1-1-2012

A Policy Analysis of the Federal Growth Model Pilot Program's Measures of School Performance: The Florida Case

Michael J. Weiss

Henry May
University of Pennsylvania

Weiss, Michael J. and May, Henry, "A Policy Analysis of the Federal Growth Model Pilot Program's Measures of School Performance: The Florida Case" (2012). *GSE Graduate Student Research*. 1.
https://repository.upenn.edu/gse_grad_pubs/1

Weiss, M. J., & May, H. (2012). A policy analysis of the federal growth model pilot program's measures of school performance: The Florida case. *Education Finance and Policy,* 7(1), 44-73. doi: 10.1162/EDFP_a_00053

A Policy Analysis of the Federal Growth Model Pilot Program's Measures of School Performance: The Florida Case

Abstract

As test-based educational accountability has moved to the forefront of national and state education policy, so has the desire for better measures of school performance. No Child Left Behind's (NCLB) status and safe harbor measures have been criticized for being unfair and unreliable, respectively. In response to such criticism, in 2005 the federal government announced the Growth Model Pilot Program, which permits states to use *projection models* (a type of growth model) in their accountability systems. This article uses historical longitudinal data from a large school district to empirically show the inaccuracy of one state's projection model, to demonstrate how projection models are very similar to NCLB's original status measure, and to contrast projection models with value-added models. As policy makers debate the reauthorization of NCLB, this research can provide guidance on ways to improve the current measurement of school performance.

Disciplines

Education | Educational Assessment, Evaluation, and Research | Education Policy

Comments

Weiss, M. J., & May, H. (2012). A policy analysis of the federal growth model pilot program's measures of school performance: The Florida case. *Education Finance and Policy,* 7(1), 44-73. doi: 10.1162/EDFP_a_00053

A POLICY ANALYSIS OF THE FEDERAL GROWTH MODEL PILOT PROGRAM'S MEASURES OF SCHOOL PERFORMANCE: THE FLORIDA CASE

Michael J. Weiss
(corresponding author)
MDRC
New York, NY 10016
michael.weiss@MDRC.org

Henry May
Graduate School of Education
University of Pennsylvania
Philadelphia, PA 19104

Abstract

As test-based educational accountability has moved to the forefront of national and state education policy, so has the desire for better measures of school performance. No Child Left Behind's (NCLB) status and safe harbor measures have been criticized for being unfair and unreliable, respectively. In response to such criticism, in 2005 the federal government announced the Growth Model Pilot Program, which permits states to use *projection models* (a type of growth model) in their accountability systems. This article uses historical longitudinal data from a large school district to empirically show the inaccuracy of one state's projection model, to demonstrate how projection models are very similar to NCLB's original status measure, and to contrast projection models with value-added models. As policy makers debate the reauthorization of NCLB, this research can provide guidance on ways to improve the current measurement of school performance.

1. INTRODUCTION

The 2001 reauthorization of the Elementary and Secondary Education Act has led to a rise in the visibility and significance of student testing in America's public schools. Statewide achievement exams, used as indicators of student achievement and school performance, are now more prevalent and consequential than at any point in the history of U.S. education. Under the No Child Left Behind Act (NCLB), student achievement scores are used to determine whether schools are performing "adequately," and failure to do so may have serious consequences.

The original NCLB measures of school performance, which are used to determine whether schools are making adequate yearly progress (AYP), focus on the status of schools at a single point in time and school-level changes in the percentage of proficient students from one year to the next. While these measures serve a purpose, critics find school-level status to be simplistic and school-level changes in proficiency rates to be flawed from a measurement perspective. Status measures are most often criticized because they do not take into account students' initial achievement levels, so schools are judged largely by that which is beyond their control. School-level changes in proficiency rates are statistically unreliable and do not reflect true school improvement in part because they compare different cohorts of students; consequently, changes in percent proficient often reflect natural sampling variability, changes in student demographics over time, interschool student mobility, or retention in grade (Kane and Staiger 2002; Linn 2004; Choi, Goldschmidt, and Yamashiro 2005; Lissitz et al. 2006). As a result, there is growing interest in alternative measures of school performance. The leading alternative is a class of models known as growth models, which measure changes in *individual* students' achievement levels over time.

Individual-level growth models have become increasingly relevant in the national education policy arena since the 2005 adoption of the federal Growth Model Pilot Program (GMPP) (USDOE 2005). The GMPP, which initially sought to allow up to ten eligible states to pilot growth models for school accountability, is now open to all states that submit proposals aligning with the core principles of the program (USDOE 2007). Since the GMPP is fairly new, there is little research regarding the growth models used under this program. The research that has been conducted suggests that growth 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 article reveals some reasons for these findings and explains why we might not expect significant changes in the number of schools making AYP given the constraints of the types of growth models allowed under the GMPP. (The work of Dunn and Allen [2008] offers further insight on this topic.)

Using existing achievement data, we first analyze the accuracy of Florida's growth model, finding that the model is inaccurate and biased.[1] More significantly, this research examines the likely impact of the type of growth model allowed under the GMPP. Results indicate that this type of growth model is very similar to the old status model and therefore is unlikely to have a meaningful impact on school accountability. Finally, we compare the GMPP's approved growth model to a value-added model (VAM), demonstrating that these two types of growth models are not only theoretically different but are practically very different as well. Researchers and policy makers need to be careful not to conflate the growth models used under the GMPP and VAMs.

2. DEFINING TERMS: GROWTH MODEL, PROJECTION MODEL, AND VALUE-ADDED MODEL

The terms *growth model* and *value-added model* are often used interchangeably (for examples see Porter and Polikoff 2007 and USDOE 2008). We wish to clarify their definitions and introduce a definition of a special class of growth models called *projection models*. In this research we focus on individual-level growth models (as opposed to school-level growth models). Borrowing from Lissitz et al. (2006), individual-level growth models refer to the entire class of models that utilize longitudinal data to track individual students' 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, or measuring the impacts of teachers or schools on student achievement using longitudinal data.

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, Sanders, and Rivers 2006). Projection models can be used to assess whether individual students have made sufficient learning gains in the past such that they appear to be on track to be proficient in the near future. The majority of states participating in the GMPP use projection models to measure growth. Using this measure, schools can be given credit for those students who have not yet achieved

[1] In this article *accuracy* refers to whether projections correspond with observed results. For example, if in fifth grade a student is projected to become proficient by sixth grade and she scores above the proficiency threshold on her sixth-grade exam, this fifth-grade projection is considered accurate. We use *bias* to refer to systematic inaccuracy in projections. For example, a projection model is considered biased if its inaccurate projections tend to be in the same direction. Florida's model is deemed biased because it systematically projects that students are on track to become proficient when they are not.

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 in the future. Of note is the fact that projection models do not attempt to measure schools' effectiveness relative to other schools' effectiveness or the "average" school's effectiveness.

Like projection models, VAMs 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).

With growth, projection, and VAMs defined, we now turn to a brief overview of the GMPP.

3. BACKGROUND: THE FEDERAL GROWTH MODEL PILOT PROGRAM

In response to requests by educators and policy makers 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 (USDOE 2005). Proposals were required to maintain the basic tenets of NCLB. The most notable of these core principles is that the models must require 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 those students who began below proficiency would always remain below proficient (USDOE 2005).

Consequently VAMs, which seek to identify how relatively effective a teacher or school is without regard to an absolute proficiency standard, are not allowed under the pilot program. Growth models must measure growth with respect to the proficiency standards, generally asking, "Are students on track to become proficient in the near future?" As explained in Tennessee's approved 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 theory established by NCLB: separate accountability decisions should be made for math and language arts, all students in tested grades should be included in the analyses, and schools are accountable for the performance of subgroups (i.e., racial subgroups, English language learners, socioeconomic status, etc.).

Projection Models

Since the 2005 announcement of the federal GMPP, fifteen states have had their growth model proposals accepted (Alaska, Arizona, Arkansas, Colorado, Delaware, Florida, Iowa, Michigan, Minnesota, Missouri, North Carolina, Ohio, Pennsylvania, Tennessee, and Texas).[2] Four proposals (Delaware, Iowa, Michigan, and Minnesota) measure growth using value tables, where schools are given extra credit for students who move from a lower nonproficient achievement level to a higher nonproficient achievement level (e.g., from below basic to basic) (Delaware Department of Education 2006; Iowa Department of Education 2007; Michigan Department of Education 2008; Minnesota Department of Education 2009). The remaining eleven states use projection models in which they attempt to assess whether each student, based on his or her growth, is on track to be proficient at a specified time point in the future. The eleven states use seven unique projection methodologies (Ohio, Pennsylvania, and Tennessee use the same method; Alaska, Florida, and Missouri use very similar methods) varying from a simple linear trajectory (e.g., Florida) to a more complicated longitudinal statistical model (e.g., Tennessee) (Alaska Department of Education and Early Development 2006; 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; Arizona Department of Education

[2] As of 21 July 2010.

2007; Missouri Department of Education 2008; Pennsylvania Department of Education 2008; Colorado Department of Education 2008; Texas Department of Education 2009).

While the methodology for calculating projections varies by state, the overall policy implementation is fairly similar across states. Generally states first assess whether a school makes AYP using the standard NCLB status and safe harbor measures. If a school does not make AYP using either of these measures, 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 examined only if a school fails to meet the status and safe harbor standards.

The growth component works as follows: For each student who is not currently proficient, the state projects whether the student has made sufficient learning gains such that he or she appears on track to become proficient in the future. Schools are then given "credit" for those students who are either currently proficient or on track to become proficient. The same rules that apply to NCLB's status measure are then applied to the percentage of students who are currently proficient or on track to become proficient. That is, if the status model required 54 percent of students to be proficient on the 2008 mathematics exam, the growth model requires that 54 percent of students be currently proficient or on track to become proficient. Some states using projection models (e.g., Tennessee) give schools credit only for those students who are on track to be proficient, regardless of their current proficiency status. This distinction can be important because projection models can be used to identify students who are currently proficient yet are not predicted to remain proficient in the future. However, the majority of states (Florida included) give schools credit for currently proficient students even if the state's projection model suggests the students are not on track to remain proficient in the future.

Generally states' growth models allow students three or four years to reach proficiency, at which time students 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 a state, the student must be proficient (i.e., the projection model does not apply to them).

Objectives of Using Projection Models

The main objective of the projection models is to give schools credit for those students who have made sufficient learning gains such that they appear on track to reach the fixed proficiency target at a specific time in the near future. While students (and therefore schools) are still held to different learning

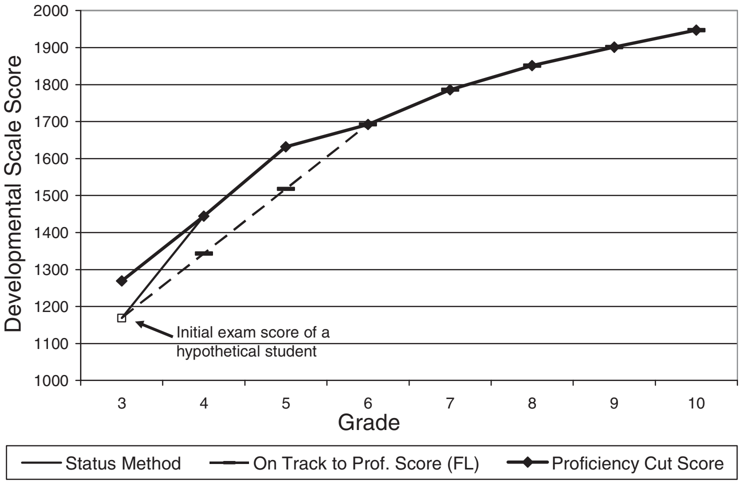

Figure 1. A Comparison of the Required Gains under a Status Model (NCLB) vs. a Projection Model (GMPP)

gains standards, the amount of time that schools have to bring students up to proficiency is longer than it is under the traditional status model. Under the traditional NCLB status model, in a school where the students were initially one year behind grade-level proficiency, the school had to make up the entire difference (two years of progress) in a single year. Using a projection model, the school would have three years to make up the difference (one and one-third years of progress per year over the course of three years). By giving students several years to reach proficiency, it is possible that the GMPP's objectives are more realistic than those under NCLB's original measures of AYP.

Figure 1 provides an illustration of how this works for a hypothetical third-grade student scoring 1,169, 100 points below proficiency, on Florida's vertically equated mathematics state exam. In this figure, proficiency cut scores are depicted as diamonds. Under both a status model (NCLB) and a projection model (GMPP), the school does not receive credit for this student in third grade because she is not proficient (no projections are made for students with data at only one time point). However, in fourth grade the status and projection models' requirements differ. Under the traditional status model this student must be proficient by fourth grade (i.e., she must score at least 1,444 on the math exam), requiring a gain of 275 scale score points. Under Florida's actual projection model, this student's fourth-grade score needs to demonstrate only that she is on track to become proficient by sixth grade, requiring a gain of 175 points (the horizontal bar above fourth grade).

The dashed line labeled "On Track to Prof. Score (FL)" represents growth targets set for this student under Florida's projection model. Florida draws a straight line from a student's initial achievement (i.e., third-grade score) to the sixth-grade proficiency cut score, reflecting an underlying assumption of linear growth. In fourth and fifth grades, if this student exceeds the growth targets (the horizontal bars on the dashed line), the school receives credit for her performance. In sixth grade this student must achieve proficiency for the school to receive credit for her, so from sixth grade on the status model is applied to this student. Notably, not all states use a linear projection model. That is, in other states the fourth- and fifth-grade growth targets are set based on different underlying assumptions. For example, Arkansas' projection model sets growth targets that require a student annually to close a percentage of the gap between her initial achievement level and proficiency. Since states set growth targets using various methods with different underlying assumptions, certain models will forecast future proficiency more accurately than others.

In order to assess the accuracy of one state's enacted projection model, we used data (described below) from a large urban school district in Florida. Consequently, the analyses regarding the accuracy of Florida's projection model may not generalize to all state projection models with respect to model accuracy. In contrast, the analyses regarding the potential impact of different projection models on measuring school performance are more likely to apply to all states, since those analyses are computed independent of a particular state projection model.

4. DATA

The analyses for this study use student-level vertically scaled scores from the Florida Comprehensive Assessment Test (FCAT) in mathematics. The FCAT is a reasonable assessment instrument for these analyses because the exam has a long history in the state, and there is some evidence that it is both reliable and valid (Florida Department of Education 2004). The data are from the 2001–2 through 2004–5 school years. They are census data from public schools in a large urban school district in Florida. The study district was one of the twenty largest school districts in the United States, serving over 129,500 students enrolled in approximately 182 schools. The ethnic composition of the students in the urban school district in the study year was 46 percent white, 43 percent black, and 5 percent Hispanic. Forty-nine percent of students receive free or reduced price lunch.

This research focuses on the 2002 cohort of third-grade students, tracked from 2002 to 2005. These grades were selected because they are the grades most likely to be affected by the GMPP. Although the GMPP is designed to give credit to schools for some students who are not yet proficient, the GMPP projection models require that all students attain proficiency three to four years

Table 1. Mathematics FCAT Scale Score Descriptive Statistics for Study District

Year	*N*	% of Initial Test Takers	Mean (DSS)	Standard Deviation (DSS)	Percent Proficient	Mean (DSS) Statewide
2002	10,007	100	1,279	297	55	1,308
2003	8,875	89	1,418	267	50	1,446
2004	8,364	84	1,594	266	50	1,616
2005	7,550	75	1,617	250	41	1,653

Note: DSS = developmental scale scores.

after their grade of first enrollment or their first instate exam. As such, the GMPP is less likely to have much impact beyond sixth or seventh grade.[3]

Sample Attrition and Descriptive Statistics

This study examines 10,007 students with third-grade achievement scores in spring 2002 (out of a population of 11,485). Due to attrition, achievement data are not available for all 10,007 students in the years following 2002. Table 1 provides frequency counts for the 10,007 students from spring 2002 through spring 2005. Of the 10,007 students with mathematics achievement data in 2002, 7,550 (75 percent) had mathematics achievement data in 2005. Interdistrict mobility is the most likely explanation for the majority of the sample attrition.

Sample Descriptive Statistics

The FCAT is a vertically equated exam, meaning that students' longitudinal developmental scale scores (DSS) are on the same metric over time and across grades. Descriptive performance statistics for the study district's 2002 third-grade cohort of students are provided in table 1. The district's average mathematics FCAT DSS in third grade 2002 was 1,279, with a standard deviation of 297. From 2003 through 2005, the district's average student from the 2002 cohort of third graders scored 1,418, 1,594, and 1,617, respectively. These scores imply average annual gains of approximately 139, 176, and 23 points on the developmental scale.[4]

Table 1 also shows that the statewide average Mathematics FCAT scores for third graders in 2002 was 1,308, for fourth graders in 2003 was 1,446, for fifth

[3] Most states give students three to four years from their first in-state achievement test 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 three to four years of enrollment in a state.

[4] 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.

graders in 2004 was 1,616, and for sixth graders in 2005 it was 1,653. 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 trajectory of students in the study district was fairly similar to the average learning trajectory of students in the state as a whole. Most notably, typical gains both in the study district and throughout the state were significantly larger from third to fourth grade and from fourth to fifth grade compared with the relatively modest gains observed between fifth and sixth grades, reflecting a curvilinear developmental scale. This pattern of annual gains (from larger in earlier grades to smaller in later grades) is common on nationally standardized achievement tests as well (Bloom et al. 2008).

5. ANALYTIC STRATEGY AND RESULTS

Florida's GMPP projection model attempts to give schools credit for those students who are not currently proficient but have made sufficient learning gains such that they appear on track to become proficient one or two years into the future. We assess the accuracy of Florida's model at forecasting future proficiency by applying it to the data described above. Before revealing the results of these analyses, we will formally describe how Florida's projection model is used to determine whether students are on track to become proficient.

Florida uses a linear projection model. Under its model a student is labeled on track to become proficient, and the school receives credit for her performance, if her observed achievement score exceeds a growth target. Each student's growth targets are set independently. 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 in-state exam). For example, a student's year 2 growth target is set using equation 1:

$$\tilde{y}_{i2} = y_{i1} + \frac{1}{3}(y_{4,cutscore} - y_{i1}) \tag{1}$$

where:

$\tilde{y}_{i2}$ = student i's year 2 growth target;
y_{i1} = student i's year 1 observed developmental scale score (typically a student's third-grade test score); and
$y_{4,cutscore}$ = year 4 proficiency cut score.

In order to meet the year 2 growth target, a student must make up one-third of the distance from his or her initial achievement score to the proficiency cut

Table 2. Projecting Proficiency One Year into the Future for Currently Nonproficient Students (Student Level)

		Actual	
		Proficient	Not Proficient
Projected	Proficient	167 (4.6%)	1021 (28.2%)
	Not Proficient	167 (4.6%)	2261 (62.5%)

score in year 4. If $y_{i2} \geq \tilde{y}_{i2}$, then the student is labeled on track to become proficient. This is a two-year projection because, based on a student's year 1 and year 2 scores (y_{i1} and y_{i2}), the state projects whether he or she is on track to become proficient by year 4—that is, two years into the future.

A student's year 3 growth target is set using equation 2:

$$\tilde{y}_{i3} = y_{i1} + \frac{2}{3}(y_{4,cutscore} - y_{i1}). \tag{2}$$

If $y_{i3} \geq \tilde{y}_{i3}$, the student is labeled on track to become proficient. This is a one-year projection since, based on a student's observed year 1 and year 3 scores (y_{i1} and y_{i3}), the state projects whether he or she is on track to become proficient by year 4—that is, one year into the future.

In year 4 the projection model is no longer used; schools receive credit only for those students who actually score at or above proficient.

Projection Model Accuracy (Student Level)

Using the Florida model and the study district's historical data, we are able to make projections one and two years into the future and check the accuracy of these projections. Using students' 2002 and 2004 achievement scores, we project whether nonproficient students (in 2004) were identified as on track to become proficient in 2005 (a one-year projection). The projections are checked for accuracy by comparing each student's projected proficiency status with his or her observed 2005 proficiency status. Similarly, using students' 2002 and 2003 achievement scores, we project whether nonproficient students (in 2003) were identified as on track to become proficient in 2005 (a two-year projection) and then compare the projections with each student's proficiency status (in 2005). Calculations are made for nonproficient students only because this is how Florida enacts their projection model. Tables 2 and 3 compare the projections with the observed results.

Several striking findings can be observed in tables 2 and 3. By adding the numbers on the diagonal we can calculate the overall accuracy of Florida's

Table 3. Projecting Proficiency Two Years into the Future for Currently Nonproficient Students (Student Level)

		Actual	
		Proficient	Not Proficient
Projected	Proficient	158 (4.3%)	1103 (30.1%)
	Not Proficient	293 (8%)	2110 (57.6%)

model. Projections one year into the future are accurate 67 percent of the time; projections two years into the future are accurate 62 percent of the time. Is this level of accuracy "good enough"? One standard for assessing the overall accuracy of a projection model is to compare the model's results with a naive model that does not use individual growth to project future proficiency. That is, consider what would happen if Florida assumed that any student who was not proficient in 2004 was not on track to become proficient by 2005 (regardless of his or her growth). One would hope that Florida's actual projection model, which uses two years of data to project proficiency based on individual growth, would be more accurate than attempting to project future proficiency by simply assuming that all students who are currently not proficient will remain not proficient in the future. The accuracy of this naive model can be attained by summing the values in the second column of table 2 or 3. Such a naive model would be accurate 91 percent of the time for one-year projections and 88 percent of the time for two-year projections. This result is critical for two reasons. First, it demonstrates that Florida's model is inaccurate, performing worse than a naive status model at projecting future proficiency. Second, this result suggests that of those students who were not proficient in 2003 or 2004, very few became proficient by 2005. Since so few students switch proficiency levels, there is little chance that Florida's projection model can capture that which is not already a part of the status measure. This becomes even more transparent in the section of this article that compares different measures of school performance.

Model Bias (Student Level)

The overall accuracy of Florida's projection model can be broken down into subcategories, as displayed in tables 2 and 3, 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. It is a concern because biased models are less likely to make accurate projections and may lead to misguided conclusions.

Under Florida's linear assumption, its one-year projection model made accurate projections 67 percent of the time. The incorrect projections (33 percent) represent the sum of the 28 percent of nonproficient students who were projected to become proficient but did not, plus the 5 percent of nonproficient students who were projected to remain nonproficient but actually became proficient. This means that six out of every seven incorrect projections involved projecting students to be on track to become proficient when they were not. Florida's model (when used to project students' sixth-grade proficiency) has a strong propensity toward falsely projecting that students are on track to become proficient because typical growth patterns on Florida's mathematics FCAT show nonlinear gains over time, while the state's projection model assumes constant gains over time. Since students' growth trajectories tend to be nonlinear, Florida's underspecified linear model overestimates the number of students who are on track to become proficient in sixth grade, resulting in disproportionately large numbers of students projected to become proficient who will not become proficient.

Projection Model Accuracy (School Level)

The previous section focused on the accuracy of projections at the individual level. However, under the GMPP, projection models are used for accountability purposes at the school level. For example, Florida calculates the percentage of students within a given school who are either currently proficient or on track to become proficient according to its projection model. For simplicity, this will be referred to as a school's projected percent proficient. This percentage is used to determine whether a school is making AYP, and we assess its accuracy in this section.

In the previous section, one measure of the accuracy of the individual-level projections was to compare projected proficiency with observed actual proficiency. Similarly, in this section, model accuracy is assessed by comparing each school's projected percent proficient with the percentage of all (initially not proficient and initially proficient) students who actually were proficient in 2005. However, unlike individual-level accuracy, school-level accuracy is not assessed by claiming each school-level projection to be either correct or incorrect. (If the model projects that 70 percent of students in a school are on track to be proficient by sixth grade and 69 percent actually became proficient, this could be viewed as a fairly accurate projection, not as inaccurate because it was off by one percentage point.) Instead of viewing the projected percent proficient as either correct or incorrect, school-level accuracy is assessed by subtracting each school's projected percent proficient from the percentage of those same students who actually became proficient in 2005. The differences

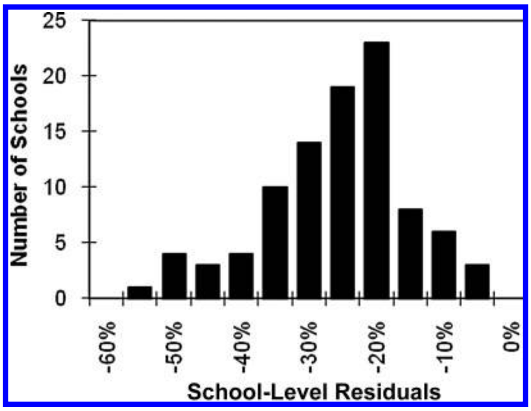

Figure 2. School-Level Residuals (Observed Minus Projected Percent Proficient): One-Year Projection

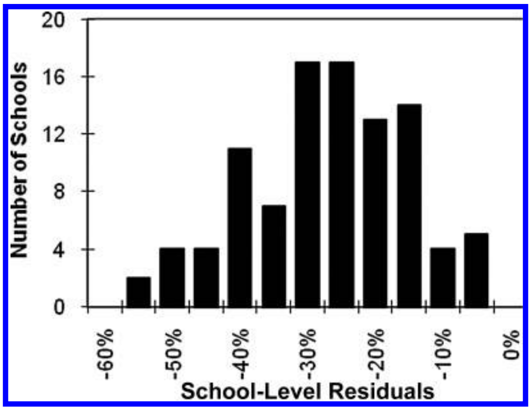

Figure 3. School-Level Residuals (Observed Minus Projected Percent Proficient): Two-Year Projection

are presented in figures 2 and 3 as the school-level residuals for one-year and two-year projections, respectively. Only those schools with thirty or more students' projected scores were included in these analyses.

In figures 2 and 3 the y-axis represents the number of schools and the x-axis represents the school-level residuals. 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 bias the model demonstrates, suggesting that the model systematically over- or underestimates the projected percent proficient.

Model Bias (School Level)

In figures 2 and 3 we see that Florida's model shows significant bias, typically overestimating the percent of students that will be proficient. Both the one-year and two-year projections are so dramatically biased that in all but one school the projected percent proficient is overestimated. In fact, the average residuals are −25 percentage points and −26 percentage points for one-year and two-year projections, respectively. This suggests that in a school whose projected percent proficient was 75 percent, on average only 50 percent of the students actually became proficient. Consequently, using Florida's projection model, many schools will make AYP not because of significant progress made by their students but because Florida's projection model is biased.

While the bias of Florida's projection model is a major problem, another way to examine the school-level projections is to consider their precision as measured by the standard deviation of the residuals (i.e., the standard error of prediction). The standard deviations of the residuals are 11 and 12 percentage points for one-year and two-year projections, respectively. These large standard deviations are reflected in the significant variation in the residuals depicted in figures 2 and 3. Consider the implications of this level of imprecision: even if Florida's projection model were not biased, a 95 percent prediction interval would still have to span over 43 percentage points (±1.96 * 11 percent). That is, if a school's projected percent proficient were 60 percent, the 95 percent prediction interval would span from 39 percent to 81 percent. Such a wide interval is not practical for accountability purposes.

Sensitivity Analysis

Florida's GMPP projection model implicitly assumes that those students who are currently proficient will remain proficient in the future. Other states (Ohio and Tennessee) participating in the GMPP attempt to assess whether currently proficient students are on track to remain proficient. In those states, schools do *not* receive credit for students who are currently proficient but whose limited growth, according to the projection model, indicates that they are not on track to remain proficient. Applying this approach to Florida's data and projection model yields the same general conclusions: Florida's model is inaccurate and biased at both the student and school levels. Similar analyses were conducted by applying projection models of Arkansas, North Carolina, and Tennessee to this same data set. For detailed results from these analyses, see Weiss (2008).

Since Florida's projection model is inaccurate, examining how many or what types of schools make AYP based on growth alone is not particularly meaningful. The previous analyses demonstrate that Florida's projection model will likely identify many schools as making AYP because of "growth." However, such findings would reflect the state's biased projection model, not schools that are getting students on track to become proficient. As a result, we choose to address a more meaningful question next: if a projection model were able to perfectly forecast future proficiency without bias or imprecision, what impact might we expect to see on an accountability system? While forecasting future proficiency perfectly is an unattainable goal, using retrospective data we can demonstrate the potential value of using projection models under optimal circumstances. If, under a perfectly accurate projection model, new and useful information about school performance is obtained, it is worthwhile to pursue the improvement of inaccurate projection models like Florida's. However, if a perfectly accurate projection model produces essentially redundant information, these models may be of limited potential. In the next section we compare schools' performances using three measures: NCLB's original status model, a projection model that perfectly forecasts future proficiency, and a VAM. These analyses will show that status and projection models produce results that are quite similar, while VAMs appear to measure a different dimension of school performance.

Comparing Measures of School Performance

Strong criticism of NCLB's aggregated school-level status measure 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. In this section we compare three approaches to assessing school performance: status models (NCLB), projection models (GMPP), and VAMs (used in some state accountability systems).

One way of comparing these three approaches is to consider schools' relative performance under each method. In this way we can assess whether the three different methods are providing unique information.

Schools' performances using each method were computed on a single analytic sample of students[5] as follows:

1. *Percent proficient (NCLB):* Under NCLB's status model schools are judged based on the percent of students who are currently proficient on the state exam. In these analyses each school's percent proficient measure

[5] Fifth grade was selected for reasons related to the second two measures of school performance considered. By using fifth-grade results, the GMPP's "percent on track to become proficient" measure could be computed using Florida's projection model as well as the percent of students who actually became proficient by sixth grade. The percent of students who actually became proficient may be more relevant since it is unaffected by the biases and imprecision of a particular projection model. Fifth grade was also selected because the VAM 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.

represents the percentage of fifth-grade students who passed the state exam in 2004.[6]

2. *Percent on track to become proficient (GMPP):* Under some states' GMPP proposals, schools are judged based on the percentage of all students who are on track to become proficient on the state exam. We look at the percentage of all students in each school who in fifth grade (2004) were on track to become proficient by sixth grade (2005). In this section, for each school we use the percent of fifth-grade students *who actually became proficient in sixth grade* as a proxy for the percent of fifth graders who were on track to become proficient. In this way, the results are unaffected by the inaccuracy and/or biases of a particular projection model. Sensitivity analyses were conducted using Florida's enacted projection model, and the results were substantively the same.[7]
3. *Value-added score:* School value-added scores were computed using the three-year historical records of fifth-grade students in 2004. The layered mixed effects model (LMEM) used here is described in Tekwe et al. (2004), including a description of the model, model equation, and SAS code found on pages 19–21 of the article. It is the foundation of the Tennessee Value-Added Assessment System (TVAAS), probably the most widely used value-added system in the country.[8] For simplicity, schools' value-added scores can be thought of as approximately equal to the average individual scale score gains from 2003 to 2004 (on a standardized scale), after adjusting for students' historical performance.[9]

Table 4 provides an overview of the correlations between schools' performances under each of the three methods described above. The correlations reveal considerable similarities among measures (evidenced by large positive correlations). Most notably, the correlation between the school-level percent proficient and the percent of those students who became proficient the next year is very high ($r = .89$). This suggests that schools' relative performance under a status model (NCLB) and a projection model (GMPP) are very similar

[6] The analytic sample used to create table 4 and figures 5 and 6 includes only those students with valid scores in years 2002, 2003, and 2005 who were in fifth grade in 2004, excluding schools with fewer than thirty students. The final analytic sample size is 6,945, and the exact same students were used for status, projection, and value added.

[7] This is probably the case because correlations are scale invariant. Florida's projection model takes the schools' observed percent proficient and generally bumps it up a bit based on those students who were not currently proficient but are deemed on track to become proficient. Doing this does not tend to change schools' relative rankings, which is largely what is captured using a correlation.

[8] The TVAAS should not be confused with Tennessee's projection model.

[9] Average individual gain scores have a .98 correlation with the value-added scores in this data set, so they can be thought of as essentially the same.

Table 4. Correlations between School Scores Using Various Measures of School Performance

	Percent Proficient 5th Grade (NCLB)	Percent Who Became Proficient by 6th Grade (GMPP)	Value-Added Score
Percent proficient 5th grade (NCLB)	–		
Percent who became proficient by 6th grade (GMPP)	0.89	–	
Value-added score	0.46	0.19	–

and that these measures are providing largely redundant information. Schools with a relatively high percentage of proficient students also have a relatively high percentage of students on track to become proficient, whereas schools with a relatively low percentage of proficient students have a relatively low percentage of students on track to become proficient. Since correlations are scale invariant, it is still 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, this high correlation implies that differences between schools making AYP under a status model versus a projection model are likely a function of the difficulty of being deemed proficient in the status year versus the projection year (i.e., the difficulty of the fifth-grade exam compared with the sixth-grade exam), an artifact of the proficiency cut scores.

While school-level status and GMPP projection models yield extremely similar results, also of considerable note in table 4 is that the correlations between schools' value-added scores and percent proficient, and value-added scores and percent on track to become proficient, are only moderate and small. The modest correlation between percent proficient and value-added scores ($r = .46$) suggests a significant difference in the information conveyed by these two metrics. Generally schools with higher percent proficient have high value-added scores and schools with lower percent proficient have lower value-added scores; however, the majority of variation in each of these measures is not associated with the other. The small correlation between the percent on track to become proficient and value-added scores ($r = .19$) may come as a surprise to those who think that GMPP projection models are very similar to VAMs. It should be clear that projection models and VAMs are not only theoretically different but result in significant differences in their assessments of schools' performances. While projection models yield very similar assessments of schools' performance compared with a simple status model, value-added models provide information that is substantially different from NCLB's status model or projection models.

6. DISCUSSION

Projection Model Accuracy

The first section of this research addressed the accuracy of Florida's projection model used under the GMPP and found the model to be inaccurate and biased. Analyses demonstrate that Florida's model, which utilizes students' longitudinal data records, is no more accurate than making projections by assuming students will remain at their current proficiency status in the future.

For school accountability purposes, the accuracy of projections at the student level is mostly irrelevant, since states consider only the aggregated results of these models when determining whether schools make AYP. However, inaccuracy at the student level is still policy relevant because several states (Arizona, Arkansas, Florida, Ohio, Tennessee) plan to report the results of individual projections to students, parents, teachers, and/or schools (Arkansas Department of Education 2006; Florida Department of Education 2006; Ohio Department of Education 2006; Tennessee Department of Education 2006; Arizona Department of Education 2007).

Under Florida's model more than one in four nonproficient students would have been labeled as on track to become proficient even though they did not become proficient. Such inaccurate individual-level reporting may result in false expectations for students, parents, and teachers. Under a model like Florida's, too many children who are on the path toward being left behind are 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.

At the school level, Florida's model did not fare any better. It demonstrates bias, systematically overestimating the percentage of students who are on track to become proficient. In addition, Florida's model demonstrates large variation in the accuracy of its school-level projections. If 95 percent prediction intervals were created around the projected percent proficient at the school level, ranges would span over 40 percentage points. This is the case even when ignoring model bias and making projections only a single year into the future. The fact that Florida's model can only claim, for example, that between 20 and 60 percent of students in a school are on track to become proficient is not precise enough to be practically useful. If prediction intervals are not used and point estimates are relied upon, large numbers of schools will be rewarded or sanctioned based on prediction error.

Why is Florida's projection model not very accurate? How can we do better? Analyzing the underlying assumptions of Florida's projection model reveals two major problems. 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 nonlinear, with average test score gains getting smaller in later grades, using a model that assumes linear growth is going to overestimate the number of students who are making sufficient growth. Using historical data, each state should be required to test its model's assumed growth trajectory and predictions.

The second problem with Florida's model is that it does not account for the distortion of gain scores. Any time a variable is measured imperfectly (i.e., any variable with measurement error or sampling variability), regression toward the mean will influence changes in that variable.[10] 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 to have distorted gain scores. Appropriate growth models should take this into account.[11]

As applied to historical data from the study district, Florida's projection model does not demonstrate impressive levels of accuracy at the individual or the school level. 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. Florida's model, which demonstrates disproportionately high numbers of false positives, is simply letting schools "off the hook" for a few years. At the designated point in the future, many of those students who were supposedly on track to become proficient will fail.

While these findings speak most directly to the inaccuracy of Florida's GMPP projection model applied to the state exam, this case highlights issues that may exist in all states participating in the GMPP. There is limited evidence to believe that any of the other states' GMPP projection models are accurate enough to be of practical use. The federal government should therefore require all states participating in the GMPP to demonstrate the accuracy of their models at the individual and, more important, school levels.

[10] We prefer the phrase "distortion of gain scores" because if the variance in students' test scores increases over time (i.e., scores fan outward), extreme scores may not regress toward the mean, although they will still be distorted.

[11] One additional hypothesized improvement would be to develop a projection model that predicts each students' probability of becoming proficient in the future. Doing so could have two advantages. First, in states where individual-level projections are reported to students and parents, they could provide more accurate information to students who are on the borderline of being on track to become proficient. A student whose expected probability of being proficient is 52 percent is quite different from a student whose expected probability of being proficient is 92 percent. In addition, it is possible that the sum of students' probabilities of future proficiency might more closely approximate the percent who become proficient, compared with the sum of the simplified dichotomous "on track" or "not on track."

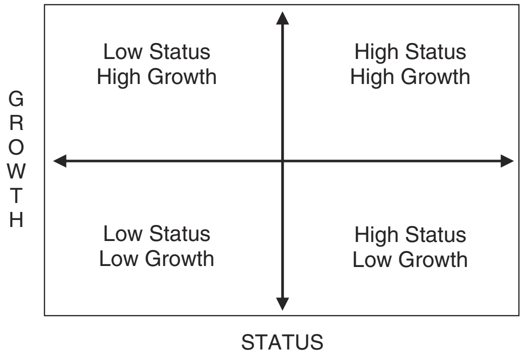

Figure 4. School Status vs. Growth

Comparing Measures of School Performance

The second part of this research assesses the similarities and differences among measures of school performance. It is important to keep in mind that a critical reason for interest in using individual-level growth models to measure school performance is the belief that there are 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 make relatively large learning gains (i.e., high growth), yet few students pass the year-end proficiency exam (i.e., low status) simply because students' initial achievement levels are very low. These schools are likely relatively effective (compared with other schools) even though they do not have as many students who demonstrate proficiency. Likewise, in some schools students make relatively small learning gains (low growth) yet most students pass the year-end proficiency exam (high status) simply because students' initial achievement levels are very high. These schools may be relatively ineffective (compared with other schools) even though they are successful at having students demonstrate proficiency. Figure 4 provides a visual depiction of how an accountability system might consider both status and growth when assessing schools' performances.

With figure 4 in mind let us examine the relationships among the three measures of school performance.

Figure 5 plots each study school's status on the x-axis and its growth on the y-axis. In this figure status is measured by the percentage of students in each school who were proficient in fifth grade. Growth is measured for these same fifth graders as under a projection model: the percentage of fifth-grade students who became proficient by sixth grade (i.e., the results of a projection model that was 100 percent accurate). As this figure demonstrates, the status and projection measures of school performance are highly similar

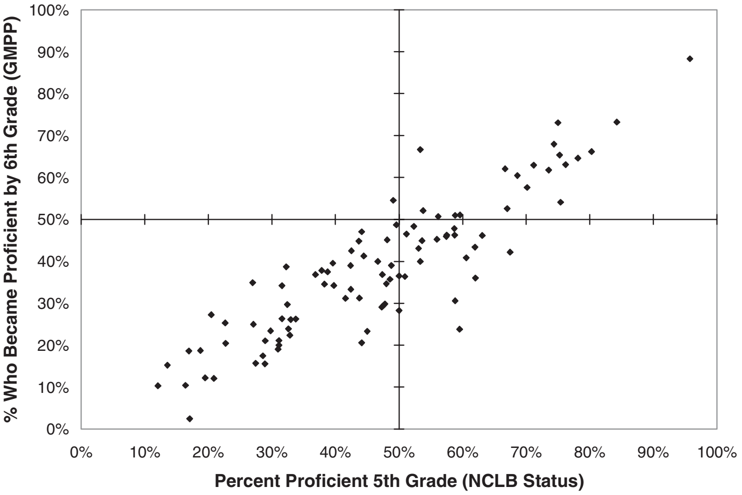

Figure 5. Percent Proficient vs. Percent Who Become Proficient (School Level)

(Dunn and Allen 2008 also note the dependency of growth to proficiency on status). This result is likely to hold regardless of the type of projection model used, as long as the model is fairly accurate at projecting future proficiency (in a sensitivity analysis we demonstrate that even Florida's inaccurate model produces substantively similar results).

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 very similar information. The reason for this has to do with how the GMPP's projection models measure growth. Rather than requiring all students to exceed the same fixed amount of growth, the required learning gain each student must make depends upon his or her initial achievement level. Therefore, just as 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 the school. For example, a school with a large number of very-low-performing students must help these students make larger gains to reach proficiency compared with the gains of students from another school who score closer to the proficiency cut score. Consequently, projection models are unlikely to have a large impact on how we measure schools' performances because they are not very different from the traditional status model.

Does figure 5 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' growth over time misguided, since it does not appear to provide much new information? Advocates of VAMs would

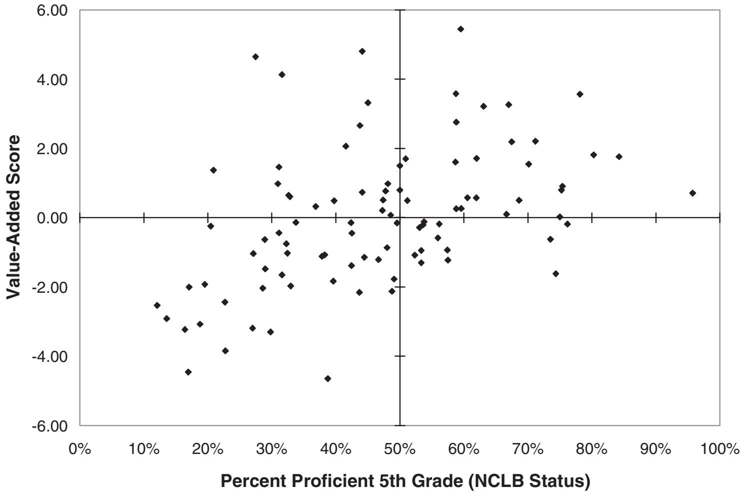

Figure 6. Percent Proficient vs. Value-Added Score (School Level)

suggest not. Like figure 5, figure 6 plots each school's status on the x-axis and its growth on the y-axis. As in figure 5, status is measured by the percentage of students in each school who were proficient in fifth grade. However, here growth is measured by the school's fifth-grade value-added score. Although the VAM is statistically complex, in this case it is a reasonable approximation to think of this measure as representing the average individual gains (on a standardized scale) that students in a particular school made between fourth and fifth grades, after adjusting for students' historical performance.

While status and value added are moderately positively related, figure 6 suggests 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, according to the VAM, students learned a lot relative to students with similar historical performances in other schools. However, because these students began the fifth-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 so they will reach proficiency, value-added advocates would argue that a nuanced accountability system might treat these schools differently than the low-status low-growth schools (lower left quadrant) that are more clearly underperforming. For example, consider the school in figure 6 with 27 percent of its students passing the state exam and a value-added score of 4.6. While the students in this school are making gains that are likely the third largest among all 90+ elementary schools in the study district, under NCLB this school would be labeled as failing. Notably, under a projection model (whether the "perfect" projection model or Florida's actual projection model) this school still ranks in the bottom third of the 90+ elementary schools in the study district. Students attending this school could be given the option to transfer, and if they chose to transfer they would likely attend a school where students are making smaller learning gains compared with the learning gains being made in their original school. In addition, if this school consistently performed the same way, it could be restructured even though it is possibly one of the most effective schools in the district.

The rules guiding the federal GMPP adhere to the core principle of NCLB: 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 their objectives sound similar to those of VAMs. Value-added models attempt to compare schools' relative effectiveness, and they attempt to judge schools based solely upon that which is within their control. Given the fact that projection models and VAMs sound similar, one of the major demonstrations in this research is that they yield very different assessments of schools' performances. Though both projection models and VAMs can be called growth models, not all growth models are alike.

It is well known that NCLB's traditional status model and VAMs represent two fairly different approaches to measuring school performance. Projection models seem like an interesting middle ground—they still hold all students to the proficiency standards (like a status model but unlike a VAM), yet they utilize longitudinal individual-level data to measure growth (like a VAM but unlike a status model). Perhaps the most important empirical finding from this research is that, although projection models and VAMs both utilize longitudinal student-level data and both measure growth, projection models are much more similar to the old NCLB status measure than they are to VAMs.

7. LIMITATIONS

There are two important limitations to this research worth discussing. First is the generalizability of this research, and second is the method used to evaluate the accuracy of projection models.

Generalizability

The analyses in this article were calculated using data from one large school district. It is possible that the findings regarding the accuracy of Florida's projection model in this district would not generalize to the state as a whole. While this concern is legitimate, the fact that the state's proficiency cut scores reflect a nonlinear scale and state average achievement scores by grade reflect nonlinear growth trajectories suggests that the bias of Florida's model is very likely to hold within the entire state. That said, the bias of Florida's model does not imply that other states' models are biased (in fact, findings from a larger study in Weiss 2008 suggest that using other states' models on Florida's data can yield more accurate, less biased, projections).

Assessing the Accuracy of Projection Models

It is critical to note that the method used in this article for assessing the accuracy and bias of projection models is only one interesting way to examine these models. While we believe it is a useful exercise to compare projected proficiency with observed proficiency when examining projection models, there is an important limitation to this analysis. Florida's projection model uses a student's gains from third to fifth grades to determine if she is on track to become proficient in sixth grade. If the projected proficiency does not reflect what is observed to happen at the end of sixth grade, this may reflect a poor projection model, but it also may reflect a particularly good or bad sixth-grade instructor. As such, it should not be expected that any projection model will be 100 percent accurate, because the sixth-grade instructors will influence the model's "accuracy" using this criterion. This may explain some of the inaccuracy of projection models, but it would require the vast majority of sixth-grade instructors be ineffective (compared with teachers in third through fifth grades) in order to produce the degree of bias demonstrated by this study. As such, it is more difficult to explain the projection model's apparent biases based on this limitation, although it is theoretically possible that a district's (or state's) sixth-grade instructors could be significantly less effective than its fourth- and fifth-grade instructors—yielding what appears to be a biased projection model. These limitations are important to keep in mind when considering the analyses presented here regarding the accuracy of projection models.

8. 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 highly criticized in large part because they do not take into account students' initial achievement levels. Because students enter schools at varying achievement levels, the required achievement gains are highly variable from student to student. As a result, schools with many initially low-performing students must be more effective than schools with many initially high-performing students. Many view these requirements (and their associated rewards, sanction, and assistance) as unfair because schools are largely judged by factors that are often beyond their control (i.e., students' initial achievement levels).

Growth models are a popular alternative measure of school performance. The federal government's GMPP allows states to use growth models but limits the type of models that are approved to those that require all students to become proficient in the near future. Consequently the projection models used under the GMPP are highly similar to NCLB's original status measure. In those states piloting projection models, the new models have resulted in little change in the percentage of schools making AYP. In Alaska, not a single school made AYP under the state's growth model that would not already have made AYP under the status model. In Arizona, less than 1 percent of schools made AYP based on growth alone. This is partly because the GMPP measures are applied only after status, safe harbor, confidence intervals, etc. are applied, but it is also because status and projection models produce very similar results. One state where the models seemed to be having a large impact is Florida. According to the Editorial Projects in Education 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. 24). This work helps to explain one of the possible reasons why Florida's model may have so many schools making AYP based on growth alone: Florida uses a biased projection model that falsely claims that many fourth- and fifth-grade students are on track to become proficient when they are not. When inaccurate and biased projection models are used to measure school performance, significant error is added to the accountability system. If states are to continue using projection models, the federal government should require them to demonstrate their model's accuracy and bias at the individual and school levels.

Even more accurate projection models are likely not to differ very much from a status model because, just like NCLB's original status measure, projection models are largely influenced by students' initial achievement levels. While status measures are reliable and accurate, projection models introduce additional noise to the system and have limited potential benefit to do anything more than mimic the status model. If the goal is to assess students' knowledge at a point in time, sticking with the old status model is probably the best bet. However, if we want to consider students' growth as a measure of school performance, the GMPP might consider allowing states to use VAMs of school performance.

That said, before the federal government begins a VAM pilot program and states start to claim that school X is more effective than the average school, more rigorous studies of VAMs of school effectiveness are needed. The majority of research on value-added modeling concentrates on estimating the contribution of teachers, not schools, to the learning gains of their students.

While this difference may seem inconsequential, one cannot assume that simply replacing the word *teacher* with the word *school* will extend the findings of researchers examining VAMs of teacher effectiveness. The application of VAMs to measure school 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 et al. 2007, p. 126). In contrast, at the school level most students tend to remain grouped together in a school from year to year. Consequently, compared with value-added models of teacher effectiveness, VAMs of school effectiveness may be less successful at separating out the unique contribution of schools to the learning gains of their students. The potential significance of this difference cannot be overstated and needs to be studied. Perhaps a more reasonable goal than attempting to compare schools' relative effectiveness would simply be to measure the average growth of students in each school. While the inferences to be drawn are far less monumental, the goal may be achievable, may even be reliable, and does not require any "heroic assumptions" (Rubin, Stuart, and Zanutto 2004, p. 111). Most important, before using VAMs of school effectiveness as part of the accountability system, these models should be put through the intense scrutiny that the GMPP models have yet to receive.

This research demonstrates the inaccuracy and bias of one state's growth model currently used under the GMPP. The results illuminate the challenge of attempting to project students' future proficiency and calls into question whether the benefits of using projection models are outweighed by the noise associated with these models. This concern is deepened by the fact that, even using an unbiased and accurate model, the results of projection models tend to be quite similar to the more reliable and accurate status measure. Finally, it is important to note that the GMPP's growth models (projection models) are both theoretically and substantively different from the value-added growth models advocated for by many researchers.

The authors would like to thank Margaret Goertz, Rebecca Maynard, Ruth Curran Neild, Russell Cole, Elliot Weinbaum, and Asya Magazinnik for their useful comments on this article. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C050041-05 to the University of Pennsylvania, as well as funding from the Carnegie Foundation. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education or the Carnegie Foundation.

REFERENCES

Alaska Department of Education and Early Development. 2006. *Executive summary: Alaska growth model proposal*. Available www.ed.gov/admins/lead/account/growthmodel/ak/akes2006.pdf. Accessed 10 October 2006.

Arizona Department of Education. 2007. *Proposal for a growth model to evaluate adequate yearly progress for schools and districts*. Available www.ed.gov/admins/lead/account/growthmodel/az/azgmp.doc. Accessed 10 December 2007.

Arkansas Department of Education. 2006. *Arkansas growth model proposal*. Available www.ed.gov/admins/lead/account/growthmodel/ar/argmp.doc. Accessed 10 October 2006.

Bloom, Howard, Carolyn J. Hill, Alison Reback Black, and Mark W. Lipsey. 2008. Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. MDRC Working Paper on Research Methodology.

Choi, Kilchan, Pete Goldschmidt, and Kyo Yamashiro. 2005. Exploring models of school performance: From theory to practice. In *Uses and misuses of data for educational accountability and improvement*, edited by Joan Herman and Edward Haertel, pp. 119–46. NSSE Yearbook. Malden, MA: Wiley-Blackwell Publishing.

Colorado Department of Education. 2008. *The Colorado growth model: Higher expectations for all students*. Available www2.ed.gov/admins/lead/account/growthmodel/co/cogrowthproposal101508.pdf. Accessed 21 July 2010.

Delaware Department of Education. 2006. *Delaware's proposal for a growth model resubmitted to U.S. Department of Education*. Available www.ed.gov/admins/lead/account/growthmodel/de/derevision112006.doc. Accessed 10 October 2006.

Dunn, Jennifer L., and Jessica Allen. 2008. The interaction of measurement, models and accountability: What are the NCLB growth models measuring? Paper presented at the National Council on Measurement in Education Annual Meeting, New York, March.

Florida Department of Education. 2004. *Assessment and accountability briefing book*. Available http://fcat.fldoe.org/pdf/fcataabb.pdf. Accessed 10 October 2006.

Florida Department of Education. 2006. *Florida's application for the NCLB growth model pilot: Peer review documentation*. Available www.ed.gov/admins/lead/account/growthmodel/fl/flrevisions2006.doc. Accessed 10 October 2006.

Hershberg, Theodore. 2005. Value-added assessment and systemic reform: A response to the challenge of human capital development. *Phi Delta Kappan* 87(4): 276–83.

Iowa Department of Education. 2007. *No Child Left Behind growth model pilot proposal: U.S. Department of Education*. Available www.ed.gov/admins/lead/account/growthmodel/ia/iagmp07.doc. Accessed 20 June 2007.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16(4): 91–114.

Klein, Alyson. 2007. Impact is slight for early states using "growth." *Education Week* (19 December): 24–25.

Linn, Robert L. 2004. Accountability models. In *Redesigning accountability systems for education*, edited by Susan Fuhrman and Richard F. Elmore, pp. 73–95. New York: Teachers College Press.

Lissitz, Robert W., Harold C. Doran, William D. Schafer, and Joseph Willhoft. 2006. Growth modeling, value added modeling and linking: An introduction. In *Longitudinal and value added models of student performance*, edited by Robert W. Lissitz, pp. 1–46. Maple Grove, MN: JAM Press.

Lockwood, J. R., Daniel F. McCaffrey, Louis T. Mariano, and Claude Setodji. 2007. Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics* 32(2): 125–50.

Michigan Department of Education. 2008. *Growth model pilot application for adequate yearly progress determinations under the No Child Left Behind Act*. Available www2.ed.gov/admins/lead/account/growthmodel/mi/migmp.doc. Accessed 21 July 2010.

Minnesota Department of Education. 2009. *Minnesota's adequate yearly progress (AYP) growth model application: Peer review documentation*. Available www2.ed.gov/admins/lead/account/growthmodel/mn/gmp1-8-2009.doc. Accessed 21 July 2010.

Missouri Department of Education. 2008. *State of Missouri's application for NCLB growth model implementation: Peer review documentation*. Available www2.ed.gov/admins/lead/account/growthmodel/mo/mogmp.doc. Accessed 21 July 2010.

North Carolina Department of Education. 2006. *North Carolina's proposal to pilot the use of a growth model for AYP purposes in 2005–06*. Available www.ed.gov/admins/lead/account/growthmodel/nc/ncgmp.doc. Accessed 10 October 2010.

Ohio Department of Education. 2006. *Proposal to the United States Department of Education for employing a growth model for No Child Left Behind accountability purposes*. Available www.ed.gov/admins/lead/account/growthmodel/oh/ohgmp07.doc. Accessed 20 June 2007.

Pennsylvania Department of Education. 2008. *Proposal to the US Department of Education for participation in the No Child Left Behind (NCLB) growth model pilot program*. Available www2.ed.gov/admins/lead/account/growthmodel/pa/pagmpfinal10-15-08.doc. Accessed 21 July 2010.

Porter, Andrew C., and Morgan S. Polikoff. 2007. NCLB: State interpretations, early effects, and suggestions for reauthorization. *Social Policy Report* 21(4): 3–10.

Raudenbush, Stephen W. 2004. What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics* 29(1): 121–29.

Rubin, Donald B., Elizabeth A. Stuart, and Elaine L. Zanutto. 2004. A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics* 29(1): 103–16.

Tekwe, Carmen D., Randy L. Carter, Chang-Xing Ma, James Algina, Maurice E. Lucas, Jeffrey Roth, Mario Ariet, Thomas Fisher, and Michael B. Resnick. 2004. An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics* 29(1): 11–36.

Tennessee Department of Education. 2006. *Proposal to the U.S. Department of Education: NCLB growth model pilot program*. Available www.ed.gov/admins/lead/account/growthmodel/tn/tngmp.doc. Accessed 10 October 2006.

Texas Department of Education. 2009. *Texas Education Agency growth model pilot application for adequate yearly progress determinations under the No Child Left Behind Act*. Available www2.ed.gov/admins/lead/account/growthmodel/tx/txproposrpt12122009.pdf. Accessed 21 July 2010.

U.S. Department of Education (USDOE). 2005. *Secretary Spellings announces growth model pilot, addresses chief state school officers' annual policy forum in Richmond*. Press Releases. Archived Information. Available www.ed.gov/news/pressreleases/2005/11/11182005.html. Accessed 9 August 2007.

U.S. Department of Education (USDOE). 2007. *Secretary Spellings invites eligible states to submit innovative models for expanded growth model pilot*. Press Releases. Archived Information. Available www.ed.gov/news/pressreleases/2007/12/12072007.html. Accessed 2 December 2007.

U.S. Department of Education (USDOE). 2008. *U.S. Secretary of Education Margaret Spellings announces No Child Left Behind "differentiated accountability" pilot*. Press Releases. Archived Information. Available www.ed.gov/news/pressreleases/2008/03/03182008.html. Accessed 19 March 2008.

Weiss, Michael J. 2008. Using a yardstick to measure a meter: Growth, projection, and value-added models in the context of school accountability. PhD dissertation, University of Pennsylvania, Philadelphia, PA.

Wright, S. Paul, William L. Sanders, and June C. Rivers. 2006. Measurement of academic growth of individual students toward variable and meaningful academic standards. In *Longitudinal and value added models of student performance*, edited by Robert W. Lissitz, pp. 385–406. Maple Grove, MN: JAM Press.

This article has been cited by:

1. Morgan S. Polikoff, Stephani L. Wrabel. 2013. When is 100% not 100%? The Use of Safe Harbor to Make Adequate Yearly Progress. *Education Finance and Policy* **8**:2, 251-270. [Abstract] [Full Text] [PDF] [PDF Plus]