EDUCATION**SECTOR** TECHNICAL REPORTS

March 2009

ARE WE THERE YET?

What Policymakers Can Learn From Tennessee's Growth Model

By Charles Barone

ACKNOWLEDGEMENTS

This publication was made possible by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

ABOUT THE AUTHOR

CHARLES BARONE is director of federal policy at Democrats for Education Reform. Before signing on with DFER, Barone spent five years working as an independent consultant on education policy issues. His clients included the Citizens' Commission on Civil Rights, Education Trust, and the National Academy of Sciences.

ABOUT EDUCATION SECTOR

Education Sector is an independent think tank that challenges conventional thinking in education policy. We are a nonprofit, nonpartisan organization committed to achieving measurable impact in education, both by improving existing reform initiatives and by developing new, innovative solutions to our nation's most pressing education problems.

ABOUT THIS SERIES

Education Sector Technical Reports are designed to give an informed audience further insight into an important aspect of education policymaking.

© Copyright 2009 Education Sector.

Education Sector encourages the free use, reproduction, and distribution of our ideas, perspectives, and analyses. Our Creative Commons licensing allows for the noncommercial use of all Education Sector authored or commissioned materials. We require attribution for all use. For more information and instructions on the commercial use of our materials, please visit our Web site, www.educationsector.org.

1201 Connecticut Ave., N.W., Suite 850, Washington, D.C. 20036
202.552.2840 • www.educationsector.org

Tennessee is one of 15 states participating in a pilot program to explore new ways to measure school performance under the federal No Child Left Behind Act. Under NCLB, states are held accountable for ensuring that sufficient numbers of schools' students are meeting state proficiency standards and improving schools that fail to measure up. In 2005, then-U.S. Secretary of Education Margaret Spellings launched a pilot program to study ways of measuring this "adequate yearly progress" that would reward schools for improving student achievement over the course of a school year.

Originally, NCLB judged schools mostly on whether sufficient numbers of their students met state standards each year. The law also included a "safe harbor" provision, which gives schools credit for improving student achievement even if they do not meet the specific performance targets. Some critics of the law charged that these provisions do not sufficiently account for the fact that some schools have students who are more challenging to educate than others. In response, Spellings sought to experiment with ways to measure annual student progress, known as growth models.

Tennessee was one of the first seven states that Spellings approved to implement a new school-rating system under a pilot program, which permits schools to comply with NCLB's adequate yearly progress requirement by having their students make enough progress each year to ensure that they meet a state-defined definition of proficiency within three or four years—a concept known in education circles as "growth-to-proficiency."[1] Since Tennessee was approved, their method, a "projected" or "expected" score approach, has been adopted by two other states–Ohio and, most recently, Pennsylvania. It is also under consideration by several other states and districts. Yet while growth models can help develop better school accountability systems, a close analysis of Tennessee's approach illustrates important issues that state and federal policymakers should be aware of as they consider expanding the use of growth-to-proficiency models.

The Tennessee approach is emerging as a model that other states may follow; therefore, this paper examines the advantages and disadvantages of that scheme. The analysis examines two broad themes. It considers the implications of the Tennessee growth model with regard to the amount and pace of progress students are expected to make on Tennessee's state tests, each year and over multiple years, in order for such progress to be deemed "adequate." It also discusses the broader implications for drawing conclusions about the achievement levels Tennessee students are expected to meet relative to national standards and other states as measured by the National Assessment of Educational Progress (NAEP), a national test regularly given to a sample of students in every state.

The Tennessee growth model has some distinct advantages over the current NCLB "status" model, under which schools are judged based on how cohorts of students (i.e., all students in a single grade or group of grades) compare to a similar cohort of students from the previous year. As such, school ratings under NCLB are currently based on changes in the performance of different students from year to year, rather than growth in the performance of the same students over time. At the broadest level, the Tennessee growth model may more fairly credit those schools and districts that have made significant progress with low-achieving students that would not be reflected in the percentage of students who have met or exceeded the proficiency benchmark. This, in turn, may have instructional benefits for individual students as schools focus additional attention on a broader group of students and track the progress of students over a multi-year period.

The Tennessee growth model system, however, has some significant potential downsides:

- By setting an interim goal short of proficiency, in a state judged by the U.S. Department of Education to have among the lowest standards of any state, it may be setting the bar so low as to evoke fairly small gains in student achievement.
- While the "expected score" system estimates a student's path to proficiency in three years, in fact, many students will not make it to proficient in three years or ever because of a statistical phenomenon known as Zeno's paradox.
- Finally, because this model relies on multiple regression analysis, one must be a statistician to understand it. Although complexity may be a necessary trade-off for more accuracy, there is a loss of simplicity and transparency for parents and the general public.

These downsides could have a significant negative impact on student learning in Tennessee and show some of the risks associated with growth models if they are not implemented with a keen eye toward ensuring results for the lowest-achieving students. Slowing down the pace at which students are expected to learn academic skills in elementary school may create long-term problems for students and create larger and more difficult burdens for public education in junior high, high school, and beyond. A wide body of research suggests, for example, that children who do not acquire language skills in the early grades have an increasingly difficult time catching up to their peers as they progress.[2] This parallels neuropsychological research that shows critical periods for brain development in language and other areas of cognitive functioning.[3]

The Tennessee growth model will also reduce the number of schools identified by NCLB as falling short academically. This could be a positive change if it allows the state to focus more intensely on the lowest-performing schools. However, it will also mean that some schools may not be able to avail themselves of resources that could help address student learning problems early enough to prevent future academic failure. These and many other trade-offs are inherent in accountability systems, and for the foreseeable future each state will have to decide for itself how to manage competing considerations within federal rules.

The growth models currently being used also vary greatly in their methods and take on unique characteristics based on the types of standards and assessment systems that their respective states employ. These specifics matter a great deal in judging what making "adequate yearly progress" or AYP means in any particular state using a growth model.

The purpose of this paper is not to recommend for or against the Tennessee growth model, an approach that reasonable people can disagree over. Rather, because the NCLB accountability system is essentially federalist in that it allows states great leeway in determining their standards, assessments, benchmarks, and definitions of adequate progress, it is important to take a state-specific focus to help consumers of these systems—teachers, parents, administrators, elected officials, and advocacy groups—explain their implications for measuring student achievement and to inform future revisions of both the NCLB law and state plans under the act. The intense and organized resistance to NCLB and the ensuing national debate about the law have to some extent obscured the real trade-offs between different approaches to school accountability. This paper is intended to help those involved in this process make more thoughtful and informed decisions by examining the approach of this important state.

Background

The measure of "adequate yearly progress" or AYP as outlined in the NCLB statute has at least two key limitations.

First, the statute delineates a "separate cohort" model of measurement in which, for example, the performance of this year's third-graders is compared to that of last year's third-graders. This method compares two completely different groups of students and does not take into account where each student started at the beginning of third grade in assessing their performance at the end of the school year. Thus, a great deal of error or "noise" is introduced into the analysis that has nothing to do with whether a *particular* student or classroom of students is making academic progress.

A second limitation is that the outcome variable, or whether or not a student reaches proficiency, is dichotomous (either-or), while the underlying measure, or test score, is continuous (can take on many different values). This is a highly inefficient method of statistical analysis. A great deal of data is lost, and it is quite

possible that a school could significantly boost the achievement of many or even most students but still not meet AYP because not enough students reached or exceeded the proficiency benchmark. Some research indicates that the proficiency benchmark creates a perverse incentive for schools to focus only on those students close to proficiency (so-called "bubble kids") and ignore those, at least in the short-term, who are farther from the proficiency goal.[4] This phenomenon has been referred to by some observers as "educational triage."[5]

A better model would be longitudinal—measuring each student's academic progress over time—and continuous—looking at how students progress based on their actual score, rather than whether or not they attain a score that puts them in a particular (and somewhat arbitrary) category, such as "proficient." This type of model is known in the NCLB arena as a "growth model." By setting an arguably more attainable goal for each student, each year, the Tennessee growth model may provide stronger incentives for schools to focus their instructional efforts and supports on boosting the achievement of *all* students below or near the proficiency cut point *every* year.[6]

The use of the separate cohort model and the proficiency benchmark in the NCLB statute was partly political—it is easy for the public to understand and sends a clearer message about the level of academic achievement that students are expected to attain. It also was partly technical—at the time NCLB was passed, very few states had data systems that allowed for longitudinal analysis of individual students. Since NCLB's passage, however, the limitations of the separate cohort/proficiency model have become more widely understood, and many states have developed and implemented data systems that permit longitudinal analysis of student achievement data.

This led Secretary Spellings to allow states to submit plans for using a growth model of AYP that is outside the exact letter (though, in its general concept at least, fits the overall purposes and aims) of the law. Currently, 15 states have been approved to use growth models for measuring AYP. The use of growth models likely will be expanded even further when Congress and President Barack Obama take up the reauthorization of NCLB, which is expected to occur in 2009–10.

As such, it is a good time to take a close look at the growth models currently being used by states, with the hope of better understanding some of the issues, advantages, and limitations associated with them.

The Tennessee Growth Model

On May 17, 2006, the U.S. Department of Education approved a proposal by the state of Tennessee to use a growth model to determine whether schools and local education agencies (LEAs) are making adequate yearly progress under NCLB. This method for determining AYP differs substantially from those prescribed under the statute.

Tennessee's growth model uses "projections" of student growth to determine whether or not students are on the path to proficiency within three years for any student scoring lower than proficiency. The model measures student progress against benchmarks based on predicted or "expected" scores that a student would need each year in order to attain proficiency within three years (i.e., from the time the prediction is made or the baseline year, to the target year, or year 3). For example, in 2008, the progress that a fourth-grader makes from his or her third-grade score will be used to determine if their rate of growth would predict that the student on average would reach proficiency when he or she reaches seventh grade in 2011. These predictions of future performance are made using inferential statistics (multiple regression) based on prior student data. Table 1 shows how determinations are made in Tennessee for each category (grade level) of students.

Table 1. Predicting Future Performance

Student Category	TCAP* Score Applied	Proficiency Standard
Third grade	Third grade	Third grade
Fourth grade	Seventh-grade projection	Seventh grade
Fifth grade	Eighth-grade projection	Eighth grade
Sixth–Eighth grade	High school projection	High school
With no prior test score	Current score	Current grade
Who take alternative assessments	Current score	Alternative standard

*TCAP – Tennessee Comprehensive Assessment Program.

Source: Tennessee Department of Education, "Proposal to the U.S. Department of Education, NCLB Growth Model Pilot Program," May 15, 2006, available online at http://www.ed.gov/admins/lead/account/growthmodel/tn/index.html

The projection model is a fundamental departure from both the status and safe harbor models of AYP that were laid out in the NCLB statute and explicated in regulations promulgated by the U.S. Department of Education. (Safe harbor provides that schools make AYP if they reduce the percent of students who are not proficient by 10 percent from the previous year, even if the percent proficient falls short of the "status" goals.) In and of itself, this departure does not make the Tennessee model invalid. But the results of the Tennessee model will differ substantially from those of the status and safe harbor models. Under the Tennessee plan, schools may meet AYP through any of the three accountability models. Schools will be more likely to make AYP under the projection model than the other two models. In fact, Tennessee estimates that 13 percent of schools that would have otherwise failed AYP will make it under the projection model.

Both the annual measurable objectives (AMOs) and safe harbor targets that schools must meet to make AYP are based on whether or not a student actually achieves proficiency in the year in which he or she is tested. In turn, whether or not a school or LEA makes AYP is based on the exact percentage of students who attain proficiency.

The projection model, however, is based on whether a student attains an *interim* score that is *estimated* to put them on a path to proficiency within three years. The interim score is recalculated each year. When Mary Smith completes third grade, her "target score" for fourth grade—one that satisfies AYP—is one that puts her on the path to proficiency within three years, or by the end of seventh grade.[7] To determine the target score that would put her on track to proficiency, the state looks at students in the past that scored similar to Mary in third grade and what minimum score in fourth grade would put her on the path to being proficient on average by the time she reaches seventh grade. It is important to recognize that this fourth-grade target is not a guarantee of her making it to proficient, but instead a score that would lead to proficiency on average (half of the time). The fact that the average student with similar scores to Mary made it to the proficiency cut point means that half of the students were above the cut point and half were below. So, reaching the cut point in fourth grade to be projected to be proficient in seventh grade will not guarantee it happens. When Mary Smith completes fourth grade, her "target score" for fifth grade is one that puts her on the path to proficiency by eighth grade based on students with similar third-grade and fourth-grade scores.

One key to understanding what distinguishes the Tennessee growth model from the status and safe harbor models is to pay close attention to the level of analysis, specifically, what is deemed adequate for an individual vs. a school or school district.

Under the status and safe harbor systems, an individual student's progress is deemed "adequate" if he or she meets a proficient level of performance as defined by a benchmark score on a state test. A school or district makes adequate yearly progress if they meet either the AMO (status model) or reduce by 10 percent the percentage of students not proficient (safe harbor).

Under the growth model, the adequate score for an *individual* is the expected score generated by the multiple regression, a score which puts them on a path to the cut off score of proficient in 3 years. It should be a score roughly one-third the distance between the student's prior year score and "proficient." So, in any one year, "proficient" is only primarily important as a variable in a mathematical sense, not as a definitive target for where a student is expected to be that year.

School and *district* AYP is now defined by the percentage of students who meet the lower bar score of "on the way" rather than proficiency per se. That percentage is still determined by the state AMO, which "ratchets up" over time. In effect, "on the way" is now equivalent to "proficient" for the purposes of AYP, even though it is a much lower bar.

How a School or District Can Make AYP Under a Growth Model

Status Model – Determines the percentage of students in the school that are proficient in math and language arts and compares it to a statewide average measurable objective (AMO) or target percent proficient. For example for Tennessee in 2008–09, 86 percent of students in math and 89 percent of students in language arts must meet proficient for the school to make AYP under the status measure.

Safe Harbor Model – A school must reduce the number of students that are not proficient in a subject by at least 10 percent to meet AYP in that subject through safe harbor.

Tennessee Projection Model – The percentage of students that are either (1) proficient in the subject or (2) projected to be on track to proficiency over three years must exceed the same AMOs used in the status measure—86 percent in math and 89 percent in language arts for 2008–09.

One potential advantage of this model is that it may focus the state, districts, and schools on the short-term goal of improving the achievement of all students. For example, under the status and growth models, if Mary Smith's score were far below the proficiency benchmark, there may be less motivation to focus on her, at least in the short-term, if it were determined that she was too far below proficient to boost her score to the proficiency benchmark in just one year. In contrast, under the Tennessee growth model, the school or district would get "credit" (i.e., be deemed making AYP) with regard to Mary Smith, if she were "on her way" to proficiency and would be estimated to reach it by seventh grade.

The disadvantage is that Mary's target will always be "on the way" to proficiency because, under Tennessee's model, Mary's goals are recalculated each year. Her target fifth-grade score is one that is estimated to put her on the path to proficiency by eighth grade. Her sixth- through eighth-grade scores are recalculated each year based on a score projected to be on the path to a goal of proficiency by high school.

In short, proficiency is a "moving target," always three or more years away, in perpetuity. There is nothing to ensure that, over the long run, Mary moves ahead toward a cumulatively higher level of performance over successive years. So while the model may provide an incentive to focus on all students in any given year in order to help them make small increments, there is no longer-range plan for getting all students to a minimum level that would be the equivalent of "proficient" or "grade level" or even "basic skills."

The model also has implications for schools in determining how AYP is assessed, which will in turn determine where interventions and resources are directed and where options are given to students for public school choice and after-school tutoring (i.e., supplemental educational services).

As long as each student makes a small amount of progress toward proficiency, the school could hypothetically make AYP through the growth model even though not a single student had actually gained proficiency. This is not only true in any given year, but also could be true for a period of more than the three years implied.

There are some checks on this happening. For example, third-grade AYP is assessed through a separate cohort model, which can only be measured against the state AMOs. If third-grade reading stayed stagnant for several years, the school would not make AYP. But the other grades could make AYP, complicating interpretations by all but the most sophisticated analysts.

For instance, things become more complex and harder to interpret when we project out from fourth grade. A student's fourth-grade target is based on third-grade scores but is a projection to seventh grade. But, a student's fifth-grade target is based on both the actual third- and fourth-grade scores and the projection to eighth grade. There are a lot of variables in this model, and because a complicated regression equation is used to estimate "projected scores," what an "adequate" trajectory means operationally over the course of a student's years in elementary or middle school is almost impossible to describe succinctly and without proprietary access to state data. It may be a while, however, before the AYP "red flag" goes up, and by the by time it does, a student may have moved on to middle school or high school.

Common sense suggests that there is no real need to obsess over the above permutations, since if a student is required to get roughly one-third of the way to proficiency each year to meet AYP, over the course of three years he or she would be at or above proficiency.

But in this case, common sense is wrong. In fact, if a child closed one-third of the gap to proficiency each year, they would, in strict mathematical terms, *never* get there, though at some point (likely later than three years) they would certainly get "close enough." Again, the core issue is that the distance to the goal is recalculated each year, based on the current year's progress.

This statistical dilemma, widely known among mathematicians, is often referred to as Zeno's paradox and was elaborated by Aristotle in such examples as the race between Achilles and the tortoise.[8] It is used in calculus to illustrate the concept of a "limit." Here, we use fractions and the example of a frog trying to reach a lily pad to explain this paradox.

Let's say a frog very much wants to get to a lily pad that is 100 meters away. He knows he can't make it all the way, but judges he can make it a third of the way (he's a championship frog). So on his first try, he jumps 33 1/3rd meters. He decides to jump a third of the way to his goal from where he stands each time, until he gets there.

When does he get there? On first glance, most of us would say in three jumps. In fact, in strict mathematical terms, he *never* gets there. After three tries, he is actually only 19/27ths (71 percent) of the way there. Table 2 shows the calculation.

When we say that the frog (or student) who meets the interim goal will never meet, in a strict mathematical sense, the end goal, we do not mean that at some point they will not get "close enough" to be considered "as if." Where that point is, is a "judgment call" that, right now, only the state of Tennessee is in a position to make. Such a decision should weigh-in what 71 percent or some other fraction of the way to the goal means relative to some widely recognized benchmark. Essentially, students deemed as "being on the way" to proficient under the model are, as shown in Table 2, somewhat less so, and that the "not making AYP" flag may not be raised at all even though a student or school is falling far short of what everyone thinks is a "100 percent proficiency" goal within three years.

It should be noted that test scores can take on far fewer values than distance, so the comparison is not quite exact. For example, for the Tennessee language arts test, 25 questions right out of 55 will result in a student being proficient, and 14 questions right would be random guessing. So students below proficient will likely be in the 14 to 24 questions correct range (only 11 values). At some point, if a student keeps progressing, he or she will likely reach the "tipping point" that separates a proficient score from lower performance levels. But that point can be beyond the three-year timeline implied in the Tennessee model. How far beyond depends on the number of test items, the cut scores used for particular tests and grades, and scoring algorithms.

The Tennessee growth model also means that schools in Tennessee will be able to make AYP in 2014 without all students being proficient. Universal proficiency by 2014 is a core principle of NCLB. There is no way to determine when, exactly, Tennessee would have to meet the universal proficiency goal, because it is difficult to model exactly what will happen with individual schools over time under the projection model without having access to the data. It is clear that a school will likely have to increase the percentage of students that are proficient over time, but how fast is an empirical question. Figure 1 shows the AMO requirements for language arts and math over time under the status model. Recall that under the projection model the percentage of students that either meet proficiency or are on track to proficiency must meet the status AMOs.

But even in 2014, a large proportion of students could be on track to proficient for many years past 2014 without getting to proficient. If a school made its AMO target every year, then eventually a school would get to 100 percent proficient but it would take some time. (This would happen because under the projection model third-graders and new arrivals would need to meet proficient, because no growth option exists for them, and if 100 percent of third-graders were proficient in 2014, then in 2015, 100 percent of fourth-graders would have to be proficient to meet the schools AMO and so on.) But if schools don't make AYP every year, then schools could continue even longer with students not making it to proficiency.

This may or may not be acceptable under former Secretary Spellings' interpretation of federal law or to Tennessee state policymakers. But it does violate the statute as written, as well as Spellings' own "first core principle" for judging growth models (i.e., that they lead to 100 percent proficiency for assessed students by 2014).[9]

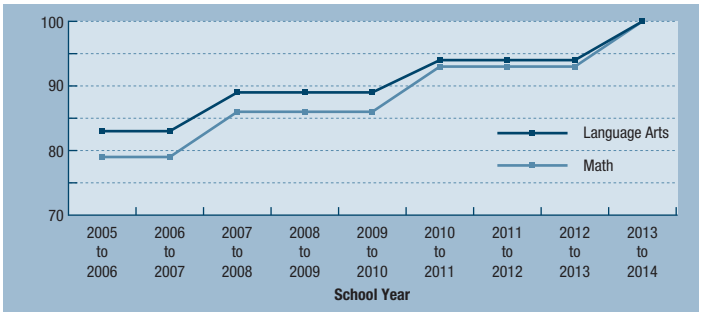
Table 2. Zeno's Paradox on the Way to Proficiency

Jump Number	Distance of the Jump = 1/3 of distance remaining	Total Distance
1	1/3rd	1/3rd
2	1/3rd of 2/3rd = 2/9	1/3 + 2/9 = 5/9
3	1/3rd of 4/9th = 4/27	5/9 + 4/27 = 19/27

Growth Model vs. Safe Harbor

Some of the same criticisms outlined above regarding growth models, such as the lack of any requirement that student achievement must progress upward over time, could also be leveled against the safe harbor provision. Like the growth model, the safe harbor provision could deem schools as making AYP in any one year, but over time the school or every individual student would never get to the AMO or the proficiency benchmark, respectively. And the safe harbor provisions extend NCLB's timeline past 2014. These are issues that Congress or the Obama administration may want to address in reauthorization.

Figure 1. Tennessee Percent Proficient to Make Status AMO



Source: Tennessee Department of Education, "Proposal to the U.S. Department of Education, NCLB Growth Model Pilot Program," May 15, 2006, available online at http://www.ed.gov/admins/lead/account/growthmodel/tn/index.html

But relative to the Tennessee growth model, the safe harbor provision has three advantages. First, it is based on real data, rather than a projection. Second, it is based on students achieving a set, policy-driven target—proficient—rather than a moving, amorphous, and norm-referenced target (i.e., a projected score), which has many more variables. Third, it is easily understandable to parents, policymakers, and the general public.

There are other differences. There is error built into all systems of student testing. Both the AMO and safe harbor models have "measurement error" associated with the difference between a student's "true achievement" and their performance on one or more tests. But the use of "projected scores" by Tennessee introduces an additional error factor. The Tennessee growth model substitutes a predicted or expected score for the AMO. Tennessee shows that the correlation of the predicted scores with actual scores is about .80 (R=.80). This means the percentage of variance accounted for is only about two-thirds (.8 squared = 64 percent); thus, one-third of the variance in actual scores is not accounted for by the predictive model. While an R of .80 is quite respectable in the research world, it may not be adequate for making real-world decisions. Many students will be counted as being on track when they are not, and vice versa.

Tennessee's own 2006 application purports that 13 percent of schools that had not otherwise made AYP under the status (AMO) or safe harbor provisions would make AYP under this model. What proportion of these schools is "misidentified" (i.e., based on accuracies in the projection of students' scores), is unknown. The report promised by Tennessee evaluating its growth model has not yet been issued, but when it is, the concerns raised above can help inform a serious and objective evaluation of what the results mean. [10]

In addition to the methodological concerns raised above, and also *because* of the complicated nature of them, the use of projected student scores under the Tennessee growth model makes the system less transparent (one has to dig through data that is not readily available to the public) and much more difficult to decipher. In order to determine a student's expected score, unlike the status or safe harbor models, one must have access to (secured) individual student data and proprietary information about the state's methodology and be an advanced statistician.

In her guidance on growth models, Secretary Spellings asked the state and reviewers to address whether the proposal has "categories for understanding student achievement at the school level and reports for growth performance and AYP judgments that are clear and understandable to the public?"[11] The Tennessee model seems, arguably at least, to violate this principle. It may be, in fact it is likely, that when a school or student in Tennessee is deemed as making AYP, the findings will have a much different meaning to the public than to the handful of statisticians that understand how the system works. What happens when a parent or a reporter asks why a student who is at "basic" or even "below basic" is nonetheless judged as being proficient for the purposes of AYP and is told that the student, and/or his or her school, met a standard set by an abstract mathematical formula using matrix algebra that puts them on a "path to proficiency" within three years (and, as we explain here, actually further) in the future. What will their reaction be when they actually see the formula?[12]

As school accountability models become more sophisticated, policymakers will face a trade-off between the values of transparency and simplicity and the attractiveness of more fine-grained measurement systems. Tennessee illustrates this dynamic.

Growth Model and State's Performance Benchmarks

Tennessee has one of the least stringent set of standards for defining what is "proficient" as measured against the National Assessment of Educational Progress (NAEP),

ranking at or near the bottom relative to other states, depending on subject and grade level. Tennessee's definition of proficient, like most states, is below the standard of NAEP proficiency. Tennessee, however, like many (or most, depending again on subject and grade level) states, has a proficiency standard that is below even the NAEP standard of *basic*.[13] In fourth-grade reading, for example, the NAEP benchmark for "basic" in reading is a score of 243; for "proficient" it is 281. The NAEP-equivalent score of the Tennessee standard for fourth-grade proficiency in reading is 222, which is about as far below the NAEP standard of basic as the NAEP standard for basic is below the NAEP standard of proficient. The NAEP benchmark for basic in fourth-grade math is 214; for proficient, it is 249. The NAEP equivalent of Tennessee's standard for proficient in math is 200.

Since the Tennessee growth model sets a yearly goal of only incremental progress toward proficient and resets its projection toward proficiency each year requiring, as described above, less cumulative progress over time than is implied, it will have the effect of further watering down expectations for achievement that are already very low relative to other states.

Because of the federal-state relationship on education policy, NCLB is set up to allow states to develop their own content standards, select their own tests, and set their own benchmarks for what constitutes proficiency. Some have argued that states should be given more leeway in developing their timelines toward proficiency, and that the strict timelines under current law encourage states to lower their standards for what constitutes proficiency. In the case of Tennessee, we see a state with already low standards that has sought to ease the timeline even further through the use of a growth model.

While it may not be possible or necessary under current law to compel or even prod Tennessee to raise its standards, it does seem reasonable and legally permissible that in evaluating state growth model proposals, where the yearly goal is an interim one, or something short of proficient, that the secretary of education take into account what that means for student growth relative to some absolute standard like the NAEP in deciding whether to approve the plan or ask for revisions. In other words, against the context of different states, all growth models may not be equal even if they meet the technical requirements of the regulations. Congress may also want to take this into account when it takes up all the issues associated with standards, assessments, AYP, and growth in the upcoming NCLB reauthorization.

Conclusion

The use of growth models represents an opportunity to improve upon the state accountability systems currently in use under NCLB. NCLB's focus on a single criterion, proficiency, and its lack of focus on continuous academic progress short of proficiency, fails to recognize schools that may be making significant gains in student achievement and may encourage so-called "educational triage."

However, the growth models currently being used by 15 states under the federal pilot program vary greatly in their specific characteristics. This paper attempts to illustrate how those specifics matter by using the case of one of those states, Tennessee, which was among the first approved to use a growth model in 2006.

The model does offer some advantages. By setting goals short of, but on a statistically projected path to, proficiency, the model may provide an incentive to focus efforts—at least in the short-term—on a wider range of students, including both those close to and those farther away from the proficiency benchmark. It also may more fairly credit, again in the short-term, those schools and districts that are making significant progress that would not be reflected in the percentage of students who have met or exceeded the proficiency benchmark.

However, because the projection toward proficiency is recalculated every year, there is not necessarily a significant progression, over time, of students toward proficiency. Students are, in fact, always expected to be "on the way" to proficiency, without being expected to ever really get there. And, even if a student meets his or her projected score of proficiency three years in a row, he or she will not necessarily, as implied by the model, reach proficiency within three years (Zeno's paradox). This may delay the focus of intensive instructional approaches and the direction of resources beyond the point in a child's development of academic skills at which they would be most effective. Those states that, thus far, have adopted Tennessee's approach—Ohio and Pennsylvania—have also inherited this potential disadvantage to their models.

Furthermore, because Tennessee's standard for proficiency is so low, both in absolute terms compared to the NAEP, as well as relative to other states, interim progress part way toward that goal may represent a fairly low attainment of basic skills. Finally, the interpretation of all of these factors is made somewhat more difficult by the reduced transparency associated with projected scores.

These factors should be considered more carefully when the secretary of education reviews the next set of growth models over the coming months, and Congress should study them closely and thoroughly as it considers revising accountability guidelines as part of NCLB/ESEA reauthorization.

Endnotes

[1] A list of state applications, decision letters, and other information related to the growth model pilot projects can be found at http://www.ed.gov/admins/lead/account/growthmodel/index.html

[2] R. Lyon, "Reading Disabilities: Why Do Some Children Have Difficulty Learning to Read?" *The International Dyslexia Association's Quarterly Periodical* 29, no. 2 (Spring 2003); K. Rayner, B.R. Foorman, C.A. Perfetti, D. Pesetsky, and M.S. Seidenberg, "How Should Reading Be Taught?" *Scientific American* 286, no. 3 (March 2002): 85–91.

[3] H.T. Chugani, "A Critical Period of Brain Development: Studies of Cerebral Glucose Utilization with PET," *Preventive Medicine* 27 (1998): 184–188.

[4] R. Reback, "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," Working Papers 0602, Barnard College, Department of Economics (2006); J. Booher-Jennings, "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal* 42 (2005): 231–268.

[5] The other major study on the "bubble kids" issue found no evidence of perverse incentives or educational triage. See M. Springer, "Accountability Incentives: Do Schools Practice 'Educational Triage'?" *Education Next* 8, no. 1 (Winter 2008), available online at http://www.hoover.org/publications/ednext/10895041.html

[6] No one has as yet offered a clear accountability solution to address the criticism that any NCLB accountability model (i.e., one targeted only at "proficient") does not provide any incentives to raise performance once students have met the proficient benchmark. This is likely due to the fact the federal government has traditionally focused its efforts on the most at-risk students. Hopefully, states and localities are addressing this issue through other means. But this may be something that will be dealt with in the next ESEA reauthorization.

[7] In this sense, Tennessee is arguably a four-year projection model, because a student's path to proficiency in seventh grade is based on third-grade scores.

[8] http://en.wikipedia.org/wiki/Dichotomy_paradox

[9] Peer Review Guidance, January 25, 2006.

[10] The U.S. Department of Education Web site indicates that at least one such report has been issued, but the report or reports do not appear there. The department did not respond to a request made by the author to obtain these reports.

[11] Peer Review Guidance, January 25, 2006; p. 6, 1.3.2.C.

[12] Projected_Score = MY + b1(X1 – M1) + b2(X2 – M2) + ... = MY + xiT b where MY, M1, etc. are estimated mean scores for the response variable (Y) and the predictor variables (Xs).

[13] *Mapping 2005 State Proficiency Standards onto the NAEP Scales* (Washington, DC: National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education, June 2007), available online at http://nces.ed.gov/nationsreportcard/pdf/studies/2007482.pdf

APPENDIX 1

RESPONSE TO *ARE WE THERE YET? WHAT POLICYMAKERS CAN LEARN FROM TENNESSEE'S GROWTH MODEL*

By William L. Sanders
SAS EVAAS®
SAS® Institute, Inc.

Several years ago, I was encouraged by various educational policy leaders to initiate studies on how the AYP part of NCLB could be augmented so that educators working within schools with entering populations of very low achieving students could be given credit for excellent progress with their students and avoid being branded as a failing school. From the Tennessee longitudinal database, it did not take long to find "poster child" schools. One Memphis City school at that time had 3rd graders with a mean achievement level that would map to the 26th percentile of all Tennessee 3rd graders. Yet this school had this same cohort of students leaving 5th grade at approximately the 50th percentile relative to the state distribution. Still this school was failing AYP because the 3rd, 4th and 5th grade scores had to be composited. Relative to NCLB, this was a failing school. In our view, this was not a failing school; rather it was an outstanding school and should not suffer the indignity of the failing school label.

Additionally, in our early investigations, we found schools that had passed AYP, yet had many of their students who had received the proficiency designation to have trajectories that would lead to a non-proficiency status in the future. It was our intent to develop a process that would give positive recognition to those schools that were truly ramping up their students' achievement levels, while not giving credit for those schools whose all ready proficient students were being allowed to slide. In other words, we endeavored to create a methodology to give the current schools credit for changing academic trajectories so that their students would have the opportunity to meet various academic standards in the future if effective schooling was sustained into the future. We sought methods consistent with the intent of NCLB to rectify the mislabeling of very effective schools as failing. We were encouraged to develop a process to be an augmentation of the AYP process, not a replacement for the USDE approved AYP process.

At the time we initiated these studies, we had many years of experience working with longitudinally merged student test data and knew of many difficult problems that would have to be addressed in order to achieve a process that would have fairness and reliability. We considered and rejected several approaches. One of the first to be rejected is one of the simplest. For example, one such approach would be to take a student's current test score, subtract that score from the proficiency cut score three years in the future, divide the difference by three to yield how much progress that student must make per year. If, in an intervening year, the student's score exceeds the target score, then that student would be deemed to have made appropriate growth. Then the percentage of students making growth could be calculated for each schooling entity (i.e. district, school or subgroup) and the approved AYP rules could be applied as for the AYP status procedure.

We rejected this approach because (1) the error of measurement in any one score for any one student is so relatively large that the setting of an improvement target by this method will inevitably send the wrong signals for many students, and (2) by definition vertically scaled tests provide scores that are intrinsically nonlinear over grades, resulting in uneven expectations of growth at the student level. We believed there to be a process that avoided these two major problems and would result in greater reliability for the final estimation of whether or not a school had earned the right to be considered a non-failing school even though it had not met the regular AYP requirements.

One of our first objectives was to avoid making judgment about the progress of an individual student based upon

APPENDIX 1 (continued)

one test score—like what is done with simple approachs similar to the one outlined above. To minimize the error of measurement problem associated with one test score, we elected to use the entire observational vector of all prior scores for each student. In some of our earlier work, we had found that if at least three prior scores are used, then the error of measurement problem is dampened to be no longer of concern.[1] *This number independently of us was also found by researchers at RAND, Inc.*

The goal is to give the **current** school credit for changing the trajectory of its students so that they can meet various academic attainment levels in the future. How is this to be measured? Consider an evaluation to see if a school's fifth graders are on pace to meet or exceed the proficiency designation in 8th grade. By using the data from the most recent 8th grade completing cohort, models can be developed which will allow projections for the current 5th graders as to their likely 8th grade scores assuming the same future schooling experience as the current 8th grade cohort received. Thus, the projected score enables an evaluation to see if each student is likely to exceed the 8th grade proficient standard. If so, the **current** school receives a positive credit. The percent of projected proficient students is calculated and all of the regular AYP rules are applied to see if this school has made AYP.

What this approach accomplishes is to use all of each student's prior scores—instead of just one score as in the simple case—to give a more reliable measure of the impact of the **current** school on the rate of progress of its student population. In other words, this approach uses multivariate, longitudinal data from students from the current school to provide estimates to map into a future scale without the test error problem. Additionally, this approach avoids the inherent non-linearity problem of vertically scaled test data in that this approach only requires the assumption of linearity between the prior scores and the future score; an assumption that is easy to verify empirically.

The author raised questions about the reliability of the projected values.

> *But the use of "projected scores" by Tennessee introduces an additional error factor. The Tennessee growth model substitutes a predicted or expected score for the AMO. Tennessee shows that the correlation of the predicted scores with actual scores is about .80 (R=.80). This means the percentage of variance accounted for is only about two-thirds (.8 squared = 64 percent); thus, one-third of the variance in actual scores is not accounted for by the predictive model. While an R of .80 is quite respectable in the research world, it may not be adequate for making real-world decisions. Many students will be counted as being on track when they are not, and vice versa.*

The projected scores have much smaller levels of uncertainty than progress measures based upon one prior score. It is true that the projected values in the Tennessee application do not consider future schooling effectiveness and will reduce somewhat the relationship between the projected scores and the observed scores in the future. However, the objective is to give the current school credit, not to hold the current educators' evaluation hostage to what future educators may or may not do! Additionally, it is most important to acknowledge and give Tennessee credit in the fact that all students' projections are used in the determination of percent projected proficiency, not merely those students who did not have the proficiency designation. In other words, students who are currently designated to be proficient but whose projected values fall below the future proficiency designation will count as a negative, providing an incentive to focus on the progress rates of all students and to minimize the focus on just the "bubble kids."

---

[1] This is true because the covariance structure among the prior scores is not related to test error. For the Tennessee application, if a student does not have at least three prior scores no projection is made and a student's current determination of proficient or not is included in the percent projected proficient calculation.

APPENDIX 1 (continued)

Response to Zeno's paradox assertion

The author spent considerable energy and space in the paper asserting that the Tennessee projection model is unwittingly trapped in Zeno's paradox. He asserts that students can make small amounts of progress, yet have their projected scores exceed the future proficiency level. Since the next year the projection targets are reset to another grade, this will allow schools to "get by" with suboptimal growth, resulting in students not obtaining proficiency due to the modeling itself. We dispute the author's assertion! The author states:

> *The disadvantage is that Mary's target will always be "on the way" to proficiency because, under Tennessee's model, Mary's goals are recalculated each year. Her target fifth-grade score is one that is estimated to put her on the path to proficiency by eighth grade. Her sixth- through eighth-grade scores are recalculated each year based on a score projected to be on the path to a goal of proficiency by high school.*

As was previously stated, the goal is to evaluate the progress made in the **current** school. The totality of Mary's data provides the estimate of Mary's future attainment. If Mary has a 'good' year, then her projected achievement level goes up. The future distribution that Mary's projected score maps into has the same proficiency standard as has been approved for regular AYP determination. Additionally, and most importantly to the Zeno paradox argument, if the cut score for 7th and 8th grade is essentially at the same place in the statewide distribution, then it does not matter to which distribution her projected scores are mapped—so *de facto* there is no remapping. This is essentially the case for Tennessee's 7th and 8th grade Math and Reading/Language Arts proficiency cut scores. The author's resetting argument has no relevance and Zeno's paradox does not apply.

Other responses

He further asserts:

> *As long as each student makes a small amount of progress toward proficiency, the school could hypothetically make AYP through the growth model even though not a single student had actually gained proficiency. This is not only true in any given year, but also could be true for a period of more than the three years implied.*

The idea that a school can have all students making a small amount of progress toward proficiency and yet make AYP with no proficient students is unreasonable. Just because a student makes a little progress, it does not mean that his or her projection will be greater than the target value. A school with all students not proficient would have a large number who are very low in academic attainment. These students would need to have made substantial academic progress to be projected to being proficient within three or fewer years. This would require more than a small amount of progress from each student. If the very low achieving students are projected to proficiency in three years, then their growth trajectories must have changed substantially. The author's conjecture that only small amounts of growth are necessary to meet projected proficiency is just not accurate.

Another comment:

> *But relative to the Tennessee growth model, the safe harbor provision has three advantages. First, it is based on real data, rather than a projection. Second, it is based on students achieving a set, policy-driven target—proficient—rather than a moving, amorphous, and norm-referenced target (i.e., a projected score), which has many more variables.*

This whole passage is misleading. As was previously mentioned, percent proficiency calculations, as used in safe harbor, are estimates based on test scores with errors. The projections are estimates based upon much more data and contain more information than is in the one test score for the safe harbor estimates. The statement, "*rather than a moving, amorphous, and norm-referenced target*," is totally wrong. There is not a norm-referenced target: the projections are to established proficiency cut scores for future grades.

APPENDIX 1 (continued)

The author further states:

> *The Tennessee growth model also means that schools in Tennessee will be able to make AYP in 2014 without all students being proficient.*

This statement could be applied to all growth models, safe harbor, and the present AYP status model as well if test errors are taken into account. To single out the Tennessee model with such a declarative statement without careful consideration of all of the uncertainties around the estimates from other models is inappropriate. As was stated earlier, many of the opinions expressed in this paper ignore the innate test errors in one set of AYP approaches yet attempt to magnify uncertainties in the Tennessee projection model. In fact, this is an area in which the projection model has advantage over the others.

The following is a statement that is clearly and provably wrong:

> *Schools will be more likely to make AYP under the projection model than the other two models.*

In Tennessee, many more schools make AYP with the status approach than either with safe harbor or projections, and many more schools make AYP through safe harbor than do through the growth model.

Not in the current paper, but in a recent blog by the author the author stated:

> *Some of the methodology and most of the data are proprietary, meaning they are privately owned, i.e., no public access. This all makes it very difficult for even Ph.D. and J.D.-level policy experts to get a handle on what is going on (which I found as a peer-reviewer last year), let alone teachers, parents, and the general public.*

Exact descriptions of the methodology deployed in the Tennessee projection calculations have been published in the open literature since 2005.[2] Additionally, the proposals to utilize this methodology have been reviewed and approved by four different peer review teams assigned by the USDE. Also, at the request of Congress, the early approved proposals were reviewed by a team from the GAO. In that review, the software for the Tennessee calculations was reviewed and evaluated to give an independent evaluation as to computational accuracy.

Agreements with the author

We agree with some of the author's comments.

> *The use of growth models represents an opportunity to improve upon the state accountability systems currently in use under NCLB. NCLB's focus on a single criterion, proficiency, and its lack of focus on continuous academic progress short of proficiency, fails to recognize schools that may be making significant gains in student achievement and may encourage so-called "educational triage."*

> *The model does offer some advantages. By setting goals short of, but on a statistically projected path to, proficiency, the model may provide an incentive to focus efforts—at least in the short-term—on a wider range of students, including both those close to and those farther away from the proficiency benchmark. It also may more fairly credit, again in the short-term, those schools and districts that are making significant progress that would not be reflected in the percentage of students who have met or exceeded the proficiency benchmark.*

We also agree with the author that Congress and the Secretary should review and learn from the growth models which have been approved. After working with longitudinal student achievement data generally for nearly three decades and working with various models to be used as augmentations for AYP specifically, I have

---

[2] Wright, Sanders, and Rivers (2005, "Measurement of Academic Growth of Individual Students toward Variable and Meaningful Academic Standards", in R. W. Lissitz (ed.) *Longitudinal and Value Added Modeling of Student Performance*, Maple Grove, MN, JAM Press).

APPENDIX 1 (continued)

formed some opinions that I hope are worthy of serious consideration:

- Simplicity of calculation, under the banner of transparency, is a poor trade-off for reliability of information. Some of the more simplistic growth models sweep under the rug some serious non-trivial scaling, reliability and bias issues. The approved models for Tennessee, Ohio and Pennsylvania represent a major step in eliminating some of these problems.
- Reauthorization of NCLB should put more focus on the academic progress rates of all students, not merely the lowest achieving students. Our research has shown for years that some of the students with the greatest inequitable academic opportunities are the early average and above average students in schools with high concentrations of poor and minority students. Too many of these students are meeting the proficiency standards, yet their academic attainment is sliding.
- Serious consideration should be given to setting future academic standards to various attainment levels. For instance, for Tennessee we provide projections to proficiency levels (regular and advanced), to minimal high school graduation requirements, to levels necessary for a student to avoid being vulnerable to taking a college remedial course, and to levels required to be competitive in various college majors. Some or all of these could be included in an AYP reauthorization with some careful thought. States which presently have these capabilities should be encouraged to move forward. Moving to these concepts will tend to avoid the conflict over what cut score the word 'proficiency' should be attached.

APPENDIX 2

RESPONSE TO THE COMMENTS OF DR. WILLIAM SANDERS RE: *ARE WE THERE YET? WHAT POLICYMAKERS CAN LEARN FROM TENNESSEE'S GROWTH MODEL*

By Charles Barone

First, I appreciate Dr. William Sanders taking the time to converse about the "Are We There Yet?" paper.

The Tennessee growth model, like those of the other 14 states in which growth models are in use, is a pilot program being conducted through time-limited waivers of federal statutory requirements. The purpose is to try something out, learn from it, and use the results to inform future policy efforts. This was the very reason I wrote "AWTY?" and why Education Sector published it.

I actually think that the paper addresses all the points raised in Sanders' response, and here, for the sake of brevity, I will focus only on the key points. In most, though not all cases, it is, in my opinion, a matter of emphasis rather than real difference.

The Fallacy of "Failing Schools." There are a couple of points raised in the opening paragraph that I will address later, but there is an overarching point that I think is implicit in this debate about NCLB in general and AYP in particular that I want to bring into the open.

In talking about a school that was doing well "normatively" i.e., relative to other schools (in terms of percentile ranks) at some grade levels, Sanders states:

> *Relative to NCLB, this was a failing school. In our view, this was not a failing school; rather it was an outstanding school and should not suffer the indignity of the failing school label.*

Nothing in NCLB labels a school as "failing." Why this is a common misperception (and why Sanders has bought into it) is a topic for another discussion, but it's indisputable that many perceive the law as ascribing this label to schools "in need of improvement." It seems to me that the school Sanders cites was likely neither failing nor "outstanding" but somewhere within the wide gulf between those two poles. The whole point of growth models, I thought, was to calibrate the differences between extremes, not to throw schools into one of two "either-or" (or dichotomous) categories.

The real purpose of federal law is to identify areas where students are falling short—by grade and subject area—and to direct resources to them early and as intensively as is necessary and appropriate. Doing otherwise is a state and local choice and, I would add, a misguided one.

Those involved in creating the NCLB law felt that, **prior to enactment of the law in 2002**, schools were able to hide behind average across groups and, it appears in Tennessee, across grade levels i.e., **were used in a way that obscured areas in need of improvement rather than illuminated them**. Average elementary school scores can hide deficiencies in third grade that would be better to address early. Average scores of all students can hide gaps between black and Latino students and their non-minority peers. Composites across subjects can hide subject area-specific shortcomings.

By bringing those problems to light, and funneling resources to those areas as early and as surgically (or radically) as needed and as is possible, it is hoped that students will get a better education and that potential long-term problems will be addressed sooner rather than latter.

Hence, in the paper we make this point:

> *The Tennessee growth model will also reduce the number of schools identified by NCLB*

APPENDIX 2 (continued)

> *as falling short academically. This could be a positive change if it allows the state to focus more intensely on the lowest-performing schools. However, it will also mean that some schools may not be able to avail themselves of resources that could help address student learning problems early enough to prevent future academic failure.*

It sounds like what we had in Tennessee was a labeling problem—calling all schools that did not make AYP "failing" rather than an AYP problem per se. I think most educators seeing a third grade with scores in the 26th percentile statewide (with one of the lowest set of standards in the nation) would want to address that problem promptly in the antecedent years (i.e., by improving what happens in pre-K, kindergarten, first, and second grade) rather than waiting two years to see what happens in fifth grade. Other states have gradations of not making AYP and target their interventions accordingly (such as at one grade level or in one subject) including interventions at grades prior to the grades in which testing begins. The law offers wide leeway to do so.

The third grade case cited by Sanders is particularly in need of attention, as stated in the "AWTY?" paper:

> *Slowing down the pace at which students are expected to learn academic skills in elementary school may create long-term problems for students and create larger and more difficult burdens for public education in junior high, high school, and beyond. A wide body of research suggests, for example, that children who do not acquire language skills in the early grades have an increasingly difficult time catching up to their peers as they progress. This parallels neuropsychological research that shows critical periods for brain development in language and other areas of cognitive functioning.*

Statistical Error. Sanders states that Tennessee rejected looking at non-statistically-derived scores (i.e., hard targets, rather than estimates) in part:

> *Because (1) the error of measurement in any one score for any one student is so relatively large that the setting of an improvement target by this method will inevitably send the wrong signals for many students.*

Here, as at other points in the paper, Sanders seems to assert that the projected score model gets rid of measurement error. It doesn't. **Measurement error is inherent in any test score** (as in any grading system). Sanders' method uses the same tests as every other AYP model in use in Tennessee and the other 49 states.

What the projected score model does is introduce an additional source of error, "prediction" error (the difference between a projected score that a multiple regression analysis *estimates* will put a student on the path to proficiency and the actual score that *would* do so).

This is pointed out in the paper, but unaddressed in Sanders' comments:

> *…the use of "projected scores" by Tennessee introduces an additional error factor. The Tennessee growth model substitutes a predicted or expected score for the AMO. Tennessee shows that the correlation of the predicted scores with actual scores is about .80 (R=.80). This means the percentage of variance accounted for is only about two-thirds (.8 squared = 64 percent); thus, one-third of the variance in actual scores is not accounted for by the predictive model. While an R of .80 is quite respectable in the research world, it may not be adequate for making real-world decisions. Many students will be counted as being on track when they are not, and vice versa.*

Sanders goes on to state that:

> *To minimize the error of measurement problem associated with one test score, we elected to use the entire observational vector of all prior scores for each student. In some of our*

APPENDIX 2 (continued)

> *earlier work, we had found that if at least three prior scores are used, then the error of measurement problem is dampened to be no longer of concern.*

But what he does not mention is that current law allows this option (using "rolling three year" averages of scores) whether or not a projected model is used.

Zeno's Paradox Issue. The "AWTY?" paper concludes that many students under the Tennessee model will take longer than three years to reach proficiency even if they meet their minimum "projected" score three years in a row.

Sanders states, through reasoning I could not quite follow:

> *The author's resetting argument has no relevance and Zeno's paradox does not apply.*

I stand by the conclusions of the paper. I challenge Sanders, or anyone else for that matter, to show me an instance where:

1) there is a long-term goal (e.g., X distance in Y years)
2) there is an interim goal that is some fraction of X progress for some fraction of Y years and;
3) the interim goals are re-calculated each year for a fraction of the remaining distance to Y;

in which it doesn't take longer than Y years to get there.

Sanders could of course clear all of this up by taking, say, 100 cases where we can see the projected scores for each student, each year, and where the student exactly makes each interim goal, to show us what happens in Tennessee in this instance over three successive years. As the paper shows, however, since the data and exact methods are proprietary, none of us can do this on our own, or we would have simulated such an instance in the paper.

On this point, Sanders states:

> *Exact descriptions of the methodology deployed in the Tennessee projection calculations have been published in the open literature since 2005. Additionally, the proposals to utilize this methodology have been reviewed and approved by four different peer review teams assigned by the USDE.*

It is true that the multiple regression formula that Sanders uses can be found in Tennessee's application for the federal waiver, as well as in most elementary statistics books. Tennessee's materials also include descriptions of some of the methods and adjustments that are specific to the Tennessee growth model.

But the details—including standard deviations and standard errors of measurement of students within in a school, and the histories of individual students over multiple years—are not available. Thus no one, at least no one that I have talked to, can do a real replication.

In addition, I sat on a growth model peer review panel in 2008 in which other states submitted models based on that of Tennessee. Not a single person at the Department with whom I inquired understood that in 2013, the goal for Tennessee was still proficiency in 2016, not 2014, and I think any casual observer of former Secretary Spellings' comments over the last few years can attest to that.

Size of Adequate Yearly Progress. Sanders disputes the paper's contention about the interaction between the growth model's incremental progress (extending the proficiency target over a number of years rather than proficiency each year) and Tennessee's low standards. But he merely skirts over the latter point.

First, I don't understand the artificial focus on those grades in which testing is done. If proficiency within three years is a doable goal, why not start with proficiency in fourth grade as a goal beginning in first grade (or kindergarten or pre-K) where research shows schools (and programs like Head Start or high-quality child care) can have an incredible impact? The state and its localities have all these resources at their disposal to

APPENDIX 2 (continued)

impact the years within and outside the 3-8 testing box imposed by NCLB. Why not do so? Is more federally imposed standardized testing, in earlier grades, what is required to bring this about? (I, for one, hope not.)

Second, no matter what grade you begin in, the Tennessee standard for proficiency is low compared to the NAEP standard—lower than virtually any other state.

Again, let me re-state a section of the paper which Sanders does not address:

> *In fourth-grade reading, for example, the NAEP benchmark for "basic" in reading is a score of 243; for "proficient" it is 281. The NAEP-equivalent score of the Tennessee standard for fourth-grade proficiency in reading is 222, which is about as far below the NAEP standard of basic as the NAEP standard for basic is below the NAEP standard of proficient. The NAEP benchmark for basic in fourth-grade math is 214; for proficient, it is 249. The NAEP equivalent of Tennessee's standard for proficient in math is 200.*

So if reaching proficiency in Tennessee is a low goal relative to other states (which the state of Tennessee acknowledges and is trying to change) relative to NAEP, then fractional progress toward that goal is, by definition, even lower.

How could it possibly be otherwise?

Linearity. Sanders asserts, regarding the Tennessee model, that:

> *This approach avoids the inherent non-linearity problem of vertically scaled test data in that this approach only requires the assumption of linearity between the prior scores and the future score; an assumption that is easy to verify empirically.*

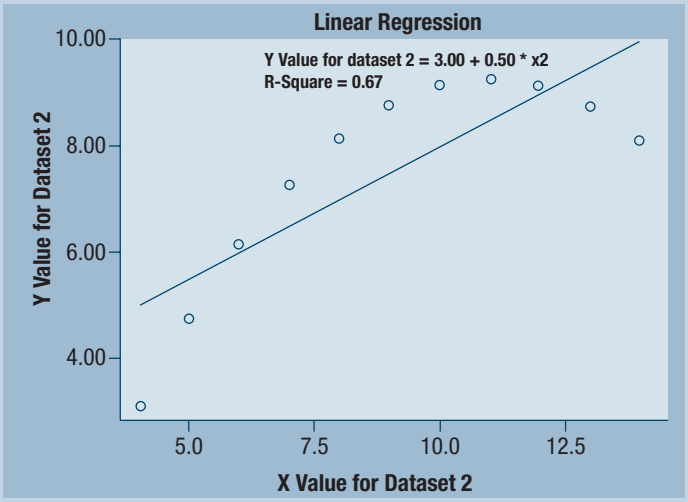
I chose not to go into this in the paper (for obvious reasons), but since the issue is being opened here, I think it should be addressed.

Linearity is a double-edged sword (stay with me until at least the chart on the next page). With vertical scaling, different tests can be equated across grades by re-scaling scores to make them comparable. We can't go into all the relative advantages and disadvantages of vertical scaling here. (Sanders is right that there are disadvantages.)

But I must point out that Sanders' assertion of the linearity of non-vertical scores in Tennessee—which he says are easy to verify empirically—may not always be true. (Note that Sanders does not supply empirical verification but only asserts it can be verified.) In turn, applying a linear regression, as Tennessee does, to estimate future scores may distort the relationship between real growth in student scores and those scores projected through the statistical model.

Let's say that over time, non-vertically scaled scores for some students are not linear but are parabolic (curvilinear) with accelerated growth in early years and a leveling off, and then a decrease in later years (a phenomenon not unknown in education research). Then let's say we try to map a linear regression onto this model (with an R squared of .67, similar to the Tennessee model with an R squared of .64).

The chart below illustrates this scenario.



Source: From SPSS Textbook Examples, Applied Regression Analysis, by John Fox, Chapter 3: Examining Data. UCLA: Academic Technology Services.

APPENDIX 2 (continued)

Here, the projected scores in the early years would be lower than the actual scores that would be seen over time. In this scenario, the linear model would set AYP goals below that which we should expect for students between ages 6 and 11. Conversely, the model would overestimate what we should expect for students over age 11.

This is just one of the many (virtually infinite) scenarios possible depending on student characteristics, age, and patterns of variance of scores for students in a particular school. The point is a linear regression only approximates, and in some cases can distort, educational reality.

Transparency. In closing, I would like to address the issue of transparency. In his remarks, Sanders says:

> *Simplicity of calculation, under the banner of transparency, is a poor trade-off for reliability of information. Some of the more simplistic growth models sweep under the rug some serious non-trivial scaling, reliability and bias issues. The approved models for Tennessee, Ohio and Pennsylvania represent a major step in eliminating some of these problems.*

This paper only speaks to Tennessee, and so we will leave the issue of other states aside.

But, as the paper shows, and as demonstrated here, the Tennessee growth model is not necessarily more reliable, accurate, or valid than those of other states using other growth models or the statutory "status" or "safe harbor" models. All represent tradeoffs.

While eliminating some problems, the Tennessee model creates others. For now, each state can reach its own conclusions about the relative strengths and weaknesses, and it was the hope that the "AWTY?" paper, and this discussion, will help better inform those decisions.

I do not, however, think transparency is an issue to be taken lightly. Real accountability only takes place when all participants in the education system—including parents, advocates, and teachers—can make informed choices.

I talked to a reporter from Texas this week (which is implementing an adapted form of the Sanders model, with at least a couple of key improvements per points raised here) who recalled her school days of independent reading assignments through the "SRA" method.

For those of you who do not remember, SRA was a box of large (roughly 8 x 11) cards, with readings and structured questions. The box progressed in difficulty from front to back (easiest to most difficult) with color-codings for varying levels of difficulty.

What the color-coding did was make understandable where you were—for yourself and the teacher—in progressing through a set of skills. The reporter pointed out that with the traditional method you would know, for example, that if you were at say red (the lowest) rather than violet (the highest) you knew you were farther back than you wanted to be by a certain time. Depending on the assigned color of where you were at (say red or orange) you also knew where you were *relative to* the end goal.

She then pointed out that with the Tennessee growth model method, we never know what the target color (or level of difficulty)—i.e., the interim "projected" score for a student by the end of the school year—is. It could be any color of the rainbow from red (below basic) to violet (proficient), and all we would know is that it was *somewhere* in between.

I think that all players in the educational policy and practice arena—educators, consumers, parents, advocates, and taxpayers—want, just as this reporter does, something a little more like the color-coded SRA system.[2] That is, they would like quite a bit more clarity than "trust us, your child is on a projected path to proficiency" within Y years (which, as we see here, is really Y + unknown # of years) according to the following formula:

$$\text{Projected Score} = MY + b1(X1 - M1) + b2(X2 - M2) + \ldots$$
$$= MY + xiT\ b^3$$

And, as much as I love statistics, I would assert that given that these individuals—educators, consumers, parents, advocates, and taxpayers—are the primary sponsors *and* intended beneficiaries of the educational system, we owe it to them to strive, as much as humanly possible, to meet their needs and expectations toward the goal of a better education for each and every child.

Endnotes

1 SRA is now the "Open Court" reading system. Its citing here by the author does not represent or imply any appraisal or endorsement.

2 Where MY, M1, etc. are estimated mean scores for the response variable (Y) and the predictor variables (Xs).