students up to proficiency, regardless of their initial achievement levels.

The push for growth under NCLB elucidates a key policy tension in designing an education accountability system. Should the primary school performance measure attempt to “fairly” compare schools’ relative effectiveness, or should it measure schools’ effectiveness at bringing their particular students up to proficiency (Elmore, Abelmann, & Fuhrman, 1996)? If the goal is to measure relative effectiveness, then value-added models need to be the centerpiece. If the measurement goal is to assess schools’ effectiveness at bringing their students up to proficiency, then status and projection
models are appropriate. Designing a single accountability system where these two perspectives are simultaneously upheld requires great nuance and some likely inconsistencies (or at least confusion). The implications of this overarching tension are significant, as each design may be appropriate for distinct policy initiatives.

Value-added measures are suited to promote competition and comparisons among schools. As a result, these measures would be desirable in a market-driven, school choice system. In addition, since value-added measures represent the best comparison of schools’ relative effectiveness, an accountability system based on value-added can more reasonably include rewards and/or sanctions as a consequence for performance according to this closest proxy to merit.

However, a major drawback of designing an accountability system around value-added measures of school performance is that those researchers who are most knowledgeable about these models are not certain how accurately they capture what they intend to estimate (McCaffrey et al., 2003; Raudenbush, 2004; Rubin et al., 2004). While there is a significant research base on value-added models of teacher effectiveness, there has been far less research on value-added models of school effectiveness (for more on why this matters, see the section titled The (Potential) Added Value of Value-Added Models). In addition, as noted by Elmore et al. (1996), some argue that a system that accounts for prior achievement “institutionalizes low expectations for poor, minority, low achieving students. Controlling for social background or prior achievement, in effect, holds schools with large proportions of such students to a lower standard of performance” (Elmore et al., 1996, p. 93).
In contrast, an accountability system based on status or projection models can be highly effective at shining a light on those schools with students scoring at particularly low (or high) achievement levels. As a result, these models can be used for the targeting of resources, or to determine which schools need assistance bringing students up to proficiency. Advocates of this type of system hold “that policy and resources should be used to create incentives for schools with high proportions of poor, minority, low-achieving students to improve learning at a faster rate than other schools” (Elmore et al., 1996, p. 93-4). In other words, they advocate a system where growth expectations are greater at schools where students enter at lower initial achievement levels.

However, status and projection models are not useful for making comparisons regarding schools’ relative effectiveness. Thus, it is extremely difficult to justify using these models for the purpose of rewarding or sanctioning schools, since this will lead to the rewarding of many relatively ineffective schools and the sanctioning of many relatively effective schools.

Ultimately, the theory of action behind an accountability system is likely to imply which measures of school performance are most appropriate. When designing an accountability system it is essential to 1) understand the uses and misuses of each measure, 2) acknowledge what can and cannot be measured, and 3) recognize the levels of uncertainty associated with the selected measures.

Policy Implications

For policymakers there are several significant implications of this research. Within the context of the GMPP, policymakers should require states to demonstrate the
accuracy of their models at the individual and school levels. States should be prohibited from using egregiously biased models. This may already be beginning to happen, as Arizona was asked to demonstrate the accuracy of their individual-level single-year projections prior to the approval of their model. However, Arizona was not asked to demonstrate the accuracy of their model’s longer term projections, nor their model’s accuracy at the school level. This should be a requirement in order for states to be approved to use growth models for accountability purposes.

That said, policies designed to improve the accuracy of the projection models used under the GMPP will only tinker with a measurement that is unlikely to have a major impact. This research shows that even if projection models were perfectly accurate, what they measure is so similar to what NCLB’s old status model measures that the impact of these models will be minimal. What makes this worse is that the projection models are not perfectly accurate. In fact, at the school level they are highly inaccurate. Consequently, at their best projection models mimic NCLB’s status model; at their worst they randomly allow some schools to make AYP that were otherwise on the cusp, simply because of measurement error. When one considers how little impact these models would have even if they were perfect, it is difficult to justify the noise they add to the system given their inaccuracies.

A recent analysis of six of the states currently piloting projection models under the GMPP confirms the finding in this research that the GMPP’s measures will likely have minimal impact. The analysis, conducted by the Editorial Projects in Education (EPE) Research Center, found that the GMPP’s growth models are not having a significant impact on the proportion of schools making AYP (Klein, 2007). In Alaska not
a single school made AYP under the growth model that would not already have made AYP under the status model. In Arizona less than one percent of schools made AYP based on growth alone. The one state where the models seemed to have an impact? Florida. According to the EPE Research Center “about 14 percent of the schools that made AYP in Florida made it under the growth model but not the status model” (Klein, 2007, p. 25). My research reveals the reason for this finding. It is not because Florida schools uniquely demonstrate tremendous growth; rather, it is because Florida’s projection model is highly biased, tending to predict that many more students are on track to become proficient than actually will become proficient. Moving forward with projection models has little potential to improve the current accountability system.

If policymakers are serious about measuring growth, they must understand that growth toward a standard (projection models) is merely a status model with bells and whistles. A substantial policy shift would allow states to use value-added models to measure the relative contribution of schools to the learning gains of their students. Such a system would likely need new policies designed to incorporate two dimensions along which a school is assessed – status and growth (value-added). A two by two matrix like Figure 5.2 could be used to allocate differential rewards/sanctions.45 Notably, the federal government recently announced a plan to begin a pilot program that would allow states to test out “differentiated accountability” (U. S. Department of Education, 2008b). Under this new program states will be allowed “to vary the intensity and type of interventions to match the academic reasons that lead to a school’s identification” as in need of improvement (U. S. Department of Education, 2008a). However, it does not appear that

45 The two by two matrix is one of many possible designs. It is probably oversimplified, and more tiers would be preferable (e.g. low, middle, and high status/growth), or use of a continuous metric.
the plan will permit the use of value-added models in the differentiation process, especially since they are not permitted under the GMPP. While the new plan has merit, it does not change the fundamental measures of school performance that require different schools to produce different learning gains by their students.

Under a system that measured two dimensions of school performance, schools consistently demonstrating low status and low growth would require the most intensive interventions, sanctions, and/or assistance. In contrast, consistently low status high value-added schools would likely not be sanctioned, since it is difficult to justify punishing schools where students are learning at a pace that is greater than normal. Such schools would certainly not be shutdown or restructured. This type of differentiation could lead to a more optimal allocation of resources, since effective schools would not be punished for the initial achievement levels of their students.

However, policymakers should be cautioned regarding the use of value-added models for school performance. Prior to their use, rigorous evaluation of these models is recommended (as should have been done with the projection models under the GMPP). For example, states could be required to assess the reliability of their value-added methods prior to implementation. Individual gain scores are often not very reliable (I estimate the FCAT's reliability of individual gains falls within the range of .50 - .68 in the study district). In this study, value-added scores were very highly correlated with average individual gains; thus, it is possible that value-added scores might not

46 Reliability of the FCAT across subjects and grades is generally around .90 (Human Resources Research Organization & Harcourt Assessment Inc., 2007). In the study population, the correlation between scores in from 2002-2005 was generally in the range of .70-.80. The ratio of the standard deviation at any two time points was generally between 1.0 and 1.25. Using this information, the reliability of raw gain scores was calculated as described by Rachor and Cizek (1996).
demonstrate high levels of reliability. Future research needs to be conducted to assess the reliability of average individual gain scores and/or value-added scores under a variety of conditions.

In sum, the projection models used under the GMPP do not stand to add much value to the current accountability system, yet they add sufficient noise that they may weaken the current measures of school performance. Status and projection models measure mostly redundant constructs, but status is reliable and accurate, whereas the implemented projection models are inaccurate. As such, projection models need to be improved if they are to be of any use. In contrast, value-added models would drastically change the current accountability, but much research is needed to determine whether school-level value-added measures can be reliable and meaningful.

Summary

Through empirical analyses this research addressed several critical issues regarding measures of school performance. Examined first was the use of projection models to project whether students would become proficient in the future. Findings suggest that the state proposed projection models are inaccurate at projecting future proficiency. Next, was an assessment of how reasonable accountability goals are in the study district under a status and under a projection model. Findings indicate that the goal of attaining proficiency is highly dependent upon students' initial achievement levels, regardless of whether one uses a status or a projection model. For initially low performing students in the study district, the proficiency goal does not appear realistic under a status or a projection model. Finally, this research compared the use of status,
projection, and value-added models for measuring school performance. Results indicate that status and projection models yield extremely similar results at the school level, whereas value-added models measure a different construct. The results from this research have implications for the federal Growth Model Pilot Program and for the design of accountability systems more generally.
Appendix A: Analyzing Projection Models

A description of each of the state projection models, as implemented, is provided in Chapter 3: Data and Methods. As noted in that section, Florida, Arizona, North Carolina, and Arkansas all use projection models that set growth targets which must be exceeded in order for a student to be deemed on track to become proficient. However, sometimes these models label students as on track to become proficient when in fact they do not become proficient. Other times these models label students as not on track to become proficient when in fact they do become proficient. In order to better understand when and why these mistaken projections are made, I utilize the underlying rational behind each projection model in order to obtain projected scale scores several years into the future, rather than the binary on track to become proficient or not on track to become proficient. Notably, using the implicit logic employed by each state model to project future scores, one can compare the projected scores to the proficiency cut score in order to determine whether students are on track to become proficient. Doing so yields precisely the same results as those obtained using the growth targets. However, the purpose of these analyses is not replication. Rather, by obtaining projected scores based on the fundamental assumption of each model we can gain a more detailed understanding of when and why these models do and do not work. The greater detail is afforded because rather than analyzing the blunt binary outcome obtained when applying each state’s model as explicitly described, obtaining projections on a continuous metric provides a richer depiction of the consequences of each state’s model’s implicit assumptions.
Described below are the equations used to project future DSS under each state’s model:

**Florida’s Projection Model**

Florida uses a linear projection model, where gains are assumed to be constant. Under Florida’s model growth targets are set on a linear path towards proficiency three years after a student’s grade of first enrollment (or first instate exam). If a student scores 1092 on the 3rd grade exam and the proficiency cut score in 6th grade is 1692, then she needs to gain 600 total points over the next three years. Based on the linear growth assumption, Florida sets her 4th grade growth target at 1292 (1092 + 1/3 * 600). If she exceeds 1292 then she is deemed on track to become proficient. The underlying assumption (linear growth) is that if she gained 200 points between 3rd and 4th grade, she will continue to gain 200 points between 4th and 5th grade, and again between 5th and 6th grade. Applying this guiding principle, two-year projections can be calculated using equation (A-1):

\[(A-1) \quad \hat{y}_4 = y_1 + 3 \times (y_2 - y_1)\]

Where:

- \(y_t\) = Developmental scale score in year \(t\)
- \(\hat{y}_4\) = Projected developmental scale score in year 4

This equation produces a projected year 4 DSS (\(\hat{y}_4\)) which assumes that a student will make linear growth. Similarly, single-year projections can be described using equation (A-2):
Arizona’s Projection Model

Arizona’s projection model involves a two-part process. The first part of the model involves a regression analysis with school fixed effects. Equation (A-3) describes this regression model:

\[ y_t = \alpha_s + \beta y_t + \epsilon \]

Where:

\[
\alpha_s = \text{Fixed effect for school } s
\]

\[
\beta = \text{Regression parameter for year } l \text{ developmental scale score}
\]

This regression model is run using a student’s past score \( y_t \) to predict her current score \( y_t \). For each student a year two score is predicted\(^{47}\) using the estimated coefficients obtained from the regression model applied to equation (A-4):

\[ \hat{y}_t = \alpha_s + \beta y_t \]

Where:

\[ \hat{y}_t = \text{Predicted scale score in year } 2 \]

In order to make a more conservative estimate of student performance, the lower bound of a 95 percent interval around \( \hat{y}_t \) is used as a student’s year two score, calculated using equation (A-5):

\(^{47}\) I use the word “predicted” here rather than “projected” because the outcome is a known value which was used to develop the regression model.
The standard error is defined as $SE(\hat{y}_2) = \sqrt{h_i s^2}$, where $h_i = x_i (X'X)^{-1} x_i'; X$ is the matrix of regressors, $x_i$ is the $i^{th}$ row of $X$ and $s^2$ is the estimate mean squared error.

The second part of Arizona’s growth model (as implemented by the state) involves calculating linear growth targets. These growth targets are set using the same linear assumption as Florida’s model. Consequently, projected year four scores can be calculated in the same way as was done under Florida’s model. The key difference is that Florida uses observed year two scores ($y_2$), whereas Arizona uses the lower bound of predicted year two scores ($\hat{y}_{2,Lower97.5}$). Mathematically, two-year projections are expressed by equation (A-6):

\[(A-6) \quad \tilde{y}_4 = y_4 + 3 \times (\hat{y}_{2,Lower97.5} - y_4)\]

Since Arizona updates its growth targets each year, this needs to be considered when making single-year projections. One-year projections are calculated using Arizona’s model as expressed in equation (A-7):

\[(A-7) \quad \hat{y}_4 = y_2 + 2 \times \left(\frac{\hat{y}_{3,Lower97.5} - y_2}{2}\right)\]

Predicted year three scores ($\hat{y}_{3,Lower97.5}$) are calculated using year two scores and school fixed effects as the predictors. If a student projected year four score is greater than the year four proficiency cut score, then the student is deemed on track to become proficient.

**North Carolina’s Projection Model**
North Carolina uses a linear projection model much like Florida’s model; however, North Carolina applies the linear projection model to standardized scores (C-Scale Scores) rather than to the developmental scale scores. Standardized scores are obtained by taking a student’s score, subtracting off the mean scale score (for that grade) and then dividing by the standard deviation (for that grade), as described in equation (A-8):

\[ y_{C\text{-}Scale} = \frac{y - \bar{y}}{s} \]

Where:
- \( y_{C\text{-}Scale} \) = C-Scale score of a student
- \( y \) = DSS of a student
- \( \bar{y} \) = The mean developmental scale score of all students in the standard setting year
- \( s \) = The standard deviation of the developmental scale scores of all students in the standard setting year

North Carolina’s one-year and two-year projections can be described mathematically by simply converting all developmental scale scores into C-Scale scores. Two-year projections can be described by equation (A-9):

\[ y_{C\text{-}Scale}^{4,\text{C-Scale}} = y_{1,\text{C-Scale}} + 3*\left(y_{2,\text{C-Scale}} - y_{1,\text{C-Scale}}\right) \]

One-year projections can be described by equation (A-10):

\[ y_{C\text{-}Scale}^{4,\text{C-Scale}} = y_{1,\text{C-Scale}} + 3*\frac{\left(y_{3,\text{C-Scale}} - y_{1,\text{C-Scale}}\right)}{2} \]

Arkansas’ Projection Model
Arkansas’ proportional growth model assumes that a student’s growth will mimic the proportion of change in the proficiency cut scores over time. Consider Florida’s mathematics FCAT. The proficiency cut scores from 3rd to 6th grade are 1269, 1444, 1632, and 1692 respectively. The change in the proficiency cut score from 3rd to 6th grade is 423 points. As such, the change from 3rd to 4th grade of 175 points represents 41.37 percent of the change from 3rd to 6th grade. Arkansas’ model implicitly assumes that a student’s gains from 3rd to 4th grade represent 41.37 percent of her expected gains between 3rd and 6th grade. This can be described mathematically by equation (A-11):

\[(A-11) \quad y_2 - y_1 = 0.4137 \times (y_4 - y_1)\]

Where:

\[y_t = \text{Developmental scale score in year } t\]

By rearranging equation (A-11) algebraically, we can solve for \(y_4\) to determine projected year four scores. Equation (A-12) describes these two-year projected scores.

\[(A-12) \quad \tilde{y}_4 = y_1 + \frac{y_2 - y_1}{0.4137}\]

Through similar logic, one-year projections can be calculated using equation (A-13):

\[(A-13) \quad \hat{y}_4 = y_1 + \frac{y_3 - y_1}{0.4137 + 0.4444}\]

**Tennessee’s Projection Model (EVAAS®)**

Tennessee’s model is used as described in Chapter 3: Data and Methods.
Appendix B: Residuals

Figure B.1 Plot of Residuals vs. Projected Values for Arizona’s One-year Projection Model (Student Level)

Correlation = 0.23
Figure B.2 Plot of Residuals vs. Projected Values for Arizona’s Two-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
2 Year Projection (Arizona)

Residuals
2000
1000
0
-1000
-2000

Projected DSS Score
200 400 600 800 1000 1200 1400 1600 1800 2000

Correlation = 0.38
Figure B.3 Plot of Residuals vs. Projected Values for Arkansas' One-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
1 Year Projection (Arkansas)

Correlation = 0.49
Figure B.4 Plot of Residuals vs. Projected Values for Arkansas’ Two-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
2 Year Projection (Arkansas)

Correlation = 0.82
Figure B.5 Plot of Residuals vs. Projected Values for Florida’s One-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
1 Year Projection (Florida)

Correlation ≈ 0.62
Figure B.6 Plot of Residuals vs. Projected Values for Florida's Two-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
2 Year Projection (Florida)

Correlation = 0.88
Figure B.7 Plot of Residuals vs. Projected Values for North Carolina’s One-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
1 Year Projection (North Carolina)

Correlation = 0.6
Figure B.8 Plot of Residuals vs. Projected Values for North Carolina's Two-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
2 Year Projection (North Carolina)

Correlation = 0.89
Figure B.9 Plot of Residuals vs. Projected Values for Tennessee’s One-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
1 Year Projection (Tennessee)

Correlation = 0.08
Figure B.10 Plot of Residuals vs. Projected Values for Tennessee’s Two-year Projection Model (Student Level)

Plot of Residuals vs. Projected Values
2 Year Projection (Tennessee)

Correlation = 0.2
Figure B.11 Plot of Residuals vs. Projected Values for Arizona’s One-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
1 Year Projection (Arizona)

Correlation = 0.88
Figure B.12 Plot of Residuals vs. Projected Values for Arizona's Two-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
2 Year Projection (Arizona)

Correlation = 0.89
Figure B.13 Plot of Residuals vs. Projected Values for Arkansas’ One-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
1 Year Projection (Arkansas)

Correlation = 0.39
Figure B.14 Plot of Residuals vs. Projected Values for Arkansas’ Two-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
2 Year Projection (Arkansas)

Correlation = 0.33
Figure B.15 Plot of Residuals vs. Projected Values for Florida’s One-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
1 Year Projection (Florida)

Correlation = 0.35
Figure B.16 Plot of Residuals vs. Projected Values for Florida’s Two-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
2 Year Projection (Florida)

Correlation = 0.37
Figure B.17 Plot of Residuals vs. Projected Values for North Carolina’s One-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
1 Year Projection (North Carolina)

Correlation = 0.34
Figure B.18 Plot of Residuals vs. Projected Values for North Carolina’s Two-year Projection Model (School Level)

Correlation = 0.34
Figure B.19 Plot of Residuals vs. Projected Values for Tennessee's One-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
1 Year Projection (Tennessee)

Correlation = 0.12
Figure B.20 Plot of Residuals vs. Projected Values for Tennessee’s Two-year Projection Model (School Level)

Plot of Projected Percent Proficient vs. Residuals
2 Year Projection (Tennessee)

Correlation = 0.06
Appendix C: Arizona’s Special Case

As described in Chapter 3: Data and Methods, year two predicted scores under Arizona’s model can be obtained using the formula described in equation (C-1):

\[(C-1) \quad \hat{y}_2 = \alpha_s + \beta y_1\]

Where:

\(\hat{y}_2\) = predicted year two score

\(\alpha_s\) = Fixed effect for school \(s\)

\(\beta\) = Regression parameter for year 1 developmental scale score

Note that a student’s predicted year two score is a function of what school she attends and her year one score. Individual students are deemed on track to become proficient if their predicted year two score exceeds a growth target.\(^{48}\) The growth target is calculated under the assumption of linear growth, as described in equation (C-2):

\[(C-2) \quad \tilde{y}_2 = \frac{y_{4,\text{cutscore}} - y_1}{3} + y_1\]

Where:

\(\tilde{y}_2\) = Year two growth target

\(y_{4,\text{cutscore}}\) = Year four proficiency cut score

\(y_1\) = Year one observed developmental scale score

Note that a student’s growth target is a function of the year four proficiency cut score and her year one observed score. As such, a student’s predicted score and her

---

\(^{48}\) Arizona does not compare a student’s observed year two score to a growth target, but instead compares their “predicted” year two score to a growth target even though the observed year two score is known. This may help reduce problems regarding regression to the mean, but it does not take advantage of all available information.
growth target are both a function of her year one score. For a student to be on track to become proficient her predicted year two score, \( y_2 \), must be greater than her growth target, \( \bar{y}_2 \). By substituting equations (C-1) and (C-2), this can be described by equation (C-3):

\[
(C-3) \quad \alpha_s + \beta y_1 > \frac{y_{4, \text{cutscore}} - y_1}{3} + y_1
\]

If the condition set forth in equation (C-3) is true, then a student is labeled on track to become proficient. This equation can be rearranged algebraically to solve for \( y_1 \) as shown in equation (C-4):

\[
(C-4) \quad y_1 > \frac{\frac{y_{4, \text{cutscore}}}{3} - \alpha_s}{\beta - \frac{2}{3}}
\]

In the case of the two-year projection model applied to the study district data, the regression analysis estimated that \( \beta = .6656 \). Consequently, the denominator in equation (C-4) is very small. Given that the 6th grade mathematics cut score \( (y_{4, \text{cutscore}}) \) is 1692, through substitution equation (C-4) can be rewritten as equation (C-5):

\[
(C-5) \quad y_1 < \frac{564 - \alpha_s}{-.00106}
\]

Because the denominator is so small, the right hand side of equation (C-5) is rarely in a range where \( y_1 \) can feasibly cross. Consequently, whether students are labeled

\[\overset{49}{\text{49}}\]

Note that if \( \beta < 2/3 \), then the inequality sign would be reversed, and (oddly enough), students with lower \( y_1 \) scores would be on track to become proficient while students with higher \( y_1 \) scores would not be on track to become proficient.

\[\overset{50}{\text{50}}\]

Note the inequality sign has reversed because \( \beta < 2/3 \).
on track to become proficient is rarely a function of \( y_i \) and \( \alpha_j \); rather it is a function of \( \alpha_j \) alone. In schools where \( 564 - \alpha_j > 0 \), the condition in equation (C-5) is never met because \( y_i \) cannot be negative; therefore, all students are deemed not to be on track to become proficient, regardless of their \( y_i \). In schools where \( 564 - \alpha_j < -2.37 \), all students are labeled on track to become proficient. This equation is derived because the maximum value of \( y_i \) was 2225, and plugging this value into equation (C-5) yields

\[
564 - \alpha_j < -2.37.
\]

In this dataset there were no cases where \( 0 < 564 - \alpha_j < 2.37 \), hence all schools had either no students on track to become proficient or all students on track to become proficient.

While this finding alone is somewhat condemning of Arizona's model, it is a rare situation which has occurred with data from Florida, a situation which only exists when \( \beta \) is very close to \( \frac{2}{3} \). However, what may be of even greater concern is the fact that if \( \beta < \frac{2}{3} \), then within any given school those students who are deemed on track to become proficient will all have lower \( y_i \) scores than those students who are deemed not on track to become proficient. This occurs because the inequality in equation (C-4) reverses if \( \beta < \frac{2}{3} \). In Arizona's GMPP proposal they provide a sample of their analyses in which \( \beta = .8111 \). It is possible that using Arizona's achievement data this would never be a problem, but states should be cautioned regarding Arizona's method.
References


Alliance for Fair and Effective Accountability. (2004). *Joint organizational statement on 'No Child Left Behind' Act*


No Child Left Behind Act, Subpart 1. Part A. Sec. 1111.b.2.(I) Annual Improvement for Schools (i) (2001a).


Ohio Department of Education. (2006). *Proposal to the United States Department of Education for employing a growth model for No Child Left Behind accountability*


