



Value-Added Assessment from Student Achievement Data: Opportunities and Hurdles

CREATE NATIONAL EVALUATION INSTITUTE July 21, 2000

WILLIAM L. SANDERS
SASinSchool© SAS Institute, Inc. Cary, NC

Abstract

Let me share with you how honored I am to receive an award named after the late Dr. Jay Millman. In 1983, after completing the first of our research studies that began our continuing work in value-added, our report was sent by officials in the Tennessee Department of Education for review by Dr. Millman. It is no secret that many in the Department at the time were assuming that his anticipated critical review would put an end to such a preposterous idea—that student achievement data could be used as part of teacher evaluation. Days turned into weeks; each time that I would inquire of the Department as to when we would hear from the review, I was always told that they had not received it. One day I called Dr. Millman and explained my frustration of not hearing from the review and inquired as to when it might be available. He immediately interrupted and explained that he had sent the review several weeks previous to that day and that he would be glad to send me a copy of his remarks, obviously very angry that they had not been passed on to me. Upon receiving and reading his review, it became obvious why I had not received a copy from the Department. Even though he raised many important questions, his review was most objective and generally very positive. Later, he asked us to submit chapters to the book on student outcomes assessment models that he edited. In all of my interactions with Jay, I developed the utmost respect for this distinguished scholar, and I am glad that fate let our paths cross.

In the last years of the twentieth century, education was called to account for the failures of large segments of the population of American students to achieve the minimal academic expectations of society as a whole. The beginnings of the twenty-first century find educators struggling with the questions society is asking: How can the level of academic achievement of all students be raised, and how can the responsibility for their success or failure be appropriately attributed to schools, districts, teachers, and the students themselves? There continue to be very divergent views as to how these questions should be addressed and the appropriate pathways to their answers. In my presentation today, I will offer a synopsis of attempts to address these questions in the field of educational assessment through the development of various methods of analyzing student test data. I will examine the criticisms of value-added assessment in particular, and of all assessment that relies on student test data, in general, and I will provide a brief description of the model of educational assessment developed by myself and my colleagues to meet these criticisms. Finally, I will talk about equity and how this critical issue can and should be addressed in the service of improving academic attainment for all students, regardless of ethnicity or socio-economic status.

Few would disagree that today in the United States there is more emphasis on improving academic achievement for all students than ever before. Across the entire political spectrum, there has emerged a reasonable consensus that the improvement of academic achievement is a shared goal. Virtually all states have initiated processes to that end, although approaches differ among them. Even within states, there are drastic differences in opinion as to how such improvement should be accomplished, often resulting in heated debates in philosophical, strategic and tactical tones. For example, some would argue that the way to improve academic achievement for all students is to appropriate more resources and free educators to utilize their own professionalism. Others, with equal fervor, will argue that the route to academic improvement for all students is to set very high standards in the belief that, if held accountable for them, the educational community will find ways for students to reach those standards.

In the first case, the evidence has accumulated that a *laissez faire* approach to public education, which *de facto* has been practiced in this country for decades, has resulted in huge variability in effectiveness among school districts and schools, even when socio-economic differences in communities has been accounted for (Graphical Summary, 1997). Huge sums of money have been delivered to the educational community by federal, state and local governments and private philanthropic foundations to support numerous initiatives and programs whose purpose was to improve student academic achievement. The results from these investments have been mixed, at best (Hanushek, 1997a and 1997b). The perceived failure of the *laissez faire* approach has led to competing strong voices from outside the traditional educational community, not only demanding improvement, but offering proposed solutions.

Vouchers, privately funded scholarships, charter schools, choice among public schools, and home schooling are some of the articulated and implemented alternatives to traditional *laissez faire* publicly supported educational efforts. However, the one effort originated and cultured from outside the traditional educational community that has gathered the most momentum is the standards movement.

The standards movement, as operationalized currently, embodies the concept of a stair-step approach to curricula and its companion, assessment. “What should fourth graders know and be able to do?” is a question defining specific goals for many states and districts—“these are first grade skills; these are second grade skills, etc.” As a working definition, this has led to testing regimes that purport to measure the percentage of students within grades who are at mastery, proficient, basic, non-mastery (or whatever language is dangled beside the test results). Inevitably, when the results of these tests are presented, it becomes obvious that differences in results among schools and districts are strongly related to socio-economic measures of the demographics of the student population of a school or district.

If consequences have been attached to test results (i.e., retention of students in grade, mandatory summer school, etc.), political pressure often builds to either lower the height of the “step”, to diminish the role of the tests, or to eliminate them all together. Eventually and invariably, these arguments distill to debates and disagreements over how academic achievement is to be measured and the proper role of standardized testing of students within the total framework of student assessment. However, there is another approach to standards, based upon value-added assessment, that eliminates much of the debate.

Value-Added Assessment Eliminates Much of the Debate

If a curriculum is viewed as a ramp—not as stair-steps—and if the desire is for each student to move up the same ramp (especially in elementary and middle school), and if it is further recognized that all students will not be at the same place at the same time in the same grade, then many problems in assessment and measurement can be mitigated. Our research work indicates that differences in schooling effectiveness is the dominant factor affecting the speed that students move up the “ramp”.

An accountability system that is based upon the rate of academic progress of populations of students is one that will hold people accountable for things over which they have control, rather than for things over which they do not. For instance, teachers in the fall have no control over the achievement level of their incoming students. However, teachers do have primary control of the rate of academic progress of their students.

Rather than attempting to have all students reach a designated achievement level at a specific time, as is the case in most of the present implementations, if standards are defined in terms of rates of academic progress, then much of the debate will be defused because it has been demonstrated, by our work and that of others, that rates of academic progress can be estimated nearly, if not entirely, free of socio-economic and ethnic confoundings (Graphical Summary, 1995; Graphical Summary, 1997; Darlington, 1997; Sanders, Saxton & Horn, 1997). However, even if standards are defined this way, still remaining are many questions as to what data should be included in the estimation of academic progress, what analytical procedures should be used, and to what level of the educational hierarchy can estimation of measures of academic progress be carried (state, district, school, or teacher).

What Data Can Be Used?

If achievement test data exist for each student each year, and if the scales are highly correlated with curricular objectives, and if the scales have sufficient stretch to measure progress of both previously low and high scoring students, and if the scales have the appropriate repeatabilities, then all data that meet these conditions can be used in a value-added assessment system, regardless of test source. If a data structure is available that meets these conditions, analytical procedures exist that will allow a multivariate longitudinal analysis that exploits the total information available in the array. Some of the immediate advantages are as follows:

1. The tests do not have to be nearly so closely aligned as they do when judgements are made from a single year of test results each year.
2. Errors of measurement are greatly reduced because the non-zero covariance structure that exists among test subjects (i.e., math, reading, etc.) and over time is exploited.

At this time, most states and many districts do not have the historical data that meet these conditions. What has emerged as a result, based on the limited data available, are

different approaches to providing summative information for educational accountability, many of which could be classified as value-added approaches.

Various Value-Added Approaches

Various value-added assessment approaches have been developed, but not all of these approaches will yield equivalent results. Some of the major differences will be described later, but first it must be stated that any of these approaches is superior to the reporting of simple raw test averages. The worst possible use of test data for public reporting is the presentation of simple test averages by districts and schools! It is well known and well documented that these simple averages are so confounded with socio-economic factors outside of the control of schools that any sensible interpretation of these reports as to the effectiveness of schools is impossible (Adcock, 1995; Wang, Haertel, & Walberg, 1993). Students within a school, serving primarily a low socio-economic community, could be making wonderful academic progress, yet their average test scores could be considerably lower than the district's average, leaving the erroneous impression that this is a woefully ineffective school. Students from another school, serving a population from more advantaged homes, could be "sliding" and "gliding," nevertheless leaving the naïve impression that this second school is "better" than the first because its average test scores are higher than the first school.

The obvious inappropriateness of the reporting of simple averages has led to other slightly more sophisticated attempts to dampen this unfairness. The first attempt often employed is to disaggregate the simple averages and report them by various socio-economic strata, often with an accompanying cluster of schools within each stratum. Even though this is a slight improvement over the presentation of simple raw averages, meaningful comparisons are not directly available because the stratification scheme obscures many other dimensions of confounding. For example, if the means are reported by ethnic group, then the influence of the educational attainment of the parents is hidden, etc.

The second most frequently observed attempt to render a more fair presentation of test results, and the first of the value-added approaches to be commented on herein, is to provide results from various regression models using commercially available statistical software. These models often take one of two basic forms. The first model of this type, usually fitted to district or school level data, includes predictor variables that attempt to account for differences in test scores due to various socio-economic factors. The predictors might be a measure of ethnicity of the population, the percentage of students eligible for free/reduced price lunch, mean educational attainment of the community, etc. (Jordan, Mendro, & Weerasinghe, 1997). The second model of this type, applied at the individual student level, purports to accomplish the same as the first model. In this model, various data at the student level may be included, usually including but not restricted to prior achievement test data. Then, in various ways, residuals from these models are used to provide a measure of the effectiveness of schools and districts (Bingham, Heywood, & White, 1991).

The school and district results from these models are a vast improvement over the presentation of raw test score averages and the various stratification approaches. These are often labeled as “value-added” results, signifying that the district and school effects are measured over the effects expected from the various socio-economic predictors. Even though these measures are a major improvement over the aforementioned approaches, I have some major concerns about their use even at the district and school level.

First, by including socio-economic variables as predictors, different expectations can unwittingly be set for students coming from different households even though these students may have the same ability and prior level of achievement. As an example, I was reared in a community that probably had less household income than any district in our county. If household income had been included in a prediction model then a young Bill Sanders and some of his classmates would have been expected to achieve less academically than other, wealthier students with comparable abilities when we attended the county seat high school. Fortunately, that was not the case.

Early high achieving students come from housing projects, remote rural areas, and million dollar homes, even though the proportions of these students across neighborhoods can be quite different. This is why I strongly believe that any of these models should not include socio-economic or ethnic accommodations but should only include measures of previous achievement of individual students, unless there is considerable and compelling evidence that additional predictor variables are needed to insure fairness.

My second concern with the more traditional regression approaches is more technical. Traditional multiple regression approaches require complete information on each observational vector (i.e., each kid). But kids move. Kids get sick. Kids miss tests for numerous reasons. Disproportionately, lower scoring kids miss more tests than higher scoring kids do. Thus, the data fed into these analyses is often a truncated sample of the district or school’s student population that often results in an over estimate of student achievement.

However, there is another approach that will provide unbiased measures of district, school and teacher effectiveness. If scales of measure, either from traditional norm-referenced tests or criterion-referenced tests,¹ are available for students over time, and if these scales of measure are highly correlated with curricular objectives, then there exists methodology which will yield the desired measures of schooling effectiveness. The process that we have developed, based upon statistical mixed model theory and methodology, enables a multivariate, longitudinal analysis, no matter how sparse or complete the data record for each student, which will, in turn, eliminate the shortcomings of other models described previously (Sanders & Horn, 1994; Sanders, Saxton, & Horn, 1997).

A conceptual view of this process may be obtained by first imagining that the “dimples” and “bubbles” around each student’s own academic pathway are measured. Then imagine that these deflections are aggregated over students to obtain measures of the district and school effectiveness. By this process, each student serves as his or her own control. We have demonstrated repeatedly that the school and district effects thusly obtained are virtually unrelated to various socio-economic indicators. Yet the undesirable

consequence of setting different expectations of academic growth for different sub-populations of students has been avoided.

Special Problems with Using Student Achievement Data to Estimate Teacher Effectiveness

Our research work, based upon millions of student achievement records, clearly indicates that differences in teacher effectiveness is the single largest factor affecting academic growth of populations of students (Sanders & Rivers, 1996; Jordan, Mendro, & Weerasinghe, 1997; Haycock, 1998). The cumulative and residual effects of teacher effectiveness on student academic achievement are measurable and huge. Notwithstanding claims to the contrary, teacher effects on the academic progress of student populations can be measured with appropriate levels of sensitivity and reliability from longitudinal analyses of student achievement data. However, analyses at the teacher level require the utmost care and caution and present even more burden on the statistical methodology, the computing software, and the data archiving process itself.

At the district and school level, the number of students' records offer some protection against spurious estimates, even when the simpler regression approaches are used. However, when the analyzes are shifted to the classroom with the accompanying smaller number of student records, there is considerable risk that individual teachers will receive false negative reports due either to a small number of student records or some spurious quirk in the data. At a minimum any statistical process that purports to yield estimates of teacher effects must possess the capability to provide best linear unbiased estimates (in some literature referred to as "shrinkage estimates") of the teacher effects (Henderson, 1975). Unlike the estimates obtained from traditional regression approaches, with "shrinkage" estimation, each teacher's effectiveness is assumed to be the average of the population (the district in our applications) until the weight of the data pulls the estimate away from the average. Thus, if the number of student records is small or if the testing regime provides data that are too "noisy", then the worst case is that the process will not distinguish among any individuals. In other words, the process shuts down.

We have added another dimension to the analytical process that offers teachers additional protection against the likelihood of misclassification while increasing the robustness of the estimation. We call this process the "layered" model. Since all of a student's achievement test data (up to five years) are used simultaneously over all subjects tested, we have found that by linking each year's data to the current and previous teachers, additional sensitivity is obtained. To illustrate one advantage obtained from the "layered" model, pretend that a teacher introduces material to students that does not appear until next year's test. With the "layered" model, credit will be appropriately assigned, because each test score affects the estimates for both current and previous teachers. This should alleviate teacher concerns that allowing students to progress beyond a specific grade level's bounds will adversely affect teacher estimates.

Additionally, teachers are protected against drastic changes in student academic trajectories resulting from extreme external factors. For example, consider the case of a

student who had been making appropriate academic progress each year, but who now is performing at levels below expectation because of recent drug involvement/chronic illness, or other problem. Without the layered model, the teacher's value-added estimate would be adversely affected by this precipitous change in circumstance. However, since future student data is linked to the previous teachers, and since a new reference trajectory will be established based upon test scores in subsequent years, each teacher's estimate is held harmless from this radical departure in student performance.

As we have developed this system, we have had to engineer the flexibility to accommodate other "real world" situations: the capability to accommodate different modes of instruction (i.e., self-contained classrooms, team teaching, etc.), "fractured" student records, and data from a diversity of non-vertically scaled tests. The data warehousing and data merging requirements to support this approach are not trivial, and the computer resources necessary to complete these analyses for a state or large district are substantial, since this process requires the iterative solution to many thousands of equations.

Why Bother?

The variability in teacher effectiveness is huge, ranging from the most effective teachers, who facilitate excellent academic gains over the entire distribution of students within the classroom, to teachers whose students make very little, if any, gain during the year. Especially in math, the cumulative and residual effects of teachers are still measurable at least four years after students leave a classroom (Rivers Sanders, 1999; Sanders & Rivers, 1996). If anyone is serious about improving the academic achievement levels for all students, then this improvement will be obtained only by reducing the likelihood that students will be assigned to relatively ineffective teachers.

Teaching ineffectiveness is not necessarily a permanent condition. I submit that most teachers are sincere, dedicated individuals who want to do a credible job in their chosen profession. However, historically, teachers have worked in a vacuum with very little summative or formative feedback as to how their students are progressing relative to other students with comparable levels of prior achievement. We have observed that once a measurement process is in place that offers feedback on the outcomes of instruction at the classroom level, many teachers begin to develop their own strategies for improving areas in which they are deficient. However, this process can be accelerated when the leadership within districts and schools provides the opportunity for individuals to learn to use and interpret the results of value-added assessment in positive diagnostic ways.

As to summative uses, a rigorous value-added approach is the fairest, most objective way to hold districts and schools accountable. At the teacher level, the value-added estimates of teacher effectiveness should be a part of formal teacher evaluation, but they should not be the sole basis upon which teachers are evaluated, because there are too many other duties, dimensions, and responsibilities that cannot be measured by a process such as has been advocated herein.

In Regard to the EVAAS Model of Value-Added Assessment, Critics Have Voiced Several Concerns:

“The process is too complicated; people do not understand it; if the process is to be used for evaluation, people should understand it.” “The process is too black box.”

I have to confess that this criticism both befuddles and agitates me. There has to be a clear distinction between simplicity of conceptual understanding and the complexity of the methodology that is necessary to provide reliable information. Most everyone can use a cellular telephone, but virtually no one knows, or needs to know, how to build the phone. Nor do they have a thorough understanding of how voice is converted into signals and how the signal is delivered from transmitter to receiver. If it were necessary for each user to know how to build the device prior to appropriate use, then all of our phones would be restricted to tin cans and string.

Likewise, the statistical methodology underlying our value-added approach is complex, but the concepts are simple. To understand and use the resulting information for positive diagnostic purposes does take understanding of output, but it is easy to grasp by those who exert only a modest amount of effort. The Tennessee experience suggests that where local leadership has provided the opportunities for teachers and principals to learn to use the reports provided, then cynicism has been replaced by teachers asking why more information can not be supplied more quickly.

Another Criticism of EVAAS and Value-Added Assessment in General is “There is Too Much Reliance on a Single Test”

We use up to five years data for each student. In the Tennessee testing regime, there are five subjects tested each year. Presently each sub-test has 40+ items. Thus, over each student’s observational vector there could be as many as 1,000 items that collectively contribute to the informational array for that student. These items, spanning over years and subjects, represent a sampling over many curricular domains. Contrast that with a writing assessment that provides a snapshot of a student’s current writing skills based upon one writing sample, often based upon only one writing prompt. There is far more information entered into the multivariate longitudinal value-added approach than can be provided by any one test for a specific year.

“This Much Testing Results in a Narrowing and Misshaping of the Curriculum”

In contrast to regimes that do not test each student each year, I believe that annual testing of each student with fresh, non-redundant, equivalent tests to support the value-added approach minimizes distortions in the curriculum and “teaching to the test” concerns. If teachers cannot anticipate specific items on tests, and after it becomes apparent that it is in their selfish best interest to teach each student along the curricular pathway, then more

focus will be on teaching students “from where they are” and less focus on test preparation *per se*. However, in contrast, we have observed in data from states that are testing specific subjects infrequently and only in specific grades, that major distortions can be seen. For example, we observed in one elementary school that the 4th grade scale score gain in Reading was outstanding, yet the 4th grade gain in Math was near zero (based upon traditional achievement test results). In 5th grade the Math gains were very large, while the Reading gains were near zero. Subsequently, we learned that in 4th grade the Reading time had been lengthened significantly at the expense of Math; in 5th grade the Math time had been lengthened and the Reading period virtually eliminated. No surprise, since in this state, the statewide tests in Reading are in the 4th grade and the statewide tests in Math are in 5th. In states where testing is administered over all major subject areas yearly, this type of gerrymandering serves no purpose and is therefore less likely to occur.

Summation

So now that we have measurement methodologies that can fairly estimate the effects of schools and teachers on the academic growth of students, what good is it to education as a whole and our students, in particular? Well, none, unless it is used.

If appropriate levels of academic gain are sustained for each student each year, then the achievement level for each student will be ratcheted to higher levels. What is important is NOT the achievement level of third graders, for instance, which is basically all that could be determined with previous schemes of reporting test averages. What is most important is the achievement levels of 11th and 12th graders, but I am not smart enough to envision how the achievement levels of ALL students can be raised without a total focus on sustained academic growth—and we cannot know how much—or how little—growth is occurring without reliable assessment based upon student outcomes data.

Our work indicates that the biggest impediment to ever higher achievement is the years in which individual students are not making realistic growth. Especially in inner city schools, too often it is observed that the previously lower scoring students are being given the opportunity to make reasonable progress, but within the same school the earlier higher achieving students are being held to the same pace and place as their lower achieving peers. When this pattern is repeated over grades, then it becomes a self-fulfilling prophecy that these early high achieving students lose ground. Without yearly feedback from responsible measurement, often teachers and principals do not recognize that these hurtful patterns exist. However, we certainly know of cases in which teachers, after being presented with the results from the data, have engineered for themselves strategies within their classrooms that have made instruction more equitable—addressing the needs of all students, rather than just a few.

Measurement alone cannot bring about the changes that will lead to greater equity for students and better outcomes from our schools (Rivers & Sanders, 2000). Teachers, principals, supervisors, superintendents, school boards, and commissioners of education must be taught how to use the information from assessment to the betterment of their curricula, instructional strategies, and educational programs as a whole. The whole mind-

set in regard to educational evaluation, in the classroom and in teacher preparation programs, must become focused on improving the instruction at the classroom level if educational assessment is to make any real, positive difference to our students.

In the past, and even now in many places, when the determination of whether a school was doing a good job or a bad job was based upon how high its students scored on standardized tests, this so-called evaluation and the tests themselves were reviled by the educational community as a whole—and rightfully so. This history is the major impediment to the use of educational measurement data today. But now we are using test data more responsibly. We have added a tremendous amount of sophistication to our analyzes. Now we can pinpoint which achievement level of kids in a particular teacher's classroom are doing very well and which ones are not doing so well, and we can furnish this information to the practitioner. This is information that teachers, principals, and other educational decision-makers need, if they are to do the best they can for every student in their schools.

It is time to teach teachers—and students in our teacher preparation programs—that data is not the enemy—that data is a valuable tool in their educational toolbox. With this tool, they can fine tune their instruction to provide the best opportunity for every child in their classroom to achieve his or her potential. With this tool, we can build the education of the future—individualized, equitable, and full of promise for all our kids.

Notes

1. This assumes that the criterion-referenced tests are scaled and that sufficient stretch exists within the tests to allow a measurement of the progress of previously high and low scoring students.

References

- Adcock, E.P. (1995). *Value-Added Effective Schools Study for Elementary Schools: 1994 Maryland School Performance Assessment Program Results*. Research Report No. 36-9-95. Maryland: Prince George's Country Public Schools, Research, Evaluation and Accountability.
- Bingham, R.D., Heywood, J.S., & White, S.B. (1991). Evaluating Schools and Teachers Based on Student Performance. *Evaluation Review*, 15, 191–218.
- Darlington, R.B. (1997). The Tennessee Value-Added Assessment System, a Challenge to Familiar Assessment Methods. In Millman, J. (Ed). *Grading Teachers, Grading Schools*. Thousand Oaks, CA: Corwin Press.
- Graphical Summary of Educational Findings from the Tennessee Value-Added Assessment System (TVAAS), 1995*. (1995). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.
- Graphical Summary of Educational Findings from the Tennessee Value-Added Assessment System (TVAAS), 1997*. (1997). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.
- Hanushek, E.A. (1997a). Assessing the Effects of School Resources on Student Performance: An Update. *Educational Evaluation and Policy Analysis*, 19(2), 141–164.
- Hanushek, E.A. (1997b). Incentives are Key to Improved Schools. *Forum for Applied Research and Public Policy*, 12(3), 62–67.
- Haycock, Katy. (1998). Good Teaching Matters . . . a Lot. *Thinking K-16*, 3(2), 3–14.

- Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction Under a Selection Model. *Biometrics*, 31(2), 423–447.
- Jordan, H.R., Mendro, R.L., & Weerasinghe, D. (1997). *Teacher Effects on Longitudinal Student Achievement: A Preliminary Report on Research on Teacher Effectiveness*. Paper presented at the National Evaluation Institute, Indianapolis, IN.
- Rivers, J.C., & Sanders, W.L. (2000, May). *Teacher Quality and Equity in Educational Opportunity: Findings and Policy Implications*. Paper presented at the Hoover/PRI Teacher Quality Conference, Stanford University, Palo Alto, CA.
- Rivers Sanders, J.C. (1999). *The Impact of Teacher Effect on Student Math Competency Achievement*. Unpublished doctoral dissertation, University of Tennessee, Knoxville.
- Sanders, W.L., & Horn, S.P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment. *Journal of Personnel Evaluation in Education*, 8(1), 299–311.
- Sanders, W.L., & Rivers, J.C. (1996). *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement*. Research Progress Report. Knoxville: University of Tennessee Value-Added Research and Assessment Center.
- Sanders, W.L., Saxton, A.M., & Horn, S.P. (1997). The Tennessee Value-Added Assessment System (TVAAS): A Quantitative, Outcomes-based Approach to Educational Assessment. In Millman, J. (Ed). *Grading Teachers, Grading Schools*. Thousand Oaks, CA: Corwin Press.
- Wang, M.C., Haertel, G., & Walberg, H.J. (1993). Toward a Knowledge Base of School Learning. *Review of Educational Research*, 73(3), 249–294.
- Wright, S.P., Horn, S.P., & Sanders, W.L. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education*, 11(1), 57–67.