Title: Holding Schools Accountable for the Growth of Non-proficient Students: Coordinating Measurement and Accountability

Author(s): Jennifer L. Dunn, *Measured Progress,* Jessica Allen, *University of Colorado, Boulder*

Date of Initial Publication: Winter/2009 in Educational Measurement: Issues and Practice Winter 2009, Vol. 28, No. 4, pp. 27–41

Abstract/Summary

A key intent of the NCLB growth pilot is to reward low-status schools who are closing the gap to proficiency. In this article, we demonstrate that the capability of proposed models to identify those schools depends on how the growth model is incorporated into accountability decisions. Six pilot-approved growth models were applied to vertically scaled mathematics assessment data from a single state collected over 2 years. Student and school classifications were compared across models. Accountability classifications using status and growth to proficiency as defined by each model were considered from two perspectives. The first involved adding the number of students moving toward proficiency to the count of proficient students, while the second involved a multitier accountability system where each school was first held accountable for status and then held accountable for the growth of their nonproficient students. Our findings emphasize the importance of evaluating status and growth independently when attempting to identify low-status schools with insufficient growth among non-proficient students.

Subject/Keywords:

Document Type: Paper

Document Archive Number: 0017cdedu2009

Educational Measurement: Issues and Practice
Winter 2009, Vol. 28, No. 4, pp. 27–41

Holding Schools Accountable for the Growth of Nonproficient Students: Coordinating Measurement and Accountability

Jennifer L. Dunn, *Measured Progress*
Jessica Allen, *University of Colorado, Boulder*

A key intent of the NCLB growth pilot is to reward low-status schools who are closing the gap to proficiency. In this article, we demonstrate that the capability of proposed models to identify those schools depends on how the growth model is incorporated into accountability decisions. Six pilot-approved growth models were applied to vertically scaled mathematics assessment data from a single state collected over 2 years. Student and school classifications were compared across models. Accountability classifications using status and growth to proficiency as defined by each model were considered from two perspectives. The first involved adding the number of students moving toward proficiency to the count of proficient students, while the second involved a multitier accountability system where each school was first held accountable for status and then held accountable for the growth of their nonproficient students. Our findings emphasize the importance of evaluating status and growth independently when attempting to identify low-status schools with insufficient growth among nonproficient students.

Keywords: growth, NCLB, accountability

A key intention of the *No Child Left Behind* Act of 2001 (NCLB) is to hold all schools accountable for ensuring that all students achieve proficiency by 2013–2014. Given the NCLB requirement that all states administer annual assessments in reading/language arts and mathematics in grades 3 through 8, many have argued that the focus of NCLB should center on holding schools accountable for moving students toward proficiency, not simply attaining proficiency. This has led to a shift in accountability values, namely that schools be held accountable for the growth of their students. The interest in growth models was further heightened when the U.S. Department of Education (USED) announced that a sample of states would be allowed to pilot a growth-based accountability system in making adequate yearly progress (AYP) determinations (Spellings, 2005; U.S. Department of Education, 2005).

Seven core principles, outlined by USED, were used to ensure that the proposed growth models were technically sound, could be validly incorporated into school accountability systems, and were consistent with the spirit of NCLB (U.S. Department of Education, 2006a, 2006b). The growth models must: (1) ensure all students are proficient by 2013–2014 and set annual goals to ensure that the achievement gap is closing for all groups of students; (2) establish high expectations for low-achieving students regardless of student demographic characteristics and school characteristics; (3) produce separate accountability decisions about student achievement in reading/language arts and in mathematics; (4) include all students and hold schools and districts accountable for subgroup performance; (5) be based on annual assessments, comparable from year to year, in each of grades 3 through 8 and high school in both reading/language arts and mathematics that have been operational for more than 1 year and approved through the NCLB peer review process; (6) track student progress across years as part of the state data system; and (7) include student participation rates and student

Jennifer L. Dunn is a Psychometrician at Measured Progress, 100 Education Way, Dover, NH 03820; dunn.jennifer@measuredprogress.org. Jessica Allen is a doctoral candidate in Research and Evaluation Methodology at the School of Education, 249 UCB, University of Colorado at Boulder, Boulder, CO 80309-0249.

performance on an additional academic indicator.

Many argue that growth-based models are more valid than the current NCLB status and improvement models because they hold schools accountable for the amount of student learning or development that occurs beyond the achievement of those same students when they entered (Carlson, 2006; Goldschmidt et al., 2005; Hill et al., 2006). However, the NCLB core principles have modified the more traditional conceptualization of growth to align with proficiency. Instead of evaluating schools according to the amount of student learning or development that occurs beyond a student's achievement when they entered, the NCLB core principles center on evaluating growth to proficiency and the maintenance of proficiency. According to the core principles, a successful NCLB growth model should be able to distinguish schools that have nonproficient students who are closing the gap to proficiency and proficient students who are maintaining proficiency from schools that have nonproficient students who are not closing the gap to proficiency and proficient students who are falling below proficiency.

The validity of the accountability system rests on whether the defined construct is measured. Are the NCLB growth models detecting the type of growth educators and policy makers are interested in? A clear answer to this question cannot be obtained without examining the specifics of the models and how they are used to incorporate growth into the accountability system. Growth can be used to hold schools accountable in many different ways, each of which will influence the type of measurement model used. For example, growth could be used as a replacement for status. In this case, under the NCLB context, the growth measurement model would need to incorporate proficiency and growth toward proficiency. Alternatively, growth could be used in conjunction with status. In this case, a school could be deemed successful if it meets or exceeds the status requirements or the growth requirements. The growth model in this latter system might look quite different from one designed to replace status. In either case, an evaluation of whether or not growth toward proficiency is being measured depends on the measurement model used and how the scores generated by the measurement model are handled in accountability.

The purpose of this article is to explore the properties of a selection of NCLB growth pilot models applied to the same data set in hopes of gaining a better understanding of how they are measuring growth to proficiency. Of particular interest are the characteristics of the growth models, how the models are used to make accountability decisions, and the construct measured by those decisions. By carefully examining the characteristics of the growth models, the scores they generate, and how said scores are incorporated into the accountability system, we hope to provide some insight into the potential of growth models for improving the validity of NCLB accountability.

NCLB Growth Models

As of June 2007, nine states (AK, AZ, IA, OH, TN, NC, DE, AR, and FL) had been approved to use growth models in 2007–2008 NCLB accountability decisions. While traditional growth models tend to value all growth equally, the NCLB pilot models value growth to proficiency and the maintenance of proficiency. In each of the models, proficient students are either automatically counted as meeting their growth targets or are assigned the maximum possible growth score. A school growth score is defined using one of two methods: the number of proficient students plus the number of nonproficient students who meet their growth target divided by the total number of students in the school (proficient plus model), or the average number of points assigned to each student, where proficient students are assigned the maximum number of points, and nonproficient students are assigned a proportion of points depending on how much of the distance to proficiency was reduced.

Although the school-level calculations are consistent from one model to the next, the amount of change each student must demonstrate to be counted as meeting his/her growth target varies substantially. The following section reviews the student-level growth calculations for each of the pilot states.

Student-Level Growth Calculations

Alaska

Under Alaska's growth model, a student growth projection is calculated to determine if a student is on track to proficiency. Students in grades 3 through 6 are expected to be proficient in 4 years and students in grade 7 and above are expected to be proficient by the tenth grade. A projection is calculated for each student using an adjusted baseline test score and the proficiency score for the grade by which the student is expected to be proficient:

$$\text{Adjusted baseline} = \text{grand mean} + \text{reliability}^{*}(\text{observed score} - \text{grand mean}).$$

Each student is expected to make up the difference between his/her adjusted baseline score and when he/she must be proficient in equal steps. For example, if a student in grade 3 scored 401 on a test of reliability 0.93 with a grand mean of 441.19, then the adjusted baseline score for that student is 404. Students who are in grades 3 through 6 in the baseline year are required to make up one quarter of the distance to proficiency each year. Students who are in grade 7 in the baseline year are required to make up one third of the distance to proficiency and students who are in grade 8 are required to make up one half of the distance to proficiency each year. Continuing the above example, assuming a grade 7 proficiency target of 472, the student's growth target would be 421. If a student scores at or above the expected score he/she is counted as meeting his/her growth target. Regardless of how a student performs in the future, the initial projection does not change (Alaska Department of Education, 2007a,b; Alaska Response, 2007).

Arizona

Arizona's growth model incorporates growth targets and predicted scores. First, a growth target is calculated for each student using his/her baseline score and the grade by which proficiency is expected. Students in grade 3 are expected to be proficient by grade 6, students in grade 4 are expected to be proficient by grade 7, and students in grades 5 through 7 are expected to be proficient by grade 8. Students are expected to close the gap to proficiency by making equal growth increments from year to year. A student's expected growth increment is calculated by taking the proficiency score in the grade by which he/she is expected to be proficient minus his/her baseline score divided by the number of years they have to become proficient. The annual growth target is calculated by adding the growth increment to the baseline score. For example, a grade 3 student with a score of 401, would be expected

to score 466 by grade 6, resulting in a growth target of 423 in grade 4.

A predicted score is also calculated for each student by regressing the current year score on the score from the previous year. The relationship between last year's scores and the current year's scores are calculated using the following equation:

$$\text{Scale Score}_{it} = \alpha_j + \beta\ \text{Scale Score}_{i,t-1} + \varepsilon_i$$

where Scale Score$_{it}$ is the scale score of student i on the state assessment for the current year, Scale Score$_{i,t-1}$ is the scale score of student i on the state assessment for the previous year, α_j is the fixed effect for school j, and ε_i is a normally distributed error term. The coefficients from the regression equation are used to produce predicted scores for each student. A confidence interval (97.5%) is placed around the predicted score and the lower bound is used to determine whether or not a student has made sufficient growth:

$$\text{Lower bound}_{it} = \text{predicted score}_{it} - (t_{it^*}\ \text{standard error of prediction}_{it})$$

where i represents student of interest and t_{it} represents the 97.5th percentile of the t-distribution. If the lower bound is greater than or equal to the student's growth target, the student is counted as on track to proficiency. Continuing the example from before, and assuming a fixed effect of 105.12 and a regression coefficient of 0.811, the student's predicted score would be 430.33 with a lower bound of 421.96. Since 422 is less than 423, the student is classified as not meeting her growth target (Arizona Department of Education, 2007a,b).

Arkansas

Under Arkansas's growth model, all students are expected to be proficient by the eighth grade. A student's growth interval is calculated using the following formula:

$$\text{Growth interval} = [(P_b - P_a)/(P_8 - P_a)] * (P_8 - X)$$

where P_b is the proficiency scaled score in the student's subsequent grade, P_a is the proficiency scaled score for the student's current grade, P_8 is the proficiency scaled score in grade 8, and X is the student's current scaled score. A student's growth target is calculated by adding this growth interval to his/her current score. If the student's subsequent test score is greater than or equal to the growth target, then the student is counted as meeting his/her growth requirement. For example, if P_a equals 407, P_b equals 432, and P_8 equals 487, a student with a score of 401 would have a growth interval of 27 and a growth target of 428 Arkansas Department of Education, 2006a,b,c).

Delaware

A value table is used to measure growth in Delaware by assigning points to students based on whether or not they changed performance levels from one year to the next. The basic idea behind using a value table for accountability purposes is to look at the achievement level a student earns in one year, compare it to the achievement level earned the previous year, and then assign a numerical value to that change (Hill et al., 2006). Higher numerical values are assigned to movements that are more highly valued. For example, in Delaware's value table (Table 1), a student earns 300 points for scoring proficient or above regardless of his/her year 1 score; zero points for scoring below proficient, if they had previously scored proficient or above; zero points for staying in a performance category; zero points for moving down a performance category; 150 points for improving from Level 1A to 1B; 225 for improving from Level 1A to 2A; 250 for improving from Level 1A to 2B; 175 for improving from Level 1B to 2A; 225 for improving from Level 1B to 2B; and 200 points for improving from Level 2A to 2B. The school score is the average of the points earned by the students in that school (Delaware Department of Education, 2006).

Florida

All students are expected to be proficient in 3 years or by Grade 10 under Florida's growth model. The difference between the proficiency scaled score in the grade by which proficiency is expected and the baseline scaled score is divided by the number of years the student has to become proficient. A student's growth target is calculated by adding this difference to his/her baseline scaled score. Continuing our previous example, a student with a grade 3 score of 401, expected to score 466 by grade 6, would have a growth target of 423. If the growth target is less then or equal to the student's subsequent test score, then the student is counted as meeting his/her growth requirement (Florida Department of Education, 2007).

Iowa

Iowa uses a value table approach to track student movement between proficiency classifications from year to year. Nonproficient students receive credit for making growth if they move from a lower achievement category to a higher achievement category. Iowa's assessment system classifies students into one of three proficiency levels (*weak*, *marginal*, and *proficient*). To more precisely track the growth of nonproficient students, the marginal category was divided into *low marginal* (more than one standard error below the marginal cut) and *high marginal* (within a standard error of the marginal cut). Students who move from *weak* to any of the other performance levels, and students who move from *low marginal* to *high marginal* or *proficient* are classified as meeting their growth targets.

Students do not get credit if they have received credit for movement between two categories previously. For example, if over 4 years a student has the following performance pattern, *weak*, *low marginal*, *weak*, *low marginal*, the student only gets credit the first time he/she move from *weak*

Table 1. Delaware's Value Table

Year 1 Level	Year 2 Level				
	Well Below the Standard		Below the Standard		Proficient or Above
	1A	1B	2A	2B	3, 4 or 5
1A	0	150	225	250	300
1B	0	0	175	225	300
2A	0	0	0	200	300
2B	0	0	0	0	300
3, 4 or 5	0	0	0	0	300

to *low marginal*. The second time the student moves from *weak* to *low marginal*, he/she is classified as nonproficient. This lack of credit safeguards against students getting growth credit for bouncing across proficiency levels (Iowa Department of Education, 2006).

North Carolina

In North Carolina, students are expected to be proficient in 4 years or by the end of Algebra I for math; English I for English Language Arts. Although North Carolina's model officially implements a 4-year approach, it makes use of a pretest, making it mathematically comparable to the 3-year approaches. For example, the baseline assessment for third-grade students is a third-grade test given at the beginning of the year. For all of the remaining grades, the baseline assessment is the end of year assessment from the previous year. Third-grade students are expected to score proficient on the sixth-grade assessment; fourth-grade students, with third-grade baseline scores, are expected to score proficient on the seventh-grade assessment; fifth-grade students (fourth-grade baseline scores) are expected to score proficient on the eighth-grade assessment; and all other students are expected to score proficient on the Algebra I and English I assessments.

To calculate individual growth increments and growth targets, North Carolina uses a modified z-scale approach called the C-scale. Each scaled score is transferred to the C-scale using the mean and standard deviation of test scores on the year the test was initially administered. For example, a student with a scaled score of 401 has a C-scale score of -0.91 ($M = 441.9, SD = 44.30$). The difference between the C-scale equivalent of the proficiency score for the grade by which the student is expected to be proficient and the C-scale equivalent of the student's score in the baseline year is divided by the number of years the student has to meet the proficiency target score. A grade 7 proficiency cut score of 472 corresponds to a C-scale cut of -0.35 ($M = 487.30, SD = 44.16$). A student's growth target is calculated by adding this difference to the C-scale equivalent of the student's baseline score. The resulting target for our example student would be -0.77 on the C-scale or 432 on the grade 4 scaled score metric ($M = 459.79, SD = 36.65$). A student meets his/her growth target if his/her C-scale assessment score from the following year is greater than or equal to his/her C-scale growth target for that year (North Carolina Department of Education, 2006).

Ohio

Ohio uses a projection model where all available test scores for the student are used to project each student's score at some point in the future. Estimates are made using a hierarchical regression with fixed effects for districts and schools. If the projected score is equal to or greater than the proficient cut for the grade the student would be enrolled at the time of the projection, the student is classified as meeting his/her growth target. Ohio's growth model projections are recalculated for each student every year. Students are considered proficient if their projected score is at or above the proficiency standard in 4 years or the grade beyond the highest grade of their current school (Ohio Department of Education, 2006, 2007a,b).

Tennessee

Tennessee also uses a projection model that is recalculated each year for each student and each test. The same regression methodology used in the Ohio model is used for the Tennessee model. Tennessee's growth model requires fourth and fifth graders to have projections at or above the seventh- and eighth-grade proficiency targets, respectively. Students in grades 6–8 must have projections at or above their graduation exam proficiency targets (Tennessee Department of Education, 2006).

Growth Model Summary

In summary, the growth models vary in how they calculate student growth targets from state to state according to three elements: (1) the number of years by which a student is expected to be proficient; (2) the final grade included in the growth model; and (3) the statistical model used to calculate change. Each of these elements is summarized in Table 2.

It appears, from the above comparisons, that although the manner in which schools are held accountable for growth is fairly consistent, the mathematical meaning of "growth to proficiency" varies from one state to the next, at least from a theoretical

Table 2. Summary of Three Elements Across Growth Models in Nine States

Growth Model	Number of Years to Achieve Proficiency	Final Grade Included in Model	Statistical Model
DE	N/A	N/A	Value table
AZ	3	8	Vertically scaled score corrected for regression to the mean
FL	3	10	Change in vertically scaled score
NC	3	9	Change in standard score
IA	Depends on initial distance from proficiency	N/A	Value table
AK	4	10	Standard score with correction for test reliability
AR	Depends on grade	8	Change in vertically scaled score
OH	4	Depends on school configuration	Hierarchical projection (fixed school effects)
TN	3	High school graduation	Hierarchical projection (fixed school effects)

standpoint. Despite these apparent variations, some skepticism has been generated about the flexibility of the models, the reasonableness of the growth targets, and the ability of the models to capture "growth to proficiency." The models have been criticized for their lack of flexibility, their unrealistic growth targets (proficient in 3 years or less), and their similarity to status (Dunn, 2007). The purpose of this study is to investigate the validity of these claims by applying the models to the same sample of vertically scaled mathematics assessment data from a single state collected over 2 years and to explore the possibility of using the above growth-to-proficiency models independently of status instead of the widely adopted "proficient plus" model. Specifically, by comparing the models at both the student and school level, we hope to learn:

1. Do the models vary in terms of how they classify schools and students?
2. Do the models have realistic growth expectations for schools and students?
3. Do the models capture growth to proficiency at the student and school levels?
4. Does meaning of growth to proficiency vary across models?

Three of the NCLB pilot models were not included in this study. A primary goal of this study is to examine growth model classifications at both school and student levels. Delaware's model is designed as a school-level growth model. Although students are assigned points based on their change in performance level, a growth target is not calculated for each student. This, coupled with the fact that Delaware calculates an average for a school growth score, led to the exclusion of the model. In addition, because the projection model is difficult to replicate without the use of proprietary software, and the authors felt a rough approximation would be insufficient, Tennessee and Ohio's models were also excluded from this study. The growth to proficiency comparisons therefore focus on the student and school scores generated by the proposed models for Alaska, Arizona, Arkansas, Florida, North Carolina, and Iowa.

Method

Source of Information

A sample of vertically scaled mathematics student scores from a state standards-based mathematics assessment administered annually under standardized conditions to all students enrolled in grades 3 through 8 was examined across 2 years. Students were required to have test scores in two consecutive years (not necessarily the same school) to be included in the sample. Students who were retained were omitted from the study. A total of 38,933 students from 140 schools were included in this study.

Student-Level Calculations

A proficiency classification, six growth target classifications, a standardized status score, and a growth percentile were calculated for each student. Students who scored at or above the third performance level were classified as proficient while students scoring in the first and second category were classified as nonproficient. Student growth targets were calculated using the growth target specifications outlined by Alaska, Arizona, Arkansas, Florida, North Carolina and Iowa. Students who scored at or above their respective growth targets (in the second year) were classified as meeting their growth target; otherwise, they were classified as not meeting their growth target. Students classified as proficient in year two were automatically classified as meeting their growth target. A standardized year two status score, designed to reflect the distance between each student's score and the proficiency cut, was calculated by determining the difference between the scaled score and the proficiency cut and then dividing by the standard deviation of the scaled scores. A growth percentile was calculated to estimate the growth of each student (see Betebenner, 2009: this issue, pp. 42–51). Conditional growth percentiles are often used in calculating pediatric reference growth charts, which have been widely used to help parents understand their child's height or weight in relation to a population of children of the same age. In achievement settings, growth percentiles are calculated using quintile regression techniques where growth is conditioned on a student's prior test score. A student's growth percentile is given by

$$Q_{Y_i}(\tau \mid Y_i, x_i) = g_\tau(i,t) + \alpha(\tau)Y_i + x_i'\beta(\tau)$$

where Y_i is the ith student's score from the previous year, g_τ is a nonparametric trend component, and $\alpha(\tau)$ reflects the relationship between the two tests (Betebenner & Shang, 2007). It is an estimate of the percentile at which a student is growing in relation to other students with the same initial test score. The primary advantage of growth percentiles over simple change scores is that they correct for aberrant changes in student achievement.

School-Level Calculations

The student data were aggregated at the school level to obtain 13 school scores: (1) the percent of students scoring at or above the proficient level using the second year of test scores, (2 through 7) the percentage of students classified as meeting their growth targets under each model, and (8 through 13) the percent of nonproficient students classified as meeting their growth targets under each model.

To compare school classifications, three annual measurable objectives (AMOs) were calculated using the statutory process outlined by USED. Each AMO was based on a set percentile (e.g., 20th) of the total student enrollment. Schools were ranked according to the percent of students scoring at or above proficiency from lowest to highest. The cumulative number of students enrolled in each school was tallied and the score (percent of students scoring at or above proficient) of the school that corresponded to appropriate percentile of the total student enrolment was set as the AMO. Because the school classifications are likely to be influenced by the percentile selected, the above process was replicated using the 20th, 50th, and 80th percentiles. This led to AMOs of 58%, 72%, and 80% proficient or above, representing the NCLB starting point, an approximation of current criteria, and a step toward the goal of 100% proficient by 2014, respectively. Schools where the percent of students meeting their growth targets is at or above the AMO are classified as meeting AYP while schools with the percent of students meeting their growth targets below this point are classified as not meeting AYP.

Analyses

School and student scores and classifications from the six different growth models were compared using descriptive statistics across the three AMOs. The relationships among the student growth to proficiency classifications, status and growth were compared

Table 3. Relationships (Cohen's kappa) Among School-Level Classifications Across Growth Models (*N* = 140)

Growth Model	AK	AR	AZ	FL	IA
AR	0.87				
AZ	0.85	0.90			
FL	0.97	0.89	0.86		
IA	0.78	0.88	0.81	0.80	
NC	0.62	0.74	0.65	0.63	0.82

Table 4. Relationships (Cohen's kappa) Among Student Growth Target Classifications Across Growth Models (*N* = 38,993)

Growth Model	AK	AR	AZ	FL	IA
AR	0.88				
AZ	0.90	0.91			
FL	0.96	0.85	0.88		
IA	0.85	0.86	0.84	0.84	
NC	0.82	0.85	0.77	0.81	0.84

Table 5. Relationships (Cohen's kappa) Among Student Growth Target Classifications Across Growth Models for Nonproficient Students (*N* = 11,926)

Growth Model	AK	AR	AZ	FL	IA
AR	0.55				
AZ	0.54	0.52			
FL	0.86	0.48	0.48		
IA	0.49	0.49	0.29	0.50	
NC	0.46	0.54	0.17	0.45	0.54

graphically across models. Particular attention was given to the relationships that status and growth (measured by growth percentile) have with the meets/does not meet decision of the various growth models for the nonproficient students. Finally, an alternative approach for incorporating growth to proficiency into accountability decisions, where schools are first held accountable for status and then held accountable for the growth of their nonproficient students, is examined.

Results and Discussion

The relationship between school classifications across growth models was examined using Cohen's coefficient (kappa) (Table 3). Kappa (κ) assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1 - \sum_i C_{i.}C_{.i}}$$

where $C_{i.}$ represents the number of schools classified as meeting the first AMO using the first model, $C_{.i}$ is the schools classified as meeting the first AMO using the second model, and C_{ii} is the schools classified as meeting the first AMO under both models.

The different growth models do not result in substantial differences in school-level AYP classifications. In particular, it appears that the school-level classifications made using Florida's growth model calculations are almost identical to those of Alaska's growth model. Interestingly, the school-level classifications using North Carolina's growth model appear to show the most variation compared to the classifications of the other growth models.

The relationship between student-level classifications (whether or not each student met his/her growth target) under the six models was examined using Cohen's coefficient (kappa) (Table 4).

Overall, the six growth models appear to classify students in similar ways. This result is somewhat counter intuitive, given the variation in the mathematical calculations that underlie the models. The similarity of the classifications may be due to one of at least two factors: (1) the proficient students are included in the classifications and (2) the growth target expectations may be too high. By including the proficient students in the comparisons, the variability in the student-level classifications across models is restricted. All proficient students are classified as meeting their growth targets, regardless of the model used. Therefore, by definition, the models agree on the classification of proficient students. Given that almost 70% of the students in the sample are scoring at or above proficient, by definition the models agree on over 70% of the students. When this occurs, it is difficult to determine whether or not the models agree on the classifications of the nonproficient students. The relationship between the growth target classifications of nonproficient students across models is outlined in Table 5.

In contrast to Table 4, the growth models do not classify the nonproficient students in similar ways. This means that not only do the models vary in terms of how they classify nonproficient students but some of the nonproficient students are able to meet their growth targets. It appears that the proficient students may have been masking the variations between the student-level growth model classifications.

From the results thus far, the classifications of nonproficient students vary across models. The relationship between these classifications with status and growth bears further investigation. In an ideal world, examining these relationships would center on the amount of growth made by each student. Unfortunately, absolute growth is a difficult construct to measure. A number of sophisticated growth models (i.e., hierarchical linear models, value added

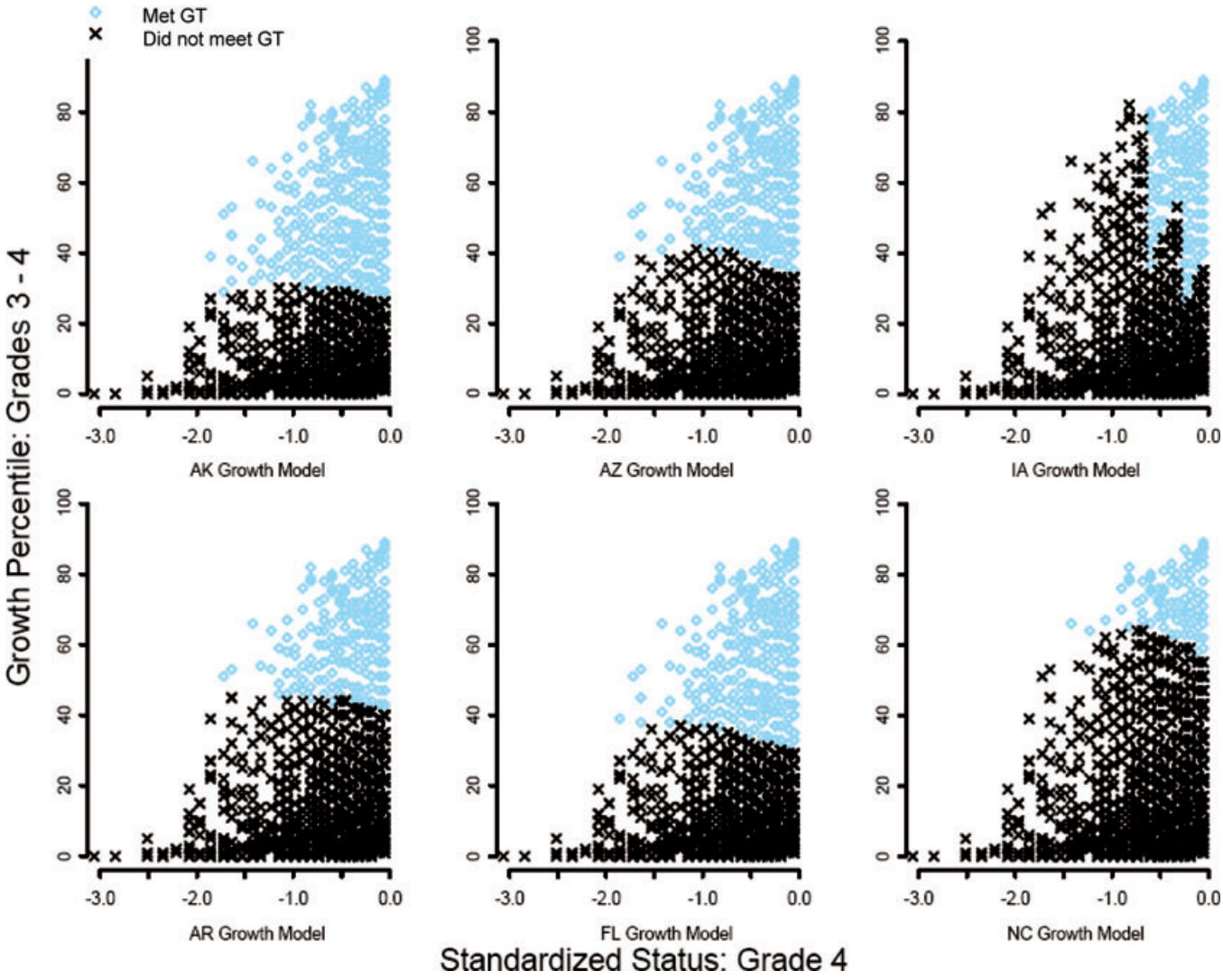

FIGURE 1. Growth to proficiency classification by status score and growth percentile, across models, for students in Grade 4.

models, and other multilevel models) have been developed to better measure change in achievement over time. These growth models quantify different aspects of change in student achievement and can have differential impacts on the estimated growth of the student (Goldschmidt et al., 2005). Growth percentiles, which reflect the change in the distribution of achievement over time conditioned on an initial measurement, were used to quantify change in student achievement. Percentiles help to counter regression to the mean, as do many of the sophisticated growth models, by correcting for aberrant changes in student achievement. For our purposes growth is conditioned on a student's prior test score resulting in a growth percentile that is an estimate of the percentile at which a student is growing in relation to other students with the same initial test score. All students were used to estimate the growth percentiles.

The relationship between the growth-to-proficiency classifications with growth percentile and scaled score were examined graphically for nonproficient students across models in Grades 4 through 8, respectively (Figures 1 through 5). Students classified as meeting their respective growth to proficiency targets are denoted by a gray diamond, while students classified as not meeting their respective targets are depicted by a black cross. The vertical axis represents the growth percentile, while the horizontal axis represents the relationship between the students second-year scaled score and the proficiency cut. Student growth classifications using Alaska's model are denoted in the upper left graph, student growth classifications using Arizona's model are denoted in the upper middle graph, student growth classifications using Iowa's model are denoted in the upper right graph, student growth classifications using Arkansas's model are denoted in the lower left graph, student growth classifications using Florida's model are denoted in the lower middle graph, and student growth classifications using North Carolina's model are denoted in the lower right graph. Although the student growth to proficiency classifications change from one graph to the next within each figure, the observations (student scores) do not. By comparing the graphs within each figure, insight into how each state has operationalized growth to proficiency and the relationship between growth to proficiency with status and growth percentile can be gained. It is important to note that the growth represented in these graphs does not reflect the

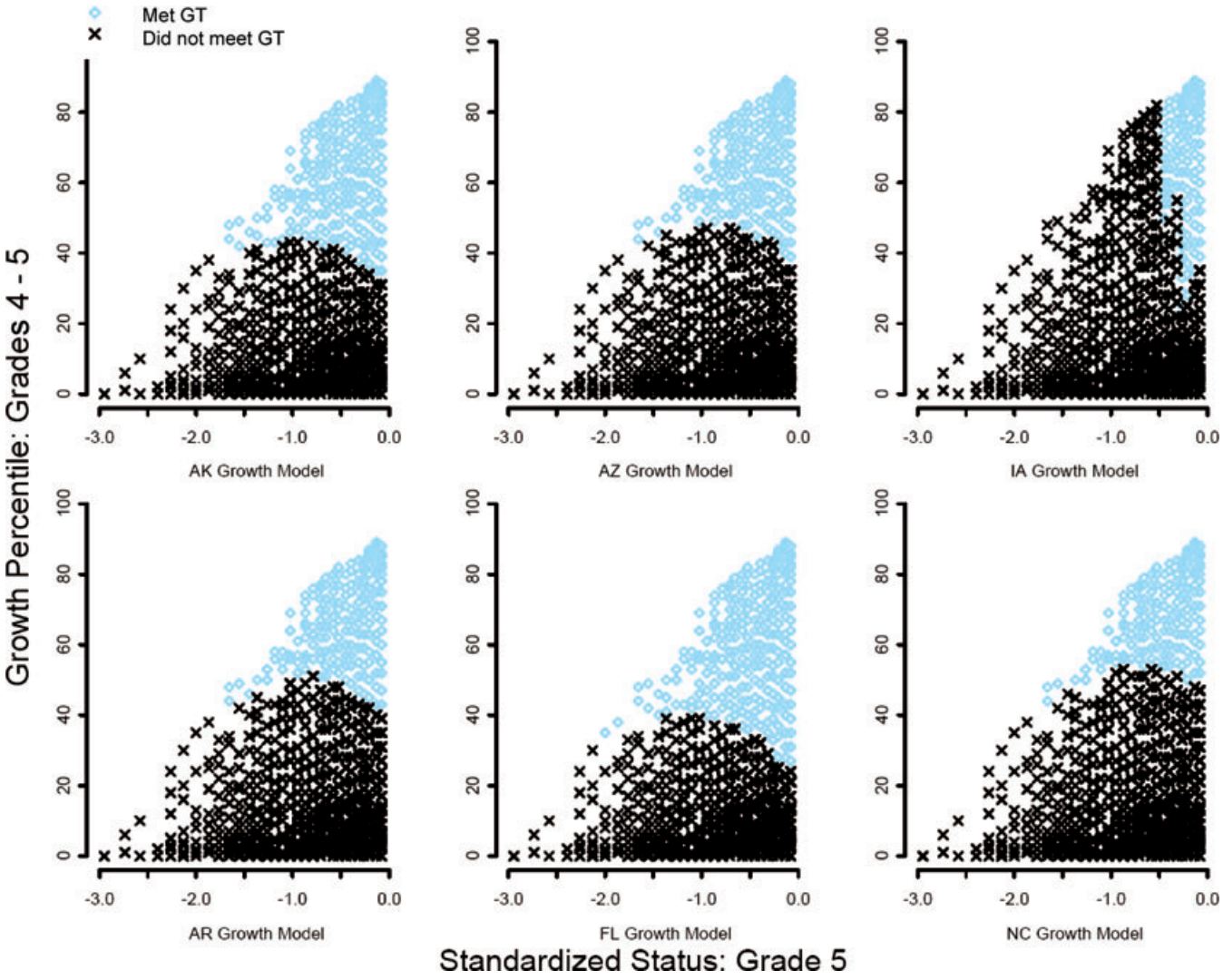

FIGURE 2. Growth to proficiency classification by status score and growth percentile, across models, for students in Grade 5.

amount of scaled score growth but rather how much each student grew in relation to other students with similar initial achievement.

Although the student classifications are not identical across grades and models, there is a high degree of similarity. In general, nonproficient students with high growth percentiles are classified as meeting their growth targets while students with low growth percentiles are not. This trend tends to hold true regardless of the distance between the student's score and proficiency (year two status). Overall, it appears low-status nonproficient students are required to meet similar growth percentiles as high-status nonproficient students.

There are some distinct exceptions to these trends. Most notably, the student classifications under Iowa's model have a stronger relationship to status than the other models. Students with high growth percentiles are only classified as meeting the growth to proficiency target when they also have high-status scores. This relationship to status is likely due to the dependency of Iowa's model on performance level classifications. Students can only be classified as meeting a growth target if they increase a performance level. Therefore, students falling just below the performance level cut in the first year will be classified as meeting their growth targets simply by scoring above the proficiency cut in the second year. Although such students increased their proficiency classification, they did not necessarily do so by growing a substantial amount.

The number of nonproficient students meeting growth targets varies across grades within each state. The growth percentile required to be classified as meeting a growth target varies across grades within state. Although the same growth model is used in each grade, the amount of required growth varies. In one grade, a student might be required to grow at the seventieth percentile to be classified as meeting his or her growth target, but in another grade, he or she would only need to grow at the fiftieth percentile. This trend is taken to the extreme in Arizona and Arkansas where none of the students are classified as meeting their growth targets in the seventh and/or eighth grades. While this trend is likely affected by the relationship between the growth model and the location of the proficiency cut across grades, the driving force appears to be the final grade by which students are expected to score proficient. In Arizona and Arkansas, all students are expected to score at or above proficient by Grade 8. By setting these expectations,

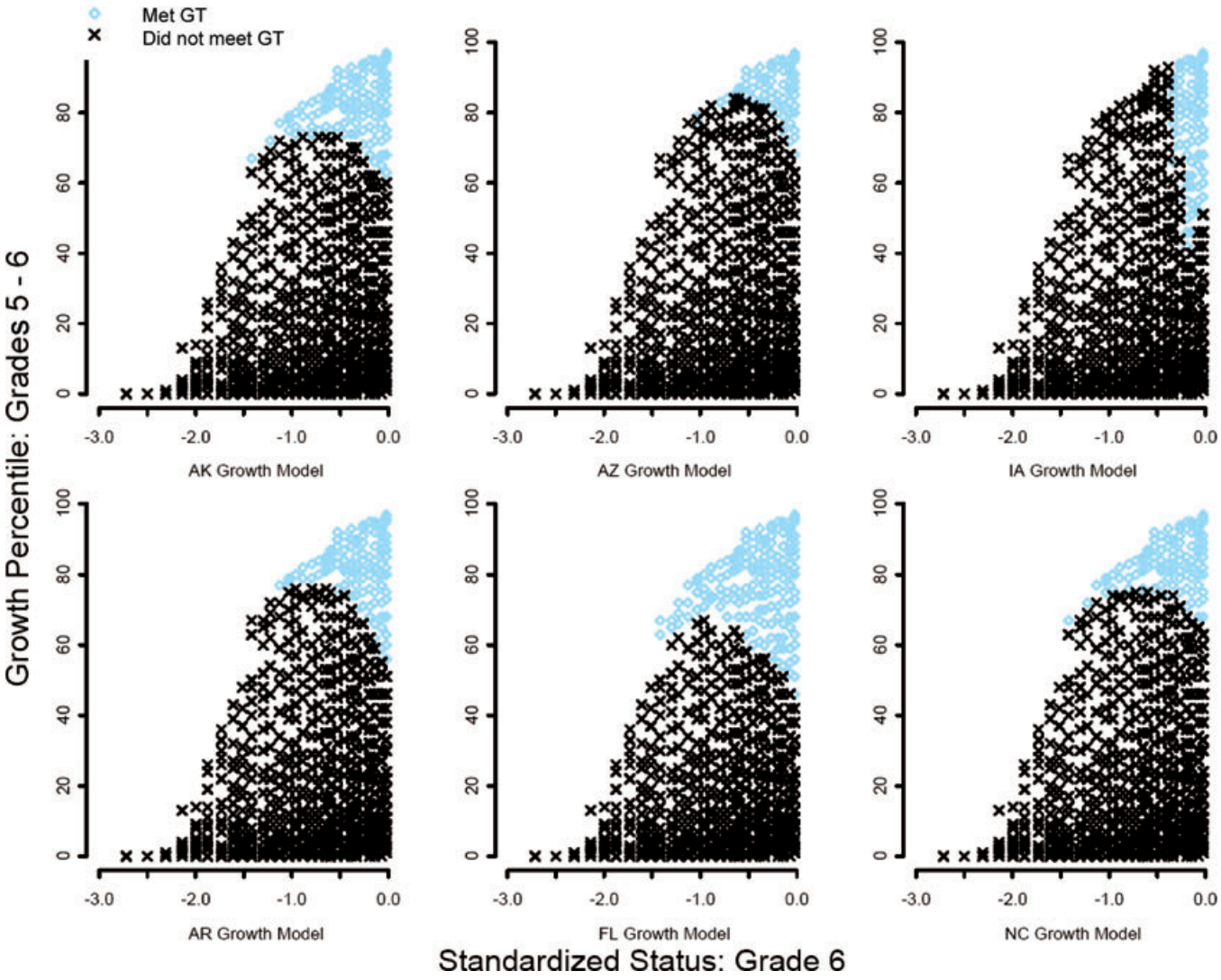

FIGURE 3. Growth to proficiency classification by status score and growth percentile, across models, for students in Grade 6.

namely requiring proficiency by Grade 8, the growth targets for the seventh-grade nonproficient students are essentially untenable. In contrast, the remaining growth models expect all students to be proficient by a later grade, making the growth targets for the nonproficient seventh graders more realistic. Although because of data availability we do not see the same phenomenon for the other growth models, it seems likely that a similar trend would arise for any state that mandated proficiency by a certain grade.

Finally, the growth percentile required to be classified as meeting a growth target varies across grades and states. While one model might have the most stringent standards for one grade, another state has the most stringent standards for another grade. North Carolina appears to have the most stringent criteria for students moving from Grade 3 to Grade 4 but the most lenient criteria for students moving from Grade 6 to Grade 7. In fact, it appears that almost all low-status students moving from Grade 6 to Grade 7 are classified as meeting their growth targets even if their growth represents a minimal growth percentile. In contrast, Florida appears to have the most stringent criteria for students moving from Grade 6 to Grade 7 but the most lenient criteria for students moving from Grade 3 to Grade 4 and Grade 4 to Grade 5. The meaning of growth to proficiency, although related to growth, varies not only across models but across grades. This result should be interpreted with caution, as it may reflect an interaction between the models and the specifics of each state assessment system. Nonetheless, it is clear that the relationship between growth and status can potentially vary across grades.

It is difficult, and somewhat dangerous, to discuss change in achievement at the lower end of the achievement continuum without addressing regression to the mean. Each of the above models is designed to measure growth toward proficiency, not growth beyond proficiency. Removing the students for whom growth is not measured from the above comparisons does not alter whether or not these models are classifying regression to the mean as growth. In theory, students at the bottom of the achievement continuum are more likely to show achievement gains irrespective of actual growth. At least two (AZ, AK) of the models have attempted to explicitly address regression to the mean through the design of the model. When the number of nonproficient students classified as meeting their growth targets is compared across initial performance level (Table 6), all but two (AZ

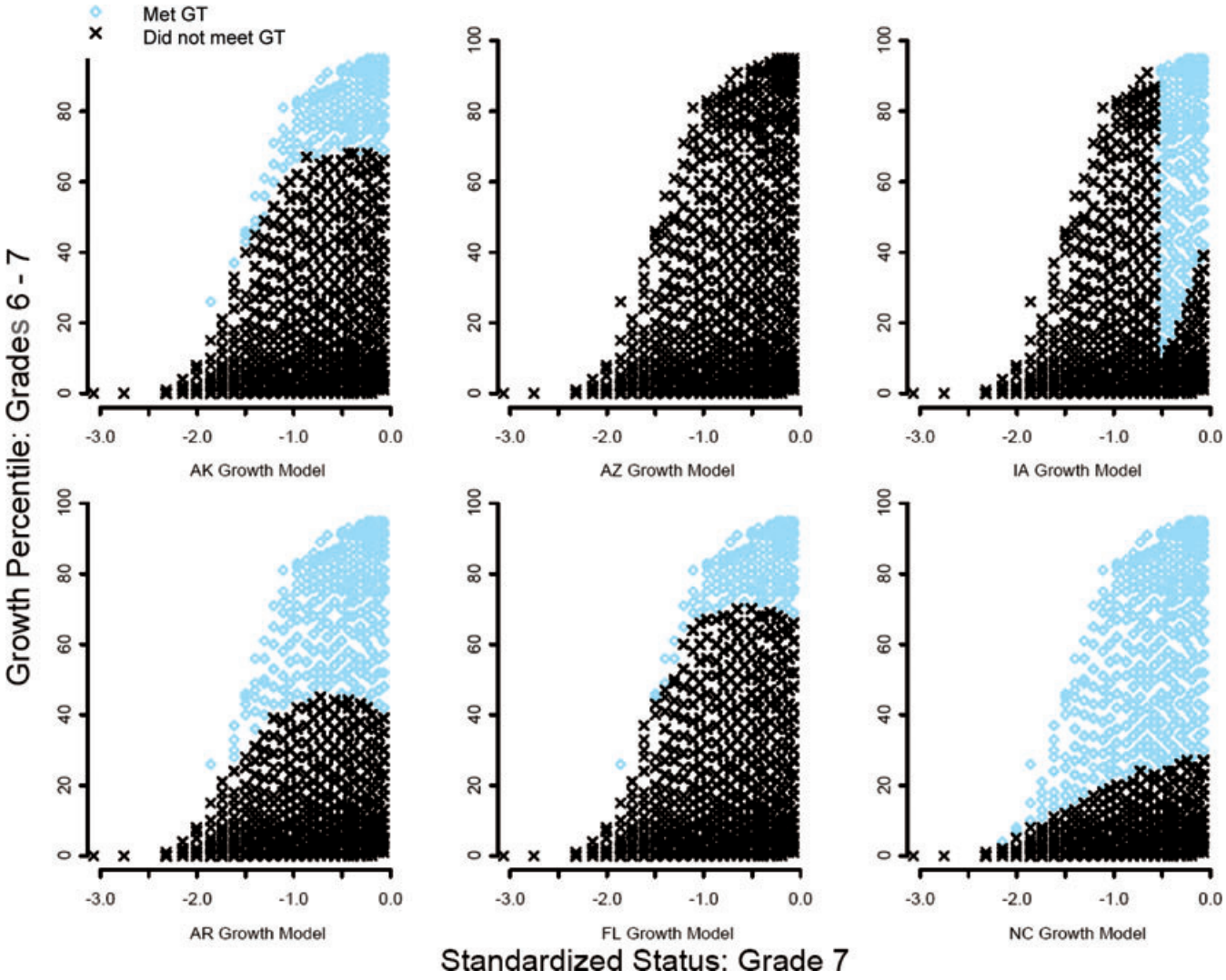

FIGURE 4. Growth to proficiency classification by status score and growth percentile, across models, for students in Grade 7.

and IA) of the models have a larger proportion of students classified as growing in the lowest performance category. Although these results do not conclusively address whether or not the models are classifying regression to the mean as growth, at least two pieces of evidence are worth noting: (1) Not all students at the lower end of the continuum are classified as meeting their growth targets (fewer than 50%), and (2) the models that have explicitly attempted to address regression to the mean do not appear to classify students with low initial status in a significantly different way from the other models.

The variation among models, although evident in the nonproficient student-level results, was nonexistent in the school-level results calculated using all students. Given that NCLB growth models are specifically designed to measure the growth of nonproficient students, it seems logical that the school accountability system should be designed to mirror this purpose. A nonproficient growth score, defined as the percent of nonproficient students meeting their growth targets, was calculated for each school. The relationships across models of the percent of nonproficient students meeting their growth targets in each school are displayed in Table 7.

The nonproficient school growth scores calculated using Arizona's, Florida's, and Arkansas's model show the strongest relationships. Interestingly, it is these three state models that rely on the vertical scale to determine whether or not a student has made sufficient growth. The nonproficient school growth scores created using Alaska's model also have a high positive relationship with these three states. While Alaska's model does not rely on the vertical scale, it does attempt to correct for the reliability of the test when calculating whether a student met his/her growth target. The nonproficient school growth scores calculated using Iowa's model have a much lower correlation with all of the growth models except Arkansas. The high correlation between Iowa and Arkansas is likely due to the similarity of school score distributions in the upper portion of the scale. More specifically, when these two models are used, there are a few schools with a high proportion of nonproficient students meeting their growth targets. Although there is not a strong relationship between the nonproficient school growth scores at the lower end of the continuum, these few schools at the upper end of the continuum are overinflating the correlation coefficient between the models.

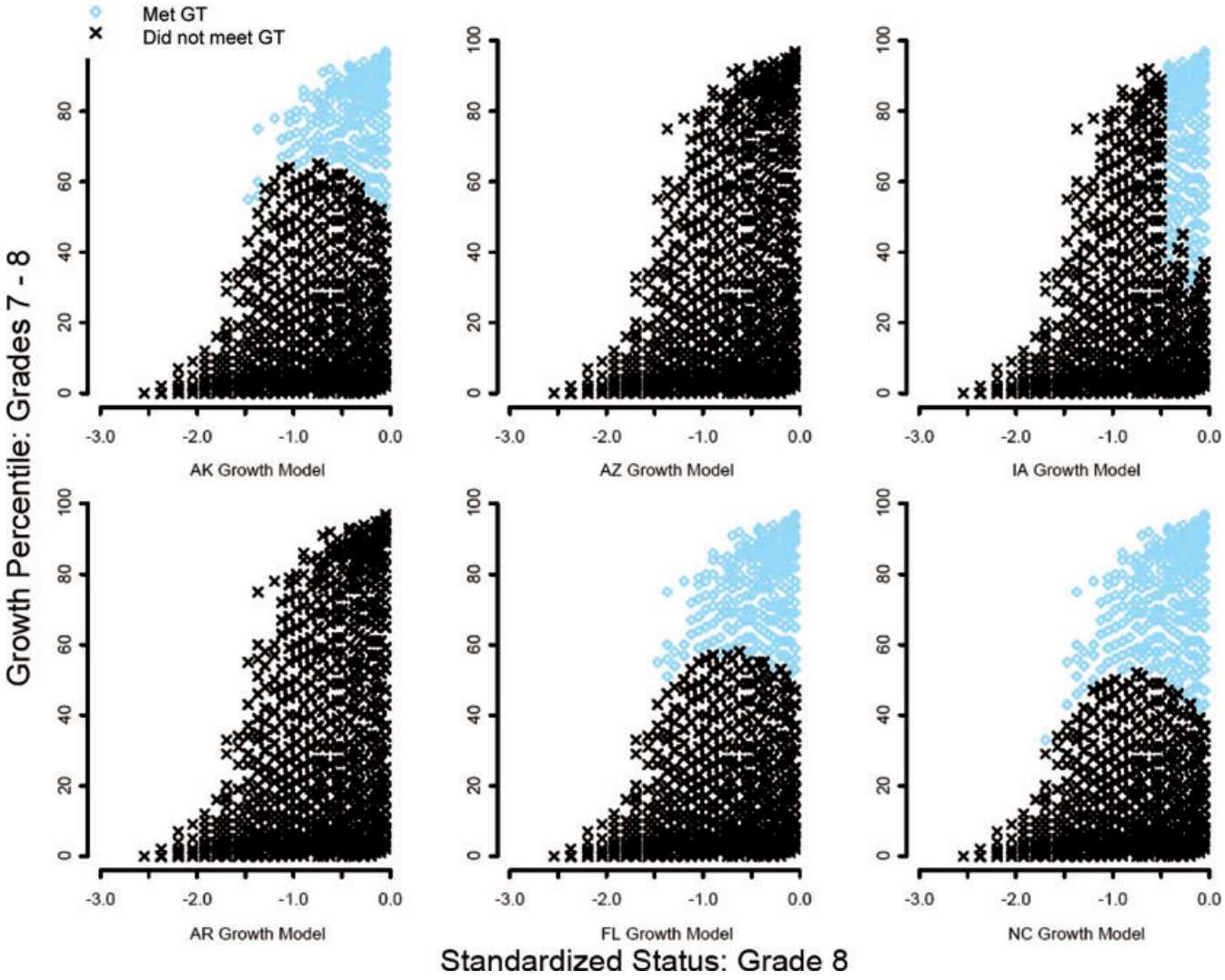

FIGURE 5. Growth to proficiency classification by status score and growth percentile, across models, for students in Grade 8.

The most surprising result of this table is that the nonproficient school growth scores calculated using North Carolina's model not only have the lowest relationship with the other models but have *negative* relationships with them. Schools with higher proportions of nonproficient students meeting their growth targets under other models have lower proportions of nonproficient students meeting their growth targets using North Carolina's model and vice versa. The model used to determine growth to proficiency clearly has an impact on school-level scores when the school-level score is based on nonproficient students.

A primary purpose of the growth model pilot program is to reward schools that are moving nonproficient students toward proficiency. According to our results, schools are accomplishing this feat. What remains to be seen is if these particular schools have large proportions of proficient students or if these schools have very few proficient students but are helping those nonproficient students close the gap. The relationship between the proportion of proficient students and the proportion of nonproficient students classified as meeting their growth targets for each school is outlined in Figure 6. Each point represents a school, the percent of proficient students in each school is represented on the horizontal axis, and the percent of nonproficient students classified as meeting their growth targets in each school is represented on the vertical axis. Each graph represents the model used to determine whether or not a student met his/her growth target. The scores generated using Alaska's model is in the upper left hand corner, Arkansas's model in the lower left corner, Arizona's model in the upper middle, Florida's in the lower middle, Iowa's in the upper right hand corner, and North Carolina's in the lower right hand corner.

Figure 6 highlights the importance of evaluating growth to proficiency independently from status. Although for the most part, the schools with higher status scores also have higher proportions of nonproficient students meeting their growth targets regardless of model, some low-status schools have a substantial number of nonproficient students who are closing the gap. Unfortunately, the proportions of nonproficient students meeting their growth targets are somewhat disappointing. The average school has between 18% and 30% of nonproficient students meeting their growth targets. Under ideal circumstances, if a school were truly closing the gap, 100% of the nonproficient students would be meeting growth targets.

Table 6. Percent of Nonproficient Students Meeting Growth Targets by Initial Proficiency Level

Growth Model	Well Below the Standard	Below the Standard
Number of students	5,884	3,165
AK	31.4	20.7
AR	28.2	17.4
AZ	11.1	13.9
FL	33.9	24.9
IA	26.5	39.8
NC	48.7	21.8

Table 7. Relationships (Cohen's kappa) Among Percent of Nonproficient Students Meeting Their Growth Targets in Each School ($N = 140$) Across Growth Models

Growth Model	AK	AR	AZ	FL	IA
AR	0.62				
AZ	0.77	0.74			
FL	0.76	0.63	0.75		
IA	0.38	0.71	0.37	0.39	
NC	−0.16	0.30	−0.30	−0.12	0.50

However, before drawing any strong conclusions, one must note that the model used to determine the proportion of nonproficient students meeting their growth targets has an impact on how each school is evaluated.

In summary, the results of this study address portions of at least four of the growth model principles set forth by USED. First, although the growth model targets are based on obtaining 100% proficiency by 2014, if these data are representative of performance across the country, the proportion of nonproficient students meeting their targets will need to increase substantially in order for 100% proficiency to become a reality (Principle 1). Second, although the growth models appear to have established reasonable yet high expectations overall, the expectations are not consistent across grades and may be too extreme in some cases. Third, including all students in a growth to proficiency model masks the growth of nonproficient students (Principle 4). If status and growth to proficiency are incorporated as separate indicators within an accountability system that applies to all students, then the status indicator could apply to all students while the growth to proficiency indicator would apply to nonproficient students. If the growth of the nonproficient students is examined in isolation from the proficient students, the results become much more informative. Fourth, the proposed models essentially track the progress of nonproficient students and the status of proficient students across years (Principle 7).

The results of this study also indicate that if applied to nonproficient students: (1) the models vary in terms of how they classify students; (2) for the most part, the models appear to have realistic expectations for nonproficient students, although the realism varies from model to model and grade to grade; (3) there appears to be a strong relationship between a student's growth percentile and whether or not a student is classified as meeting growth to proficiency requirements, indicating that the models are capturing growth to proficiency; and (4) the meaning of growth to proficiency varies from one model to another. At the school level, if growth to proficiency is calculated as a measure independent of status, but used in conjunction with status to make an overall accountability decision, then (1) the growth to proficiency models vary in terms of how schools are classified and (2) the models capture growth to proficiency. However, if growth to proficiency includes the status of proficient students then (1) the models do not vary in terms of how they classify schools, (2) the models do not capture growth to proficiency at the school level, and (3) the meaning of growth to proficiency mirrors that of status.

Conclusion

The argument for incorporating growth models into NCLB accountability systems stems from two related but distinct notions. The first argument is that a valid accountability system should hold schools accountable for the amount of learning that occurs beyond what the student knew upon entering the school. This means that schools should be held accountable for the learning of proficient and nonproficient students alike. While this is an admirable notion, it leads to the possibility of nonproficient students remaining nonproficient despite relatively large amounts of growth. The second argument is that educators and policy makers are interested in moving all students to proficient. Schools should be given credit for moving students *toward* proficiency, not simply attaining proficiency. This may be particularly useful for schools with at-risk student populations and may allow for better alignment between state and federal accountability systems.

The NCLB pilot is clearly oriented around this latter notion, which begs the question: Are the growth models developed for the NCLB pilot measuring growth toward proficiency? This study attempted to disentangle this question by examining the variation in growth model classifications at the student and school levels when the growth models are applied to the same sample of student assessment data. It is clear from the results that growth to proficiency and status represent different types of information about a school but that this distinction is masked by the inclusion of proficient students. Given that a primary goal of the NCLB growth pilot model is to reward schools that are moving nonproficient students toward proficiency in addition to those schools already meeting proficiency requirements, it seems imperative that both pieces of information are included in the accountability system.

If policy makers are willing to agree that the purpose of including growth in the accountability system is to measure growth toward proficiency, the

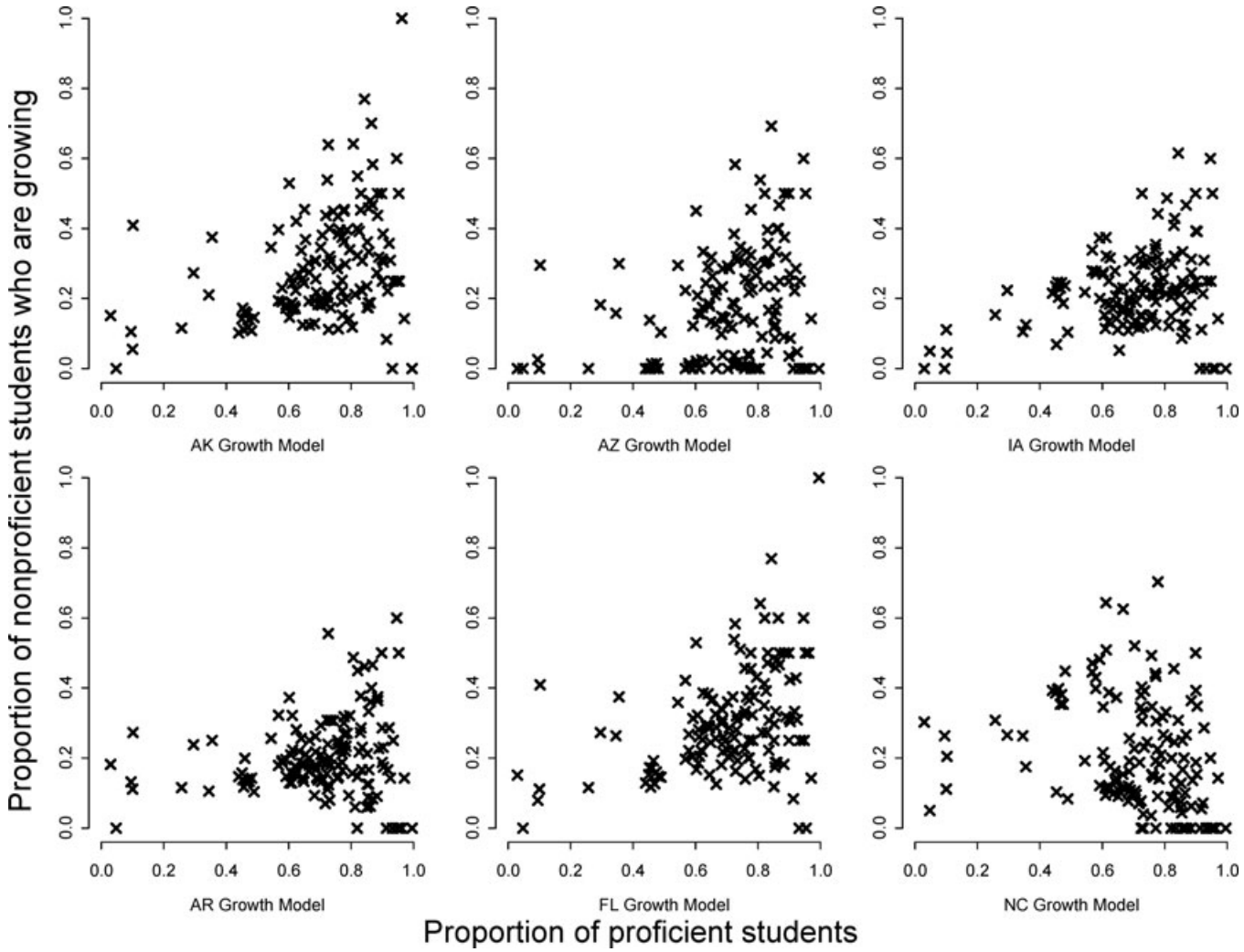

FIGURE 6. Relationship between school status (using all students) and growth to proficiency (using nonproficient students) scores across models.

incorporation of growth into the accountability system needs to be carefully considered. Under NCLB, status is indisputably a critical component of the accountability system. Once status has been accounted for, schools could be held accountable for their growth to proficiency. More specifically, a school would be deemed successful if it met or exceeded the status requirements or the growth requirements. Because the NCLB core principles indicate that all students (not just those who score below proficient) must be included in the growth model, the proposed growth pilot models were not able to make use of this system unless they included growth targets for proficient students. Since students were counted as meeting their growth targets if they scored at or above proficiency, the inclusion of the proficient students masked the growth to proficiency that they were interested in measuring.

If USED were willing to adjust the inclusion requirements for NCLB, the solution could lie in a multitier accountability model where schools are first held accountable for the status of the students and then held accountable for the growth of their nonproficient students. While the status model would include all students, the growth model would only include nonproficient students. In the first tier of the accountability system, schools would be held accountable for meeting the status requirements. This tier essentially represents the current status model in that it holds schools accountable for the proficiency of all students. The second tier would hold schools accountable for the growth of their nonproficient students. If a sufficient amount of growth toward proficiency in their nonproficient students is occurring, then the school would be classified as meeting the growth to proficiency standard. Again, a school passing either tier could be classified as meeting AYP.

Such a system is specifically designed to hold schools accountable for the proficiency of all students and the growth of nonproficient students toward proficiency. The multitier system will help ensure that schools are given credit for moving students toward proficiency, in addition to attaining proficiency. This is not to say that a multitier system is an easy, straightforward, or simple solution. The results of this study raise some important policy issues that will need to be addressed should a multitier system be adopted. First, should the growth expectations of nonproficient students be consistent from one grade to the next? Educators and policy makers exert substantial efforts in attempting to ensure that what it means to be proficient is consistent from grade to grade and that similar numbers of students are able to meet the proficiency standards in each

grade. It may seem logical to exert these same efforts to growth proficiency standards, setting the growth expectations in a way that ensures that similar proportions of nonproficient students are classified as meeting their growth targets across grades. The results of this study clearly indicate that the growth expectations of nonproficient students vary from grade to grade. However, they were set with the goal of 100% proficiency by the end of a certain grade. Is this appropriate, and should this expectation override reasonable expectations at each grade? Some important policy discussions need to take place.

A multitier system also raises some complications for the measurement community regarding the measurement of nonproficient growth in a meaningful and reliable manner. At the moment, our tests are designed to maximize the reliability of the measurements around the proficiency cut, not below the proficiency cut. While the scores below proficiency are not necessarily unreliable, they are less reliable than scores closer to the proficiency cut. In addition, although many sophisticated models have been developed to maximize the reliability of measuring growth, there are very few models specifically designed to measure growth at the lower end of the achievement continuum. Growth models are typically designed for populations or samples of students that approximate a normal distribution. Clearly the students for whom we are interested in measuring growth to proficiency do not approximate a normal distribution.

A number of issues remain unresolved when it comes to the measurement of growth toward proficiency. The multitier system presented in this article is by no means the only possibility, nor should it be. It is our hope that this article will further discussions and help generate innovative models that go beyond the current constraints of the NCLB models, yet maintain the focus on growth to proficiency. The value of future discussions will depend on the clear articulation of the following three elements and their interactions: (1) the type of growth valued, (2) the interaction between the measurement model and the underlying construct being measured, and (3) the incorporation of the measurement into the accountability system. If each of these elements is clearly articulated and the interaction between the elements is carefully considered, the potential for growth to improve the validity of the NCLB accountability system is immense.

NCLB growth models have brought us closer to examining growth toward proficiency, but the requirement that the proposed models are applied to all students has muddied the waters. By allowing states to implement a multitier system where schools are first held accountable for the status of their students and then held accountable for the growth of their nonproficient students separately from that of their proficient students, we may be able to explain the variations between the student and school classifications made by the growth pilot models and improve our understanding of what growth to proficiency means.

In today's high-stakes accountability world, we are often forced to implement new methodologies before the value of such methodologies can be fully understood. This study attempts to address that limitation by exploring how the different NCLB models conceptualize growth to proficiency. Overall, the models appear to be doing an adequate job of defining growth to proficiency and are doing so consistently across the majority of the models. However, the amount of growth required, both across models and across grades within models, varies substantially. These results serve as a starting point, useful for informing federal, state, and local policy makers about elements that should be considered when attempting to implement growth models designed within the NCLB framework. According to the results of this study, the NCLB models may be bringing us closer to rewarding schools for moving students *toward* proficiency. However, the reliability at which we are measuring growth to proficiency remains unknown.

Note

The majority of this study was conducted while the first author was an associate and the second author was an intern at the Center for Assessment.

References

Alaska Department of Education (2007a). *Alaska growth model proposal*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/ak/index.html.

Alaska Department of Education (2007b). *Peer review guidance for the NCLB growth model pilot applications: Alaska response*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/ak/index.html.

Alaska Response (2007). *Letter from education commission of Alaska to assistant secretary of United States Department of Education*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/ak/index.html.

Arizona Department of Education (2007a). *Proposal for a growth model to evaluate adequately yearly progress for schools and districts*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/ak/index.html.

Arizona Department of Education (2007b). *Addendum to Arizona's Growth Proposal: Adjustments for Regression to the Mean and error in gain scores*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/ak/index.html.

Arkansas Department of Education (2006a). *Arkansas growth model proposal*. Retrieved August 31, 2007 from http://arkedu.state.ar.us/ark_growth_mod.

Arkansas Department of Education (2006b). *Growth model amendments*. (Retrieved August 31, 2007 from http://arkedu.state.ar.us/ark_growth_mod.

Arkansas Department of Education (2006c). *Arkansas growth model proposal*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/ar/argmp.doc.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, this issue, 42–51.

Betebenner, D. W., & Shang, Y. (2007). *Reference growth charts for educational outcomes*. Paper presentation, annual meeting of the American Educational Research Association, Chicago, IL.

Carlson, D. (2006). *Focusing state educational accountability systems: 4 methods of judging quality and progress*. Retrieved November 21, 2007 from http://www.nciea.org/cgi-bin/pubspage.cgi.

Dunn, J. (2007). *The interaction of measurement, models and accountability: How values affect our growth model choices*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Delaware Department of Education (2006). *Delaware's proposal for a growth model re-submitted to U.S. Department of Education*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/de/index.html.

Florida Department of Education (2007). *Florida application for the NCLB growth model pilot program: Peer review documentation*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/fl/index.html.

Goldschmidt, P., Roschewski, P., Choi, K. C., Auty, W., Hebbler, S., & Williams, A. (2005). *Policymakers' guide to growth*

models for school accountability: How do accountability models differ? Washington, DC: CCSSO. Retrieved November 21, 2007 from http://www.ccsso.org/publications/details.cfm?PublicationID=287.

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. A. (2006). Using value tables to explicitly value growth. In R. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp. 255–290). Maple Grove, MN: JAM Press.

Iowa Department of Education (2006). *No Child Left Behind Growth Model Pilot Program*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/ia/index.html.

North Carolina Department of Education. (2006). *North Carolina's proposal to pilot the use of a growth model for AYP purposes in 2005–2006*. (Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/nc/index.html.)

Ohio Department of Education (2006). *Proposal to the United Sates Department of Education for employing a growth model for No Child Left Behind accountability purposes*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/oh/index.html.

Ohio Department of Education (2007a). *Addendum to the proposal to the United Sates Department of Education for employing a growth model for No Child Left Behind accountability purposes*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/oh/index.html.

Ohio Department of Education (2007b). *Addendum to the proposal to the United Sates Department of Education for employing a growth model for No Child Left Behind accountability purposes*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/oh/index.html.

Spellings, M. (2005). Secretary Spellings Announces Growth Model Pilot, Address Chief State School Officers' Annual Policy Forum in Richmond. *U.S. Department of Education Press Release*. Retrieved August 7, 2006 from http://www.ed.gov/news/pressreleases/2005/11/1182005.html.

Tennessee Department of Education (2006). *Proposal to the U.S. Department of Education: NCLB growth model pilot program*. Retrieved August 31, 2007 from http://www.ed.gov/admins/lead/account/growthmodel/tn/index.html.

U.S. Department of Education (2006a). *Peer review guidance for the NCLB Growth model Pilot Applications*. Retrieved November 21, 2006 from http://www.ed.gov/admins/lead/account/growthmodel/index.html.

U.S. Department of Education (2006b). *Secretary Spellings approves additional growth model pilots for 2006–2007*. Retrieved November 21, 2006 from http://www.ed.gov/news/pressreleases/2006/11/11092006a.html.

U.S. Department of Education (2005). *Key policy letters signed by the Education Secretary or Deputy Secretary*. Retrieved November 21, 2006 from http://www.ed.gov/policy/elsec/guid/secletter/051121.html.