

VALUE ADDED ASSESSMENT



Achievement tests seem to draw fire from all sides, but in one form or another, they must be administered. Parents, teachers, legislators, and businesses want a legitimate way to evaluate a school system's effectiveness. Parents and teachers in particular want to know how much children really know in comparison to academic standards. These goals require instruments that measure the achievement level of individual children as well as schools in a given group. To be fair, the instruments must be "standardized": tests must be uniformly administered, identical tests, or tests made equivalent, must be offered in different locations and years, and scores must be expressed in a standard way that allows for fair comparisons.

Norm-Referenced



Norm-referenced assessment compares a student with others in his or her grade.

Criterion-Referenced



Criterion-referenced assessment compares a student with an established benchmark of achievement.

Value-Added



Value-added assessment compares a student's current level of achievement with his or her past level.

As an accountability model, traditional “norm-referenced” achievement tests have attracted sharp criticism over the years. Results are expressed by percentile rank, comparing students with each other, meaning half of all students are by definition “below average.” While important information can be gleaned from norm-referenced tests, simply comparing students with one another has limited usefulness in promoting or measuring excellence. Furthermore, a school’s scores frequently reflect the socio-economic status of its students rather than the work of the teaching staff.

Criterion-referenced tests began to gain favor with educators who wanted tests to better reflect actual achievement. In taking a criterion-referenced test, students are expected to demonstrate selected knowledge and skills, with scores reflecting a particular student’s standing in comparison to selected levels of achievement, rather than in comparison to other students. These tests have long been used as an excellent way to measure individual student achievement. At the most basic level, the challenges were to 1) find a way to “norm” criterion-referenced tests so they would also be valid and reliable tools to evaluate aggregate groups of students and schools, and 2) agree on standards.

“High stakes” tests resulted, such as our WASL (Washington Assessment on Student Learning). These tests (a combination of performance-based and criterion-referenced tests) are very controversial for a variety of reasons, one of them being the standards themselves. Some believe they are too high, causing a large proportion of children to fail, which creates resistance to the tests. Others believe the standards do not reflect an appropriate body of knowledge to begin with. Further, scores on the tests still primarily reflect the socio-economic composition of the student population, rather than what an individual student learns about a body of knowledge in a given year.

Few people inside or outside education trust the results of the current testing system and whether the sentiment is valid or not is outside the scope of this report. The underlying issue is finding an objective tool to evaluate student, teacher and school progress. Legitimate demands for accountability mean our education system must find some way to establish trustworthy evaluation measures.

For this reason, value-added assessment is worth a hard look. This relatively new method of looking at

test scores provides a way to ensure that all students are learning, and it offers an objective teacher evaluation method. Because it focuses on student growth during a prescribed period of time rather than absolute levels of achievement, value-added assessment is not tied to who the students are but to what goes on in a classroom during a specified period of time.

This report examines how various jurisdictions have implemented the concept of value-added assessment and the effects of implementation. It looks at potential challenges in using value-added assessment, examines different applications of the data, and provides recommendations.

What is value-added assessment?

As originally envisioned, the focus of value-added assessment is measuring student academic gain. But the value-added analysis may be done at the level of an individual student, a classroom, a teacher, a school, or a district. Most often, value-added assessment is used to discover the effects a teacher or school has on students.

For purposes of this report, value-added assessment is defined as *any method of analyzing student test data to ascertain students' growth in learning by comparing students' current level of learning to their own past learning*. This is in contrast to analyzing test data to measure students against an absolute standard of achievement, or to rank them against each other, or to evaluate a school's performance for accreditation purposes.

Value-added assessment is not a different type of testing program. Standardized tests (either normed or criterion-referenced) are used to obtain the scores. It is the act of comparing students' scores with their own past scores that distinguishes value-added assessment.

Value-added assessment, depending on the exact form it takes, has the potential to provide several benefits over other methods of analyzing student scores:

- **Focus:** Value-added assessment changes the focus of education statistics from quibbling over demographic factors to asking the essential question: How well are students progressing?

- **Equitable comparison:** By focusing on student growth, value-added provides a way to recognize outstanding student growth accomplished by teachers. This is especially important in schools with high populations of learning disadvantaged students.
- **Accountability:** Because value-added assessment provides results that are less tied to student demographics and more tied to teacher effectiveness, they provide a fairer accountability measure for schools and teachers.
- **Diagnostic:** Value-added assessment alone cannot identify the cause of poor student achievement. But where the data is sufficiently detailed, it can identify where failures and successes are occurring, giving staff a starting place to begin asking questions and making data-driven decisions. The more extensive forms of value-added provide a gold mine of data for education research.

Teachers matter. Value-added assessment operates on the assumption that a good teacher can create and facilitate student learning no matter what his or her students are like when they enter the classroom. Conclusions based on value-added data confirm this thesis: Teacher quality is the most important element in determining how much students learn. Excellent teachers are able to create adequate student growth in students at all achievement levels.

To be included in this report, value-added systems must measure the increase in knowledge in a particular student or group of students. Analyses that focus on the school's progress as a whole toward making certain goals (such as the Academic Performance Index growth measure in California, or the federal requirements for Adequate Yearly Progress) were not included. This is because the focus is on the school's gain (this year's third graders are compared to last year's third graders) rather than on the students' gain (this year's fourth graders are compared to themselves in third grade).

Jurisdiction	Date	Grades Tested	Test	Subject Areas	Report Detail	Methodology
Tennessee	1993	3–8	Terra Nova (CTB/McGraw Hill)	English/language arts, mathematics, science, social studies	District, school, teacher	Mixed-model statistics
Dallas	1992	1–12	Stanford 9/Aprenda; Texas Assessment of Academic Skills; Assessment of Course Performance; other measures	Reading and math (all grades); writing, science and social studies in selected grades	School, teacher	Multiple regressions and hierarchical linear modeling
North Carolina	1997	3–12	North Carolina’s End of Grade and End of Course assessments; other measures	Reading and math (3–8); 10 courses including history and science course in high school	School	Comparison between state average and student group average growth
Texas	1996	3–8	Texas Assessment of Academic Skills	Reading and math	School	Comparison between school growth and 40 similar schools
Arizona	1999	2–8	Stanford 9	Reading and math	School	Calculate percent of students maintaining or improving stanine standing

Approaches to value-added assessment

Value-added assessment is an objective evaluation of what student tests say students know and teachers do. It does not—cannot—determine whether correct content is being measured. Presumably, that decision was made before a test was adopted or administered.

Value-added assessment expresses only the concept of analyzing test results to determine the amount of student academic growth. Different states and local school districts have implemented this concept using different statistical methods, and have given it different roles in their education system. This section provides a brief overview of the basic approaches to value-added assessment in different jurisdictions. For a more detailed explanation of the history and development of value-added assessment in each jurisdiction, consult Appendix B.

Detailed statistical models

Tennessee is the state most strongly identified with value-added assessment. Its system dates back to 1992, when it was implemented as an integral part of a comprehensive education reform measure. Using a complex statistical method developed by Dr. William Sanders, then a statistician at the University of

Tennessee, the Tennessee Value-Added Assessment System (TVAAS) provides data to the public on the performance of districts and schools, and data for appropriate administrators on the performance of teachers.

The TVAAS statistical model aggregates student growth increases using a design that accommodates missing data. Because of a philosophical belief that schools should ensure all students progress at equivalent rates, no matter their disadvantages, the model does not include other data on students. Dr. Sanders now makes the TVAAS statistical analysis available commercially under the name EVAAS.

About the same time Tennessee incorporated TVAAS, the Dallas school district began implementing its own value-added analysis based on a statistical model developed within the district. The district issues School Effectiveness Indices, which use student growth measures based on test scores, as the primary part of an index that also includes other measures, such as dropout rates. Test scores are also aggregated at the classroom level.

The Dallas model predicts student growth based on multiple factors, including prior achievement and other factors such as ethnicity and socio-economic status. School-level factors, such as mobility and

overcrowding are also factored into the analysis. Schools that achieve growth significantly beyond prediction and that fulfill other requirements are given cash awards for all staff.

Simpler statistical models

North Carolina examines student growth at the school level as part of its “ABCs of Public Education” program, enacted in 1996. Schools making expected growth or high growth are given cash awards for all staff. Low-performing schools may receive targeted assistance from the State Board of Education. The North Carolina system is primarily based on the increase in the average score of a group of matched students in two successive years, with minor statistical adjustments.

In Texas, schools have been issued a Comparable Improvement rating since 1996. The comparable improvement aggregates matched student scores in successive years, and compares that growth with the growth in a set of 40 schools selected for their similarity in ethnic and socio-economic composition.

Arizona uses a rating system developed by Department of Education staff, which matches student data and then examines what percent of students remained in the same stanine (a standard nine-level ranking system issued with standardized test scores) from year to year. If students achieve at least the same stanine level as they progress a grade, they are considered to have attained “one year’s growth.”

Challenges in implementing value-added assessment

Statistical Issues

An independent, statistically sophisticated comparison of the various systems of value-added assessment would be a valuable addition to education literature, but it is beyond the scope of this report. The primary challenge in choosing a statistical method, however, is quite simple: The more statistically robust a system is, the more variables it can allow and the more precisely it can pinpoint results. But it is also more difficult for most educators and parents to understand.

This problem is particularly troubling when it comes to the point of analyzing teacher proficiency. Research has shown overwhelmingly that the primary influence on students’ growth is not their socio-economic status, or their school, but the quality of

their teachers.¹ Yet most of the systems attempting to analyze students’ rate of learning only examine data at the building level. They lack capacity to provide fair evaluations and guidance at a classroom level. Comparing student achievement growth rates may make for a fairer standard of comparison between schools than absolute achievement, but it does not provide the most fundamental information needed to improve student learning.

On the other hand, systems of teacher evaluation are politically charged. Introducing a complex system that provides rankings quite different from usual test scores will create an automatic target. Furthermore, the information will be useless unless teachers and administrators accept it as valid.

Not much can be done to make high-end statistical equations user-friendly. However, the concepts behind them can be broken down and explained. Districts and states wishing to implement a value-added model must ensure they find a way to explain the methodology in layman’s terms and convince participants that this data will be useful to them.

The best endorsement for a value-added approach, however, will be evidence of its real-world validity. If a system really works, its evaluation of a teacher or school should make sense to those with first-hand knowledge. Over time, ratings should be logically consistent—not changing radically when methods and personnel have remained relatively constant. If the statistical system exhibits these qualities, the honest skeptic should be convinced that it does provide a fair and accurate way to assess student learning.

When using a computer, for example, people do not need to understand the entire mathematical system behind it; they simply need to observe that the machine in front of them responds in predictable ways. Similar verification can be used to build trust for value-added analysis.

Whatever system of value-added assessment is used, and whatever use is made of it, flaws are inevitable. Comparing scale scores requires an assumption that each point of growth represents an equal increase in difficulty—something impossible to prove because no one can measure “difficulty” objectively. There will always be errors in measurement of student learning, and errors in calculating the growth between them. Even though these measures may not be “perfect,” if they succeed reasonably well at measuring real student growth, emphasizing them

will encourage teachers and administrators to focus on student academic growth.

Political challenges

While it may appear that the decision to select a value-added assessment model hinges on objective facts, it is really more a matter of policy and politics. For example, should expected growth for students be adjusted based on their socio-economic status and other factors? Should we expect at-risk students to increase learning at a slower rate, or should all students be expected to achieve the same amount of learning each year, whatever their starting point? What about the argument that it is best to require higher rates of learning for students who start out further behind, to help them catch up with those who had greater advantage at the beginning?

Another difficult question is whether students who have only attended a particular school for a short time should be included in value-added analysis. On one hand, it is unfair to hold schools accountable for students they have had little opportunity to teach. On the other hand, mobile student groups tend to be lower achieving, and if they are always excluded from expectations, who will be held accountable for helping them? Tennessee balances these concerns by including mobile students (where data can be matched) in the school score but not in the teacher score. Most other jurisdictions do not include data from a student unless they have spent the better part of the school year—sometimes two years—in that particular school. For schools with a fairly mobile population, this tends to skew scores upward.

Students in special education programs and bilingual education programs provide another challenge. To be counted, they must be tested in a way that provides a fair equivalent to the general tests. For special education students, an additional problem is raised related to whether or not growth rates can be expected to be comparable. More research would be needed to determine whether this would be a fair comparison, but preliminary research conducted by Dr. Sanders indicates it would be.² Where testing exemptions can be made justly for certain classes of students, the inherent danger is that lower-achieving students will be classified solely to avoid testing them.

But the most political question of all about value-added assessment: “What do we do with the results once we have them?” That question will be addressed in a subsequent section.

Cost and time investment

Value-added assessment, by its nature, requires frequent standardized testing to plot student growth. Annual testing is necessary, especially if the data is to be gathered at the grade or teacher level. Fortunately, value-added assessment does not require its own specialized instrument. It can use scores from any reliable instrument that is sufficiently correlated with the curriculum (perfect correlation is not necessary to measure growth), and that has enough stretch to measure growth of all levels of students. This allows a jurisdiction to include value-added data as part of the reports from another testing program. The jurisdiction must also have a database that tracks student results over multiple years.

Conducting the value-added analysis requires an additional investment of resources. The most statistically complex system, EVAAS, is available on the open market for \$1 a student and \$25 a teacher. Simpler analyses would probably be less expensive, and might be done in-house. A major expense is training staff on how to understand the data and use it in decision-making. With staff turnover, this is an ongoing process.

Every system of accountability requires an investment of time and money that would otherwise be spent elsewhere. A system that requires annual testing is likely to be attacked as diverting too much time from teaching to testing. Some testing is necessary, however, and value-added assessment provides great potential for using test results to identify academic problems and to reward good teaching.

Delay in useful data

A lapse exists between administering any test and receiving results. Value-added data has some additional time lapses. Because value-added measures growth over a year's time, no data is available at the baseline grade level (usually around the third grade). North Carolina addresses this problem by administering a pretest at the beginning of the third grade.

Modifications must be made if value-added data is to be used for teacher evaluations because the information gleaned generally is not available, processed and analyzed until the fall of the following year. Yet teacher evaluations are more likely to take place in the spring, simultaneous to testing. Dallas addresses this by using the value-added data from the prior year as part of a needs assessment in the fall; teachers are then evaluated in the spring for how they have addressed the needs identified.³

With the TVAAS system, teacher evaluation data is further delayed because three years' longitudinal data is required to ensure statistical accuracy. Thus, this data is not available to evaluate teachers in their first two years of teaching. Although value-added data may not always be available, where it is available, it provides useful and objective evidence to guide teacher evaluations.

Alignment of tests to teaching

Few propositions in testing seem more obvious than requiring that students be tested on the material they are being taught. But implementing this basic concept in the context of using standardized tests is always fraught with controversy. The process of creating and standardizing a test year after year, so that questions are different enough to ensure test security, yet results can be compared across years, is difficult and expensive. For this reason, most states and school districts contract out with private companies who sell standardized tests on a nationwide basis. Critics complain that these nationally available tests do not correspond adequately to the curriculum required in a particular state or district.

Even though districts and states may vary somewhat in curriculum requirements, the basic sequence and achievement levels expected in core subjects like reading comprehension, language arts, and math are unlikely to vary widely from the scope and sequence used as a basis for tests. Dr. Sanders indicates that for a value-added analysis to be valid, perfect correlation between the test and the curricular objectives is not necessary.⁴ The test need only be highly correlated so that the number and variety of questions is adequate to identify overall student growth.

When standardized assessments are discussed, the specter of "teaching to the test" is raised. Suffice it to say here, as long as the test is properly secured and updated, the harm only occurs if the test requires knowledge or skills outside the scope of what teachers should be/can be teaching. No test can measure everything involved in good teaching, but good teaching is generally reflected in good test scores.

Value-added assessment can be used with norm-referenced or criterion-referenced tests, as long as scores can be converted to a scaled score measurable across various levels of the test. But certain criterion-referenced tests provide problems for a value-added assessment. Criterion-referenced assessments may

place their emphasis on questions close to the standard, leaving too few questions at the high and low end to assess the progress of very high and very low achieving students. The Texas Assessment of Academic Skills, for example, had this problem, and thus very high and low achieving students do not have their scores included in Texas' Comparable Improvement measure.⁵ The consequence: Schools are not held accountable for ensuring that all students are making adequate gains each year.

Variety of content in high school

In grades K-8, students usually follow the same course of study and are tested on overall development in language arts and math, and often in science and social studies as well. However, once students reach high school their paths begin to diverge—some progress further in mathematics or sciences than others, and courses may be taken in different order. It becomes difficult to measure overall growth each year. North Carolina and Dallas both provide for standardized, course-specific tests at the end of selected core subjects. North Carolina also has a comprehensive test in reading and math that is administered in the tenth grade, measuring progress over the eighth grade scores. Tennessee has included end-of-course analyses for some mathematics courses, but most of the program has been delayed due to lack of funding.

Since an end-of-course test does not measure growth from year to year, some other predictor of student achievement must be selected. For example, eighth grade mathematics scores might be used as a predictor for Algebra I; student scores could then be evaluated to see whether they learned more, less, or as much as would be expected based on their previous level of mathematics achievement.

Value-added in action

Value-added assessment has tremendous potential to produce the reams of detailed data that would bring joy to any researcher's heart. But unless school boards, superintendents, principals and teachers have the commitment and the incentive to use that data to improve student learning, gathering the information is a waste of time. The next section profiles how this data can be used in practice.

Maryville Middle School

Principal Joel Giffin,
of the award-winning
Maryville Middle School



Maryville Middle School

Three-year-average gains, 2001

Expressed as percent of national norm gains

Math	139.6%
Reading	155.0%
Language	140.4%
Social Studies	93.8%
Science	141.0%

in Maryville, Tennessee describes himself as a firm believer in the Tennessee value-added assessment system.⁷ He uses the data provided through TVAAS to drive the decision-making process throughout the school. This has required him to understand the data himself and to ensure that teachers understand it as well.

Giffin works to ensure the first use of TVAAS data is to reward and recognize staff for successes. TVAAS data is then used to pinpoint problem areas. For example, at one point all students were making adequate or even exemplary gains in math, except a small group of low-performing students. This group of students—20 out of 340 in that grade—was not large enough to drag down the overall growth rate, but the precision of TVAAS data allowed the school to target that group of students for help.

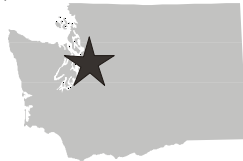
Teachers were given a list of the low-gaining students and asked to consider what they knew about them. A common list of challenges was given: low socioeconomic status, attendance problems, behavior problems, dislike of school, and lack of money for supplies. Giffin then challenged the teachers to dig deeper and identify which of the students' challenges the school could address. Finally staff identified the low-performing students as lacking adequate support in doing their homework. A second math period was added to their day to give them more individualized support in math. In the next testing cycle, that group of students showed a 356% gain in math knowledge (compared to 100% standard gain).

Teachers at Maryville are taught how to understand and use the TVAAS data for themselves, as well as in consultation with the principal. By combining TVAAS

data with the criterion-referenced data, particular strong or weak points can be identified. The data might indicate, for example, that Teacher X's students were low in math gains because the students did not understand fractions, but Teacher Y's students showed a strong understanding of fractions. Teacher X would be assigned to observe Teacher Y's instruction in fractions and vice versa. The teachers would then confer with each other and with the principal, and a plan would be drafted to change teaching methodology to improve student learning in that area. Next year's data would be used to determine whether the change was effective.

These uses of information indicate that obtaining value-added data is only be the first step. Principals and teachers must examine the data at their school to identify where they are succeeding and where they need to improve. (Unfortunately, most of the systems of value-added assessment do not provide sufficiently precise data to perform many of the analyses conducted at Maryville Middle School.) Value-added assessment does not prescribe remedies—teachers and principals must draw from their own knowledge and experience to create a plan to address the problems. Once the plan has been implemented, next year's value-added data and other feedback must be used to evaluate the success of the plans.

Seattle



During the mid-1990s, the Seattle School District began looking for ways to improve the school accountability system within

the district. Through research and personal contacts, the district staff and superintendent became aware of the work of Dr. William Sanders. Dr. Sanders was invited to speak to district staff and, in 1999, district leadership decided to contract with Dr. Sanders for review of data from Seattle schools. The Sanders approach was chosen for several reasons: It was well-documented, readily available, and it allowed the district to use data it already had on hand rather than placing a new testing burden on teachers or students.⁸

The district was ultimately able to obtain a three-year grant to pay for the costs of obtaining and institutionalizing the value-added analysis, making it a regular part of schools' planning. This grant enabled them to hire additional staff to provide training and coordinate the project. As project manager they hired

Marsha Denton, Ed. D. Dr. Denton had worked as a teacher in Tennessee and used the TVAAS data to improve her own teaching methods. She had also worked as a consultant and advisor for other teachers using TVAAS data.

Training staff to use the data is the primary challenge. Dr. Denton conducts numerous training sessions and Dr. Sanders has also visited the district to provide training. Dr. Denton describes the immediate purpose of value-added data as giving the staff a place to start asking questions on how they can improve student learning. The data allows them to begin identifying strong points and weak points, and then look to the strong points to learn how they can make improvements in other areas. Teachers are given the information to talk across different grade levels, not only about the students, but also about strategies. Some principals have also presented the data to parent groups, explaining strengths, weaknesses, and what the school plans to do with the information.⁹

Currently the district only obtains the data at the school building level, but would like to obtain teacher-level data if it proves possible to do so without the data being made public. The analysis combines data from the Iowa Test of Basic Skills and the Washington Assessment of Student Learning. EVAAS is now providing the data in a form that examines the growth of the district as a whole across time, not just comparing schools to the district average each year.¹⁰

Using value-added data

The most difficult question related to value-added assessment is, once we have the data, what should we do with it? The investment required by value-added analysis can only be justified if it pays off in improving student learning. States and districts are only beginning to explore the potential for the data, so accomplishing this goal will require careful planning and review.

Teacher evaluations

The most innovative but controversial proposal for using value-added data is teacher evaluation. On one hand, it holds the prospect of using hard data, not just personal impressions, to identify excellent or deficient teaching. On the other hand, it flies in the face of the last few decades of consensus in the education hierarchy that evaluating teachers on the

basis of student achievement is unfair because so many other factors may be reflected in student test scores.

Value-added data challenges the assumption that student test scores cannot be attributed to teachers. Certainly achievement levels reflect many outside factors besides teachers. But when student growth is examined, the primary influence is the quality of the teacher. This does not mean no other factors exist, but the correlation is strong enough to make value-added assessment a potentially revolutionary change in the way teachers are evaluated.

Using value-added data as a portion of teacher evaluations requires addressing several challenges, first of which is having a statistical system robust enough to bring data down to the teacher level. Currently only the TVAAS and Dallas systems offer this level of analysis. Second, students must be properly attributed to teachers. This requires identifying how much time a student spent under a particular teacher and for which subjects.

Getting data, however, is only the beginning. Using its results in constructive practice is far more difficult. Two possible uses exist: formative—to help teachers determine how to improve; and summative—to provide information for decisions on teacher standing. How the data is used in the formative evaluation process will always be primarily dependent on the individual administrator and teachers. The state and district can mandate value-added data be included, but its presence will have little impact if teachers and principals do not believe the data is useful. Extensive training will be necessary to get principals and teachers to understand the data and view it as a resource for identifying strengths and weaknesses, and to use that knowledge to expand strengths and shore up weak areas.

The real controversy is using value-added data in summative, high-stakes decisions, like teacher pay or termination. So far no one uses the data in this way, although in Dallas value-added scores may be part of the information used in a termination decision. Obviously, using the data in high-stakes decisions requires great caution. It probably should not be used in any high-stakes decisions until it has first been implemented, verified, and accepted in a lower-stakes context. Even then the data can never stand alone. Data cannot tell which teacher was assigned the most rambunctious students, or which one struggled with illness. Data from standardized tests do not help assess

competence in some subjects or activity areas, like art or physical education.

Value-added analysis for teachers generally includes statistical safeguards designed to prevent a teacher's score from being unfairly skewed by insufficient data. This means data tends to be skewed toward the middle, with the vast majority of teachers ranking in a broad middle band of acceptable performance. Thus, the most likely potential for high-stakes use of value-added data is at either end—the smaller percentage of teachers whose performance is so high or so low that it breaks away from the average.

For the one group, it seems unreasonable that a principal could know that, year after year, a teacher was not giving students adequate opportunity to learn, yet be unable to do anything about it. No decisions should be made in haste, on inadequate data, or without considering other factors, but ultimately the principal ought to have recourse to use the data to establish a teacher's incompetence.

On the other hand, it seems logical to recognize and even financially reward teachers who are experiencing dramatic success—success measured by an objective evaluation of how much their students are learning. By recognizing and rewarding success we make it easier for others to copy it. Again, test score gains alone may not provide an adequate basis for making this decision, but there is no reason why they cannot ultimately be part of the determination.

Whether value-added data is used for summative evaluation or not, the primary emphasis must always be on helping teachers improve. The vast majority of teachers, like everyone else, want to do well, and will respond favorably to value-added data if they perceive it as providing them with information on how to be more effective.

Building-level planning

Every current system of value-added assessment provides data at the building level. General information at the building level, however, does not provide much information about where problems are occurring. When data is at least broken down by grade level, it can provide a starting place for asking questions. For example, if fifth graders are achieving much lower gains in math than fourth graders, perhaps the math curriculum for fifth graders is not sufficiently challenging to continue growth in students who made excellent progress last year. Teachers who have this

data can discuss student challenges and strategy across grade levels.

As more detailed data is available, more specific questions can be asked. For example, EVAAS's ability to classify data by high, middle, and low-performing students allows the school to examine whether a particular group of students is not making adequate gains. Often this is the high-performing students, perhaps the consequence of teachers concentrating on helping lower-performing students catch up.¹¹ On the other hand, in the example of Maryville Middle School, detailed data was used to target low-performing students. These examples illustrate why detailed data is vital to diagnostic use. Perhaps the school as a whole is doing well, yet certain groups are falling behind. The more those groups can be identified, the better they can be helped.

If student records can also be tagged for participation in various programs, schools can identify whether these supplemental programs are having the desired effect.

Incentives and accountability

Although value-added data can provide a valuable diagnostic service, the primary reason most value-added systems have been implemented is accountability. Parents, legislators, businesses and the community at large all want some way of discovering how schools are doing. Traditional measurements generally reflect the demographics of students rather than how well students have been taught. Value-added data offers the potential to hold schools and teachers accountable for ensuring student success, no matter if or how those students may be disadvantaged.

Simply having value-added data provides teachers and principals with a strong incentive to improve, particularly if the data is specific enough to provide guidance of which areas to target. Most people want to succeed. With specific data, they can get a better idea of how to do that, and they can receive feedback on how they are improving.

But there still may be a place for financial rewards. Financial rewards do not imply that teachers are mercenaries who have no motivation for helping students other than cold, hard cash. Sometimes, providing financial rewards is simply the best way to demonstrate seriousness about improving student learning. As the saying goes, "Put your money where your mouth is." All people appreciate and respond to recognition and tangible rewards.

This raises the question of how financial rewards should be appropriated. Both Dallas and North Carolina provide cash awards to teachers and other staff at high-achieving schools. No system currently rewards teachers on a purely individual achievement basis. North Carolina has no teacher-level data anyway, but in Dallas rewarding at the school level was a conscious decision:

The intent of the performance awards made at the school level was to encourage cooperation and assistance within a school building. The Accountability Task Force rejected any plan in which a teacher might be encouraged to withhold information or assistance from a fellow teacher in a school.¹²

Another concern about providing rewards exclusively on an individual basis is the potential for abuse by building administrators. For example, an administrator might concentrate the difficult students in one teacher's class (either out of vindictiveness or simply because that teacher handled them better). If the building as a whole is evaluated and rewarded, the administrator has an incentive to see that all teachers are encouraged to excel.

Logistical problems to rewarding teachers on a value-added basis also exist. It is highly unlikely that all teachers will ever be evaluated under a value-added system. The value-added system provides no measurement for teachers in grades too young for standardized testing, the grade used as a baseline, and subjects not in standardized tests. Thus cash awards given solely on a value-added basis would exclude many teachers. If such awards were instituted, some way would need to be found to evaluate teachers outside the value-added system and also to recognize their achievements.

The problem with using group instead of individual rewards is that it violates the reality of human nature that we perform better if we know it matters to our own rational self-interest. Rewarding an entire group of people means some in that group will be unjustly rewarded because they will not have participated in the group's success.

Nonetheless, monetary rewards are worth strong consideration.

A lack of perfect correlation between value-added analysis and quality of teaching should not necessarily preclude its use. Currently, teachers are paid based on years in the classroom and degree of education—measures that have little direct correlation with student learning.

Inter-building coordination

Just as value-added data can provide a starting point for teachers to talk across grades, it can also provide a starting point for teachers to talk across schools. One finding from the Tennessee data is that learning tends to drop off dramatically when populations of students transfer from one school to another, as when students are promoted from grade school to middle school.¹³ The likely reason seems to be a lack of proper coordination between the curriculum of the feeder schools and the receiving school. If both schools have grade-specific data, they can begin looking for where problems occur, discuss how to address them, and verify whether their efforts have been successful.

Meeting the needs of all students

Much of education reform talk for the past two decades has centered around establishing and raising standards: attempting to ensure the vast majority of students reach a particular benchmark of achievement. Such an objective seems like a logical objective for a



When value-added assessment is combined with standards, schools can verify that all standards are meeting a minimum level of achievement, while higher-achieving students continue to be challenged.

public school system charged with serving all students. But standards-based reforms also have some inherent challenges. The state may want to ensure a basic education for all students, yet all students are not the same. Some may only achieve the benchmark after much struggle; some may be capable of vastly exceeding the benchmark; some may never achieve the benchmark. Setting an achievement level standard that addresses the needs of all students is impossible. Set the standard too high, and you guarantee failure for large numbers of students. Set the standard too low, and you eliminate any challenge for high-achieving students.

On the other hand, requiring schools only to meet a particular standard of growth has its own limitations. If schools simply must show average growth for all students, those who start out far behind might never reach basic achievement levels.

By combining measurement based on standards and measurement based on growth, schools can legitimately be held accountable for meeting the needs of all students. Schools can be required to bring all students to a basic level of achievement, and simultaneously be required to challenge students who have left basic achievement far behind. Once the data has been analyzed to determine what rate of achievement can reasonably be expected of lower-performing students, teachers can work to ensure students are learning at a rate that will allow them ultimately to reach acceptable levels.

Combining the two measurements will require making some policy decisions. For instance, if greater growth is needed from lower-achieving students to bring them up to a benchmark, resources need to be targeted to accomplish this. This in turn may aggravate the problem that higher-achieving students tend to make slightly lesser gains on average, perhaps because teachers are already focusing on lower-achieving students.¹⁴ Policy-makers must determine to what extent this is acceptable to ensure all students learn a minimum amount, and to what extent schools should endeavor to offer equal growth to all students. Value-added data at least allows schools to make these decisions consciously and determine how well they are succeeding with the goals they have set.

Educational research

Value-added assessment data, particularly the detailed data gathered by the TVAAS/EVAAS system, offers tremendous potential to conduct education research. Focusing on growth already controls many

factors, making it easier to identify the impact of specific variables. Already value-added data has provided insights and verification of what has the most significant effect on student learning.

Studies based on value-added data confirm the thesis that teacher quality has the greatest impact on student learning, and that high-quality teachers are able to obtain gains despite a wide mixture of ability in their students.¹⁵ Further, only the top tier of teachers, as measured by the rate of gains made by students overall, were able to show adequate levels of growth at all student achievement levels; less effective teachers tended not to obtain adequate gains in higher-achieving students.¹⁶

While individual teachers are important to students, so is the sequence of teachers. One study looked at student test scores after three years under teachers of varying effectiveness. For students initially scoring in the same range, the difference in test scores between a sequence of three low-performing teachers in a row and three high-performing teachers in a row was as wide as 54 percentile points.¹⁷

As teacher quality has proved to be the most important factor in student gains, other factors have proved statistically insignificant. The percent of minority or low socioeconomic status students, for example, was found to be unrelated to the academic growth rates a school could achieve.¹⁸

The availability of value-added data opens up new possibilities for research. Dr. Sanders has identified several topics he is currently investigating or that merit further study. This includes examining the effectiveness of teachers with different certificates or years of experience, comparing teaching abilities of teachers who quit with those who remain in the profession, and analyzing whether, when states only test certain subjects in certain years, student growth in those subjects increases the year of the test and trails off in other years.¹⁹

Conclusion

Value-added assessment is still a relatively new concept in education. The massive amounts of data tracking and analysis required made it unworkable until computer power increased. Already it shows great potential for changing the focus of education statistics from looking solely at achievement levels to examining student growth. This in turn offers an opportunity to identify practices that speed up or hamper student growth.

No amount of statistics can eliminate all uncertainties or inequities in evaluating schools and teachers. But value-added analysis does provide a better tool to focus on the real issue in education: What and how much are students learning? As we turn the focus to answering this question, annual statistical analysis may prove to be only the beginning. Teachers can target their classroom assessments to make sure their students are moving forward throughout the year. They can ensure that the lessons they are offering are appropriate for their particular group of students. Principals can determine if school configurations and programs are having their desired effect.

In the end, value-added assessment is simply a statistically sophisticated way to help good teachers and administrators do what they have always tried to do: find out where students are academically and take them as far as they can go.

Recommendations

By integrating value-added assessment into its testing program, a state or school district gives itself a powerful tool for transforming student test data into information that can potentially improve student learning. The examples of different jurisdictions that have used a value-added analysis provide some guidelines as to what elements are necessary to make this potential tool effective:

- **Sound statistical analysis:** The statistical system used must be reviewed by statistical experts to ensure fairness and accuracy for the level of detail offered. The standard of growth chosen and data included should be carefully reviewed to logically support the state's policies. The system should be capable of explanation, at least in concept, to those who will be receiving and acting on the data. Finally, the data should make sense in comparison to the real-world experiences of those involved.
- **Adequate data detail:** To provide the greatest amount of information for making data-driven decisions, data should be available at the district, building, grade, and teacher level. It would also be advantageous to be able to compare data by achievement level, and

perhaps by other considerations such as ethnicity or socio-economic status, in order to identify any specific subgroups not making adequate gains. If teachers were also able to look at individual student levels, they might be able to make better determinations of how best to help each student in their class.

- **Appropriate publicity and training:** Because value-added assessment is an unfamiliar approach to test scores for most people, implementing it will require ensuring that the media, parents, and community understand what the new scores mean. Educators must not only understand what the scores mean, but be thoroughly trained in how to use the scores in decision-making. That training will need to be ongoing to accommodate staff turnover.
- **Commitment to using data to improve student learning:** Value-added systems have generally been enacted first as an accountability measure. But accountability itself exists to further what should be the end goal of every action in the school system: increasing student learning. Obtaining value-added data is not enough. Principals and teachers, especially, must be committed to using the data to ask questions about how learning can be improved.
- **Adequate resources:** Value-added assessment will require expenditures to provide for annual testing, data analysis, and staff training. Further funds may also be necessary to provide incentives or to implement necessary changes identified after obtaining the data.
- **Incentives for high performance:** Although value-added data itself provides incentives for schools and teachers to find ways to improve learning, providing cash rewards does give the state an opportunity to communicate the depth of its commitment to increasing student learning. And the data itself is only useful to the extent that parents, educators and the community take it seriously.

Appendix A: Glossary

Achievement test

A test that measures what knowledge a student has acquired in one or more common content areas. This is in contrast to tests that may measure aptitude or readiness.

Composite

A single score that combines the scores on different tests or other performance measures.

Criterion-referenced test

A test in which every item is tied to a specific educational objective, designed to determine which of the objectives have been mastered. Scores on a criterion-referenced test are generally expressed as levels of achievement.

Mean gain

The average gain in scores of a selected group of students.

Norm-referenced test

A standardized test that compares a student or group of students with a specific reference group, usually students of the same grade.

Percentile

Percentile ranking divides all student scores into 100 groups of equal size. Rank is expressed as one of 99 point scores; for example the 57th percentile is a score higher than 57% of the overall scores.

Reliability

The reliability of a test is the indicator of its consistency. A reliable test is one in which the same individual taking the same test on different occasions or a test with different but equivalent items would get the same score.

Regression to the mean

A statistical phenomenon in which scores that are distant from the mean tend to move closer toward the mean upon retesting. Thus students scoring high at one session are likely to score not quite as high, compared to everyone else, the second time. Similarly low scoring students are likely to score not quite as low. To find the likely true gain of student scores, the numbers must be adjusted for this phenomenon.

Quartiles

Division of a population into four equal groups.

Scaled score

Scores on a single scale with intervals of equal size—thus each gain of a point represents an equivalent increase in knowledge. A scale score allows comparison across grade levels and tests. Thus a scale score of 150 in one level of the test would reflect the same level of knowledge as a scale score of 150 on the subsequent level of the test, even though raw scores would be different because the second test would be more difficult. Scale scores can properly be added, subtracted and averaged across test levels.

Standard deviation

A statistical expression used to show how far the set of scores is spread from the average of all scores. One standard deviation above and below the mean includes approximately two-thirds of all scores. A large standard deviation indicates that there is a wide variety in scores.

Standard error of measurement

The standard error is an estimate of how much a statistic is likely to deviate from the true measurement.

Standardized test

A test administered in compliance with directions for uniform administration. The items in the test must be appropriate in difficulty for the students taking the test, and conform to the planned contents. The test must be interpreted using reliable norms or standards.

Stanines

A standard score scale (see above) that divides the norm population into nine groups. The mean is stanine five.

Statistically significant

A statistically significant difference is one that is greater than could be expected from purely random variation.

Value-added assessment

A method of analyzing student standardized test scores to determine and compare the amount student knowledge has grown over time in a matched group of students.

Validity

Validity refers to whether a test is capable of measuring what it is used to measure.

Appendix B: Methods for value-added assessment

Tennessee

History and description

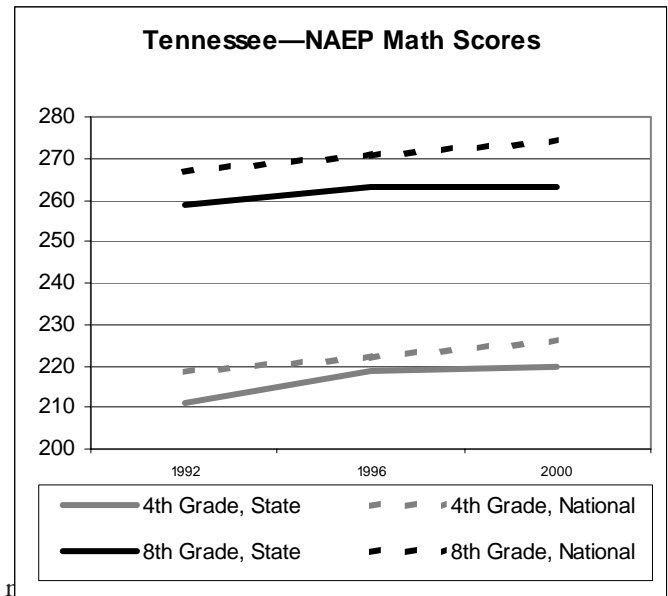
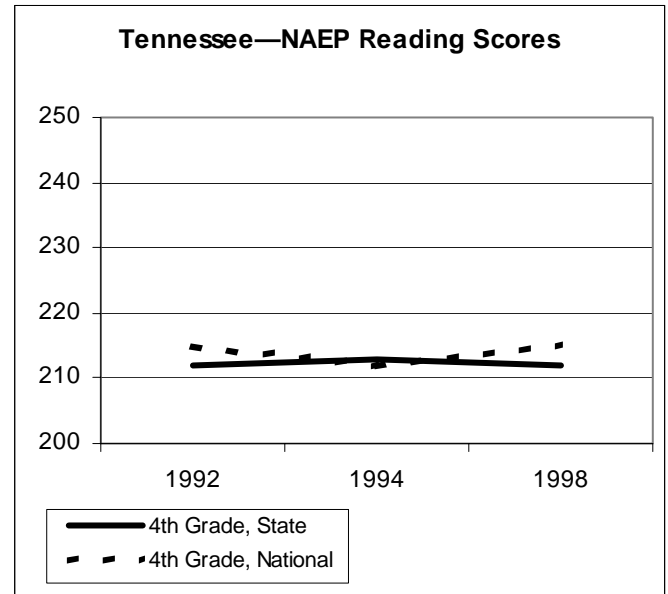


In 1992, Tennessee became the first state to adopt a value-added model statewide, the Tennessee Value-Added Assessment System (TVAAS). Its system remains the best known, most detailed, and most statistically sophisticated example of implementing value-added assessment.

The primary developer of TVAAS was Dr. William Sanders, formerly a statistician at the University of Tennessee. In the early 1980s, Tennessee was examining the possibility of awarding merit pay for teachers. In response to statements that it was impossible to evaluate teachers fairly based on student achievement, Sanders and a colleague theorized that a statistical model developed in agriculture (mixed-model) could be used to discover how much a teacher's class had learned. They gained permission to examine three years' worth of test data from the Knox County schools and found that by examining student growth rather than absolute test scores, and correlating data by classrooms, they could estimate teacher effectiveness in ways that were consistent from year to year, and that also fit with the subjective impressions of school administrators.²⁰

Despite these findings, the study failed to attract much attention at the time. However, in 1992 the Tennessee legislature undertook another round of education reform, one that would require raising taxes. Business interests were demanding that accountability for districts, schools, and teachers be part of the package.²¹ This time legislators were attracted to Sanders' proposal as a way to verify results. After inviting Sanders to speak, legislators amended the state's Educational Improvement Act to incorporate the "Sanders Model." Schools and systems would be expected to have a mean gain in student learning that would meet or exceed the national mean gain. As of 1995, data would be analyzed at the teacher level and used in teacher evaluations.²²

Through 1997, Tennessee used data from the CTBS/4 test by CTB/McGraw Hill, testing second through eighth grades. Since 1998, Tennessee has tested third through eighth grades using the Terra Nova test by CTB/McGraw Hill.²³ Terra Nova is a



and constructed response questions, and provides both norm and criterion referenced results. Students are tested in reading, math, language, social studies and science. System and school scores, expressed as an average of the last three years' gains, are made public. Scores are expressed as a percentage: a score of 100% reflects normal gains.

The evaluation system for secondary schools is still being developed, and currently includes three end-of-course tests for math, a writing assessment, and the ACT. Actual student scores are compared with predicted scores based on their Terra Nova scores in earlier grades.

In addition to the public reports, district superintendents control access to the web-based

reports that provide more detailed information, allowing authorized users to track how students at different levels of achievement are performing and to look at individual student performance over time.

If a school or system does not meet the performance standard of making gains equal to or greater than the national mean gains, and has not made statistically significant progress toward that goal, it may be placed on notice or on probation. While a school or system is on notice, the Department of Education and the Office of Education Accountability must study the school or system and, if it is placed on probation, they may restrict local powers in order to implement recommendations. If low gains continue, the Commissioner of Education may recommend that the local board and superintendent be removed from office.²⁴ This sanction, however, has not yet been implemented.²⁵

The use of value-added assessments to evaluate individual teachers makes Tennessee unique among states. (The Dallas School District also applies its value-added information to teachers.) At least three years' worth of data are used to ensure evaluations are based on long-term tendencies rather than a one-year fluke. Special-education students and students who have been in a teacher's classroom for less than 150 days are not included in that teacher's analysis.²⁶ Unlike district and school data, teacher-specific data is kept confidential.

Tennessee does not provide for the use of teacher data for high-stakes decisions such as termination or merit pay. Rather, the value-added data is one part of the teacher evaluation process: actual use of the data varies based on how the administrator and teacher want to use the data. Ideally, the data is used to provide objective feedback pinpointing problems and highlighting successes, which teachers and administrators can then use for decision-making. Critics have complained that this limited use of data in teacher evaluations does not live up to the original plan to use the data for teacher accountability.²⁷

Even in the use of evaluations, a review by the Tennessee Comptroller's Office of Education Accountability indicated that, although there were principals and teachers using the data effectively, most Tennessee schools were not using TVAAS data in a way that would help improve student learning.²⁸ The report concluded that although TVAAS data had great potential to improve student learning, for that to

happen "the state must broaden its purpose from a tool for reviewing student performance to a tool impacting student achievement."²⁹

Statistical model of TVAAS

A conceptual view of TVAAS can be obtained by imagining an "academic growth chart" for each individual student, charting the student's rate of growth over multiple years. Like a physical growth chart, this chart shows times of more rapid growth and times of slower growth. When the charts of all students in a classroom, school, or district are correlated, educators can spot areas where learning is taking place more slowly or more rapidly.

The statistical model required to correlate this growth is much more complicated, because real life testing scenarios are much more complicated. Students miss tests, move between schools, or have a bad day on test day. These and other complications require sophisticated statistical analyses to insure reliable measures of the influences of districts, schools and teachers on the rate of academic growth of students.

The most significant differences between TVAAS, which uses mixed-model statistics, and less sophisticated methodologies are treatment of missing data; approach to non-teacher variables; and accommodation of different real-world teaching scenarios.

1. Unlike most other value-added models, students with only partial testing data are not dropped from the TVAAS analyses. By giving these students' scores proper weight in analyzing the school effects, students who only test sporadically are not simply allowed to drop through the cracks of the accountability system.

2. TVAAS limits itself to examining achievement test data. Rather than trying to quantify and factor in all the non-teacher variables that may affect student learning—from socio-economic status to family crises—the statistical analysis only looks at a student's past achievement levels, since those levels already reflect the student's situation in life. Each student's past achievement serves as his or her own "control." Research has indicated that when the data is examined in this way, such factors as socio-economic standing, race, etc., do not have a statistically significant effect on rates of student growth.³⁰

Two reasons are given for omitting non-teacher variables: (a) the variables are often difficult to quantify and some data simply may not be available; (b) including

these factors would mean setting lower standards of growth for students considered to be at-risk. This is regarded as unfair to students with the potential to be high achievers who happen to be in at-risk categories.

3. The teacher analyses accommodate different teaching situations, such as multiple teachers for the same student in a given subject as well as teachers of self-contained classrooms. Teachers in higher grades who offer departmentalized instruction can also receive analyses.

In addition to reporting the growth rates of all students, TVAAS can break down that data by achievement levels, charting the progress of high, low, and average scoring students separately. This allows schools and teachers to discover which groups of students are not making adequate progress and make curriculum adjustments accordingly.

EVAAS

Originally, Dr. Sanders provided the TVAAS analysis through the Value-Added Research and Assessment Center at the University of Tennessee. However, in 2000 he moved to SAS in School, part of the SAS Institute Software Company, and from there, under the name EVAAS, offers the same data analysis as TVAAS to districts around the country. EVAAS provides the data at \$1 per student for schools and districts, \$25 each for teacher reports. Web delivery is an additional \$1.50 per student, which provides a look at the data over time in different breakdowns—by grade level, prior achievement level, and even individual student. This provides teachers and administrators with specific information to pinpoint strengths and weaknesses and track improvement.

EVAAS provides training for districts on understanding and using the data. Specialized research can also be undertaken. EVAAS staff is now working with over 100 districts outside of Tennessee.

Since most states only test in a few grades (such as 4, 8 and 10), most districts must piece together data from multiple tests—often a state criterion-referenced test in some grades and a norm-referenced, nationally available standardized test in others. The researchers have found that once test scores are translated into scaled scores, there is strong correlation for growth rates from year to year, regardless of which test is used.³¹

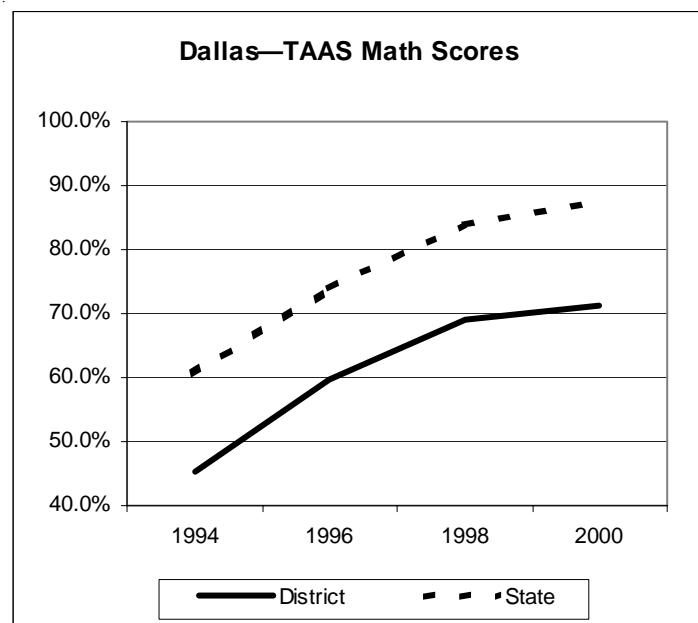
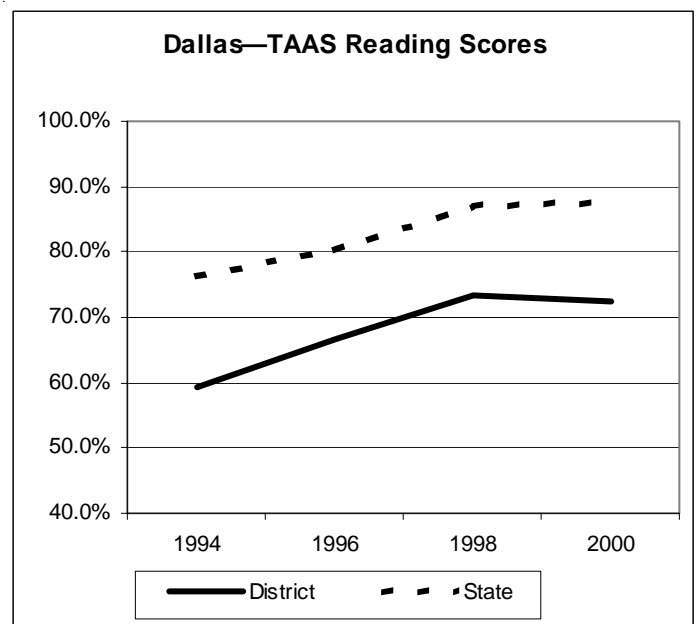
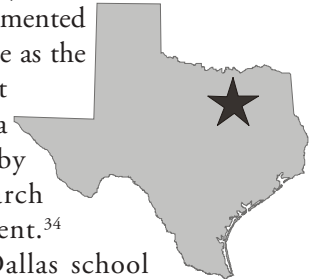
Few districts have yet obtained the data down to the teacher level, partly because they have not been keeping data correlating students with teachers,³² and partly because such data might be a public record under existing state laws.³³ However, teachers and

administrators are interested in obtaining this data, and districts are working to overcome the barriers.

Dallas

History and description

The Dallas system of value-added assessment was implemented in 1992, about the same time as the Tennessee system, but it reflected the outgrowth of a decade of data gathering by the school district's research and development department.³⁴ Beginning in 1984, the Dallas school



district ranked schools based on student growth curves. This system was eventually abandoned because of a new state accountability system. In 1990, the Dallas Board of Education established a Commission for Education Excellence. The commission recommended developing an accountability system that incorporated the previous work, including other factors but still giving priority to growth as reflected by test scores, and extending the analysis to the teacher level. The current Dallas accountability system was built in response to this report.³⁵

At the school level, the Dallas accountability system provides for School Effectiveness Indices. The district's Accountability Task Force, a group of parents, teachers, principals, and community representatives, selects and gives weight to the goals that measure the effectiveness of a school.³⁶ This incorporates multiple assessment instruments, including the national norm-referenced Stanford 9/ Aprenda (the Spanish version) in grades 1-9, the state-mandated, criterion-referenced Texas Assessment of Academic Skills (TAAS) in grades 3-8 and 10, and the Assessment of Course Performance (ACP), criterion-referenced tests for specific high school courses.³⁷ Non-test indicators are also included, such as dropout rates, graduation rates, and enrollment in advanced courses.³⁸ However, test scores have the greatest weight.

A student's scores only count toward the school's ranking if the student was enrolled during the first six weeks of school and took the test at the end of the year. Schools are required to test at least 95% of eligible student population.³⁹ The purpose for these requirements is to increase the fairness of the system by only holding schools responsible for the education of students whom they have had the opportunity to teach, while not allowing schools to skew test scores by not testing otherwise eligible students.

Schools are recognized for exceeding predicted academic growth and meeting other standards in a multi-tier system. "Gold Star Schools," those at least one-half standard deviation above prediction, are given cash awards. This includes \$2000 for the school activity fund, \$1000 for professional personnel and \$500 for support personnel, adjusted for individual attendance.⁴⁰ Awards are not provided to individual teachers outside high-performing schools, because the

Accountability Task Force thought this might undercut community and teamwork among school staff.⁴¹

Beginning in 1995, Teacher Effectiveness Indexes were calculated for elementary and middle school teachers; they were renamed "Classroom Effectiveness Indicators" later. This data is used to identify "needs" in the process of creating teachers' and principals' improvement plans for the following year.⁴² Teachers and principals are then responsible for identifying how those needs will be addressed. Although classroom effectiveness indicators, especially if low for multiple years, may provide part of the data that is ultimately used in a termination decision, they do not stand alone in making that decision.

Statistical method of Dallas system

The Dallas system focuses on the gains of individual students, as aggregated at the classroom and school level. The idea is to predict how well each student can be expected to do based on previous performance and a host of personal and school factors. The degree to which students exceed this prediction is considered the value added by the teacher or school.⁴³

The statistical method involves a two-stage analysis. Using multiple regression, student predicted and actual scores are adjusted for such factors as ethnicity, gender, language proficiency, and socio-economic status, creating residuals that reflect the difference between the expected and actual achievement. Then this data is taken through a two-level hierarchical linear modeling analysis to adjust for school-level factors such as mobility, crowding, overall socio-economic status, and percentage minority.⁴⁴

To be recognized for comparable improvement, schools must exceed prediction as well as having at least one-half of the school's tested students outgrow the national norm group on the Stanford 9 in reading and mathematics.⁴⁵

For classroom effectiveness indicators the student data, reflecting the adjustments for individual students and the school, are grouped into classrooms.⁴⁶ In addition to eliminating students who have not been continuously enrolled in the school, the classroom-level analysis does not include students who have had excessive absences.⁴⁷

North Carolina

History and description



A value-added type of assessment is a key part of North Carolina's "ABCs of Public Education" accountability system. The State Board of Education developed the ABCs in response to the School-Based Management and Accountability Program, passed by the state Assembly in June 1996.⁴⁸ The ABCs system contains both a growth (value-added) standard and a performance standard. Elementary and middle schools have been evaluated since the 1996-97 school year, high schools since 1997-98.⁴⁹

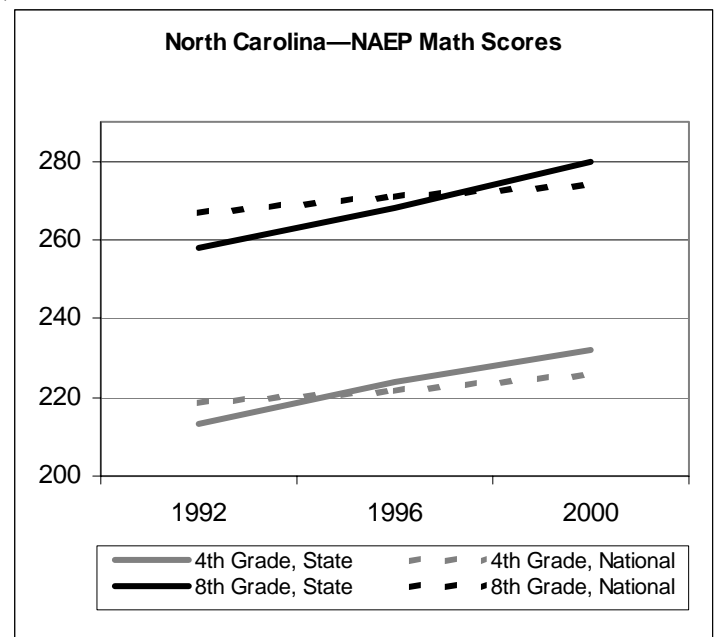
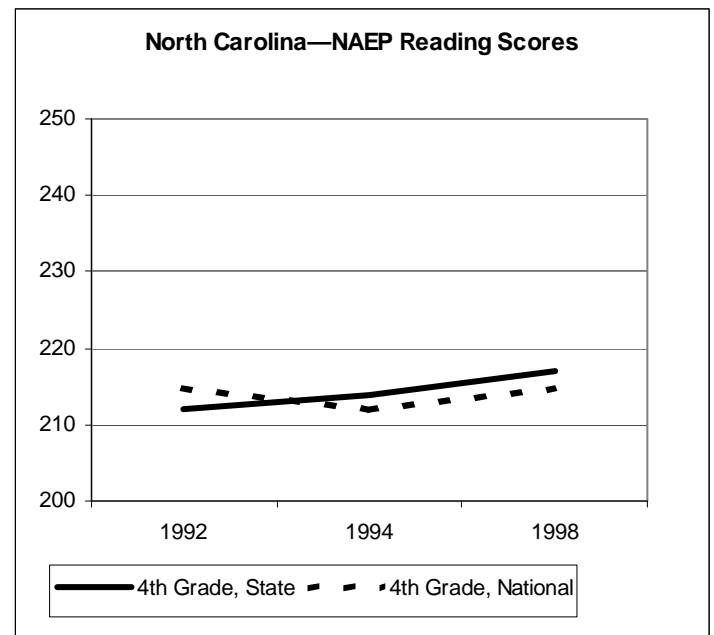
The ABC growth measure examines whether a school has met either expected growth or high growth standards, compared to a year selected as a baseline. Performance measures, in turn, look to what percentage of students in a school have met a certain benchmark. Together, the two measures are used for a comprehensive school recognition and classification program.

Recognition is provided to schools with high percentages of students meeting the performance standard, but cash grants are awarded to schools based on whether they make expected or high growth. Schools with high growth are awarded \$1,500 per certified staff and \$500 per teacher assistants. Schools making expected growth are awarded \$750 per certified staff and \$375 per teacher assistant.

Schools who do not make expected growth and who have less than 50% of their students reach the performance benchmarks are classified as "low performing schools." The State Board of Education assigns an Assistance Team to some of these schools and offers others assistance on a voluntary basis.⁵⁰ The Assistance Team is to review all facets of school operations and develop recommendations. This includes evaluating the certified personnel (principals and teachers), and may involve requiring staff to take competency tests or recommending the local board or state board dismiss staff.⁵¹ In 2000-2001, 31 schools (out of 2,137) were identified as low-performing.⁵²

Statistical method

North Carolina's testing system consists of two types of tests. Grades 3 through 8 are administered an End of Grade (EOG) test, a multiple-choice



criterion-referenced test created by North Carolina teachers and the state Department of Public Instruction to correspond to the state's Standard Course of Study. There is also a pre-test at the beginning of Grade 3, which is used as a baseline for the Grade 3 EOG test. Students must have been registered at the school for 91 days to count toward school scores on the EOG tests.⁵³ At the high school level, standardized End of Course (EOC) tests are administered in 10 core subjects.

North Carolina's data analysis model was based on feedback from the L. L. Thurstone Psychometric

Laboratory at the University of North Carolina at Chapel Hill.⁵⁴ The core measurement for actual growth in grades 3-8 is the EOG average scale score of a matched group of students in two successive years. The school's expected growth is calculated from the state average rate of growth at a particular grade level in a baseline year. This state baseline is then adjusted to create the expected growth standard for the grade: first, for proficiency, based on an assumption that higher-performing students are likely to advance faster, and also for regression to the mean.⁵⁵ The difference between the actual and expected growth is then divided by the standard deviation to accommodate different score spreads in different grades, and weighted based on number of students. The "weighted standard expected growth" for all grades is added together to get the "weighted expected growth composite." If this number is equal to or greater than zero, the school is considered to have met the expected growth.⁵⁶

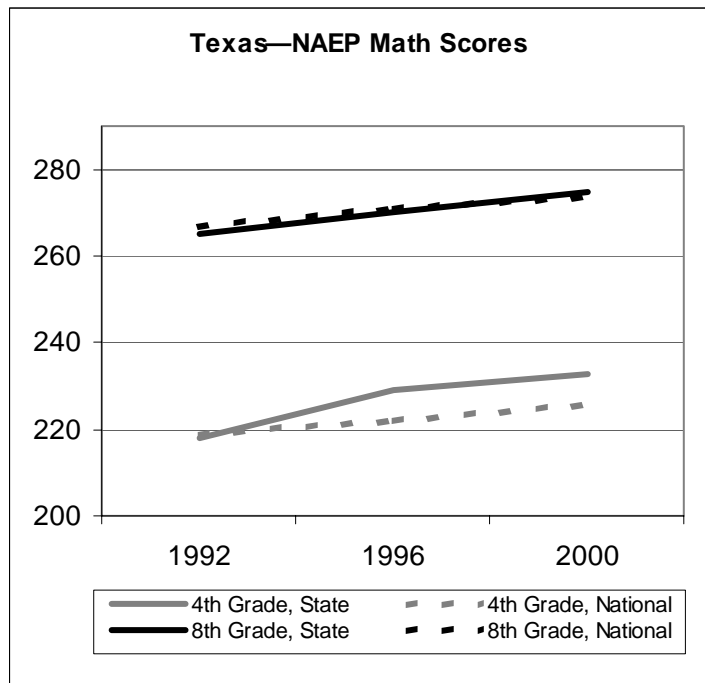
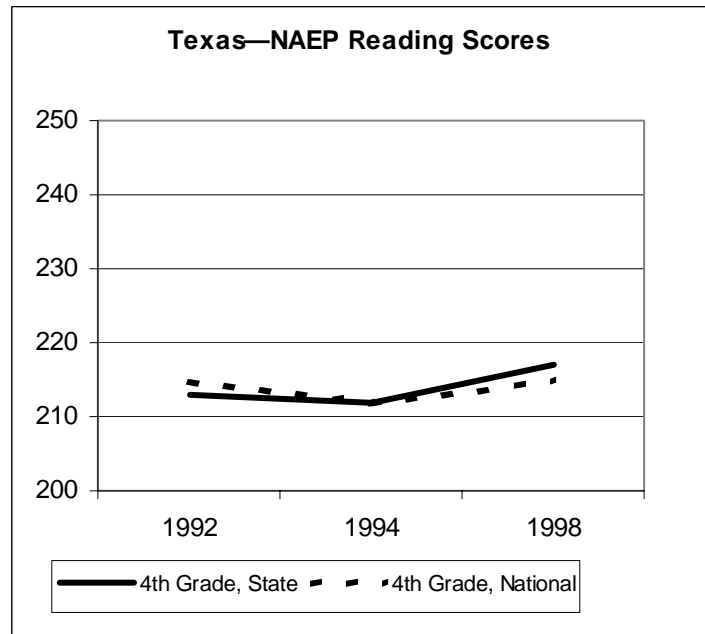
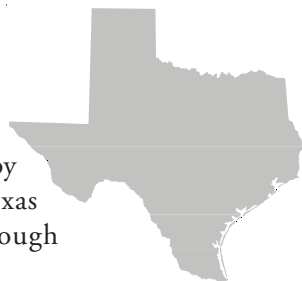
For high schools, growth is measured in different ways. To examine growth on the EOC tests, a prediction is made based on the state average as adjusted by previous scores in a selected, logically related test: for example, the predictor scores for Algebra I are the 8th grade mathematics scores. This predicted score is subtracted from the actual average score, and the result is adjusted for variety in score spreads.⁵⁷ Other factors that are compared are the growth on a comprehensive test in grade 10 over the end of grade test in grade 8, percent gain in students achieving a passing score on a competency test, comparison of participation in college prep courses across two years, and comparison of dropout rates across two years. These weighted standard expected growth scores in each area are added together; if the composite is greater than or equal to zero, the high school is considered to have met expected growth.⁵⁸

The calculation as to whether the school has reached high growth is similar, but the higher growth target (10% more in 3-8, 3% more in high school) is used for calculation instead.⁵⁹

Texas

History and description

A value-added type analysis of schools, called Comparable Improvement (CI), is required by state statute as part of the Texas accountability system.⁶⁰ Although



required since the current system was originally implemented in 1993, due to lack of student-level growth measures, it was not implemented until the 1995-96 school year.⁶¹ Texas currently uses a criterion-referenced test known as the Texas Assessment of Academic Skills (TAAS). This test is administered to grades 3-8 and exit level in reading and mathematics, to grades 4, 8 and exit level in writing, and in science and social studies at grade 8.

In the 2002-2003 school year, Texas is implementing a new testing system known as the Texas Assessment of

Knowledge and Skills. Comparable Improvement will be calculated based on the new instrument, assuming appropriate measurements are available to assess year-to-year progress.⁶²

Statistical analysis

To calculate the CI, Texas Learning Index (TLI) scores are used, which show achievement both above and below the TAAS proficiency cut-off. To count in the analysis, students must be matched from two successive school years. Students scoring very high or very low are not included in the analysis, since the test does not adequately measure growth in students scoring at the extremes.⁶³

For each matched student, the previous year's score is subtracted from the current year's score. If the result is zero, one year's growth has occurred; higher scores indicate more rapid learning. These scores are then averaged for the school in the categories of reading and mathematics.

For each school, Texas then selects a unique group of 40 schools that are most demographically similar. The characteristics used to select schools are percentages of students identified as African American, Hispanic, and White, economically disadvantaged, Limited English Proficient, and percent of mobile students. This comparison group is divided into quartiles based on their students' average growth. A school's Comparable Improvement is expressed by which quartile their school fits in. Thus, a school marked as Q1 is in the top 25% of schools in that particular group. A separate quartile assignment is given for reading and for mathematics.

Arizona

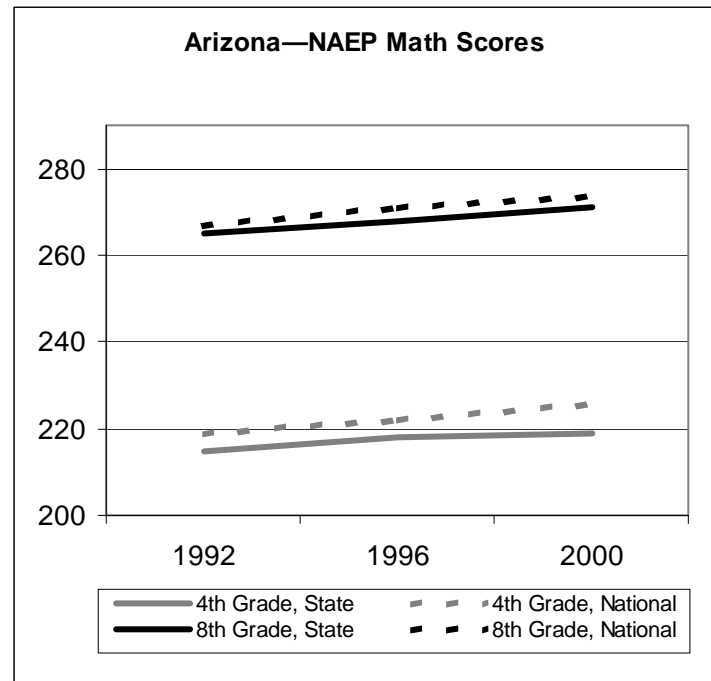
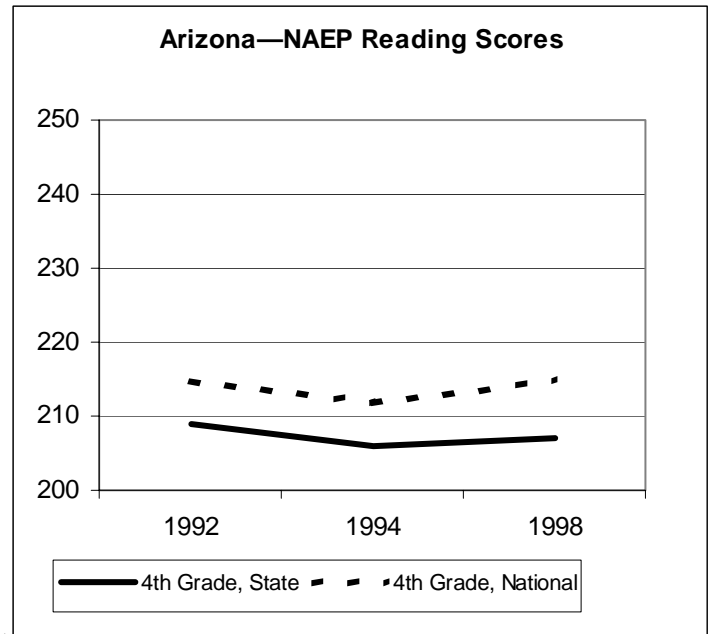
History and description



The Arizona Department of Education's Research and Policy division developed a Measure of Academic Progress (MAP) in an effort to add a value-added analysis to the state's reports on school test scores. The concept was inspired by the TVAAS system,⁶⁴ but uses a

relatively simple statistical analysis. The report has been issued since the 1998-99 school year.

Originally MAP was a staff initiative. Subsequent laws on education in Arizona (particularly Proposition 301, passed in November 2000) have made academic progress a necessary part of school accountability and



of developing proposals for performance-based pay scales.⁶⁵

MAP provides only a school ranking system, although with the relatively simple analysis involved, a teacher or principal could easily calculate the data at the classroom level.

In the first year MAP also included a Star Rating system, under which grade levels in schools were segregated into five groups based on where their adjusted growth score was in relation to others. Thus a score of five stars meant the grade level performed

better than 80% of others, while a score of one star meant the grade level is in the bottom 20% for the state.⁶⁶

The statistical simplicity of the MAP system has made it easy to explain to participants; the tradeoff for that simplicity has been a lack of precision. However, the data has proved useful in shifting the focus of education debate from absolute scores to student growth.

Statistical method

MAP is based on Stanford 9 scores, a nationally available norm-referenced test, for grades 2-8 in reading and math. The challenge for staff proved to be finding a way to match students, since no identifying numbers are used. Students are matched based on last name, first name, date of birth, and gender. Students can only be counted if they took the test in the same school two years in a row.⁶⁷ Thus, there is no measure of progress for the first year of junior high or middle school.

The statistical analysis of student scores, however, is quite simple. Originally, the analysis required calculating the mean scale score of all students in one grade, then subtracting it from the mean scale score of those same students in the following grade the next year. These scores were then standardized to adjust for regression to the mean, based on the greater ease schools with lower absolute test scores have in making gains.⁶⁸

As originally conducted, the analysis looked to whether each grade level at the school had, on average, achieved one year's growth (OYG). The OYG standard was set by calculating the number of points on the developmental scaled score between one grade and the next at the 50th percentile. Thus if the 3rd grade 50th percentile mark on the scaled score was 599 and the 4th grade mark was 625, the OYG standard is 26. A grade level at a school would then meet the OYG standard if its adjusted gain was 26 or greater.⁶⁹

Due to statute, the form of the analysis was changed starting with the 2000-2001 school year. Instead of looking at the average gain, MAP now examines what percent of students within a grade attain OYG, as measured by their position in stanines, a nine-point scale commonly used to report standardized test results.⁷⁰ Under this new system, a student who is in the same stanine score or higher as in the previous grade is considered to have achieved OYG.⁷¹

Endnotes

1. S. Paul Wright, Sandra P. Horn and William L. Sanders, "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation," *Journal of Personnel Evaluation in Education* 11 (1997): 66.
2. Dr. Sanders indicates that his data shows including special education students in a value-added analysis would not necessarily be unfair to teachers; many teachers are causing outstanding growth in these students. William L. Sanders, telephone conversation with Karen Helland, 18 April 2002.
3. William J. Webster and Robert L. Mendro, "The Dallas Value-Added Accountability System," in *Grading Teachers, Grading Schools*, ed. Jason Millman (Thousand Oaks, California: Corwin Press, 1997), 93.
4. William L. Sanders, conversation, 18 April 2002.
5. Department of Accountability Reporting and Research, *2002 Accountability Manual* (Dallas, Texas: Texas Education Agency, April 2002), 59-60.
6. Katie Cour, *Multiple Choices: Testing Students in Tennessee* (Nashville, Tennessee: Comptroller of the Treasury, March 2002), 15.
7. Information for this section obtained from Joel Giffin, a telephone conversation with Karen Helland, 9 May 2002.
8. Mike O'Connell, Seattle School District Director of Research, Evaluation and Assessment, telephone conversation with Karen Helland, 14 May 2002.
9. Marsha Denton, Value Added Project Manager, telephone conversation with Karen Helland, 10 May 2002.
10. William L. Sanders, conversation with Karen Helland, 22 May 2002.
11. Jeff Archer, "Sanders 101," *Education Week*, 5 May 1999.
12. Webster and Mendro, 88.
13. William L. Sanders and Sandra P. Horn, "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research," *Journal of Personnel Evaluation in Education* 12:3 (1998): 274-256.
14. Wright, Horn and Sanders, 65.
15. Wright, Horn and Sanders, 63.
16. William L. Sanders and June C. Rivers, *Cumulative and Residual Effects of Teachers on Future Student Achievement*, (Knoxville, Tennessee: University of Tennessee Value-Added Research and Assessment Center, November 1996), 4-5.
17. Sanders and Rivers, 3.
18. William L. Sanders, *Graphical Summary of Educational Findings from the Tennessee Value-added Assessment System (TVAAS)*, (Knoxville, Tennessee: University of Tennessee Value-Added Research and Assessment Center, 1997), 26-38.
19. Sanders, conversation, 18 April 2002.
20. R.A. McLean and William L. Sanders., *Objective component of teacher evaluation: A feasibility study*. (Working Paper No. 199) (Knoxville: University of Tennessee, College of Business Administration, 1984), quoted in William L. Sanders and Sandra P. Horn, "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment," *Journal of Personnel Evaluation in Education*, 8 (1994): 300 .
21. Patricia E. Ceperly and Kip Reel, "The Impetus for the Tennessee Value-Added Accountability System," in *Grading Teachers, Grading Schools*, 134-135.
22. Ceperly and Reel, 135-136.
23. "The Tennessee Value-Added Assessment System," <<http://www.k-12.state.tn.us/assessment/scores.asp>>, 18 October 2001.
24. Tenn. Code Ann. §49-1-602 (2001).
25. Cour, 34-35.
26. James H. Stronge and Pamela D. Tucker, *Teacher Evaluation and Student Achievement* (Washington, DC: National Education Association, 2000), 25.
27. Mickie Anderson, "Accountability? For Schools, Yes; For Teachers, No / Evaluations Not Tied to Tests, Critics Say," *The Commercial Appeal*, 30 November 1998, sec. A.
28. Cour, 36.
29. Cour, 37.
30. Sanders, *Graphical Summary*, 25-37
31. Sanders, conversation, 18 April 2002.
32. Sanders, conversation, 18 April 2002.
33. Marsha Denton, email message to Karen Helland, 15 November 2001.
34. Webster and Mendro, 81.
35. Webster and Mendro, 81-82.
36. Webster and Mendro, 83.
37. William J. Webster, "Dallas Independent School District Accountability System," presentation to the Policy Seminar on Value-Added Assessment, sponsored by Oakland Schools Education Policy Center and The Education Policy Center at Michigan State University, 17 January 2002, 30.
38. Webster, presentation, 17 January 2002, 31.
39. Webster and Mendro, 83.
40. "School Performance Improvement Awards 1999-2000," (Dallas, TX: Dallas Public Schools, February 2000), 7-8.
41. Webster and Mendro, 88-89.
42. William J. Webster, et al., "Little Practical Diference and Pie in the Sky: A Response to Thum and Bryk and a Rejoinder to Sykes," in *Grading Teachers, Grading Schools*, 128-129.
43. Yeow Meng Thum and Anthony S. Bryk, "Value Added Productivity Indicators: The Dallas System," in *Grading Teachers, Grading Schools*, 102.

44. Webster and Mendro, 82.
45. "Comparable Performance And Beyond 2001-2002," (Dallas, Texas: Dallas Independent School District, 2001), 5.
46. Webster and Mendro, 91.
47. Webster, presentation 17 January 2002, 44.
48. "History of the ABCs Program," *NCPublicSchools.org*, <<http://www.ncpublicschools.org/abcs/ABCsHist.html>>.
49. "History of the ABCs Program," *NCPublicSchools.org*, <<http://www.ncpublicschools.org/abcs/ABCsHist.html>>.
50. "State Assistance Teams," *NCPublicSchools.org*, <<http://www.ncpublicschools.org/school-improvement/assistance-index.html>>, 20 May 2002.
51. *The ABC's of public education: 2000-01 background packet*, (Raleigh, NC: NC Department of Public Instruction, 2001), 11.
52. *A Report Card for the ABCs of Public Education Volume I: 2000-2001 Growth and Performance of Public Schools in North Carolina*, (Raleigh, NC: NC Department of Public Instruction, updated December 2001), vii.
53. "History of the ABCs Program," *NCPublicSchools.org*, <<http://www.ncpublicschools.org/abcs/ABCsHist.html>>.
54. Gary Williamson, NC Department of Public Instruction Accountability Chief, telephone conversation with Karen Helland, 20 May 2002.
55. Division of Accountability Services, *Setting Annual Growth Standards: "The Formula,"* Accountability Brief Vol. 1, No. 2 (Raleigh, North Carolina: NC Department of Public Instruction, June 2000), 1-3.
56. *The ABCs Accountability Model, Determining Composite Scores 2001-2002*, (Raleigh, North Carolina: NC Department of Public Instruction, 2002), 2.
57. *ABCs Accountability Model*, 4-5.
58. *ABCs Accountability Model*, 7.
59. *ABCs Accountability Model*, 9.
60. Texas Educ Code, §39.051(c)
61. Department of Accountability Reporting and Research, *2002 Accountability Manual*, (Dallas, Texas: Texas Education Agency, April 2002), 51.
62. *2002 Accountability Manual*, 142.
63. *2002 Accountability Manual*, 59-60.
64. David Garcia and Anabel Aportela, *A First Look at Growth in Arizona Schools Technical Document*, (Phoenix: Arizona Department of Education, February 2000), 1.
65. Ariz. Rev. Stat. §15-241, Ariz. Rev. Stat. §15-918.02.
66. Garcia and Aportela, 9.
67. Garcia and Aportela, 2.
68. Garcia and Aportela, 4.
69. Garcia and Aportela, 7.
70. "Analysis of the Arizona Measure of Academic Progress," (Phoenix: Arizona Department of Education, 2001).
71. "Analysis of the Arizona Measure of Academic Progress," Appendix A.