Teacher Effects as a Measure of Teacher Effectiveness: Construct Validity Considerations in TVAAS (Tennessee Value-Added Assessment System)

CSE Technical Report 563

Haggai Kupermintz
CRESST / University of Colorado, Boulder

May 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Comparative Analyses of Current Assessment and Accountability Systems
Robert L. Linn, Project Director, CRESST/University of Colorado at Boulder

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

TEACHER EFFECTS AS A MEASURE OF TEACHER EFFECTIVENESS: CONSTRUCT VALIDITY CONSIDERATIONS IN THE TVAAS (TENNESSEE VALUE-ADDED ASSESSMENT SYSTEM)[1]

Haggai Kupermintz
CRESST/University of Colorado, Boulder

Abstract

This report examines the validity of measures of teacher effectiveness from the Tennessee Value-Added Assessment System (TVAAS). Specifically, the report considers the following claims regarding teacher effects: that they adequately capture teachers' unique contributions to student learning; that they reflect adequate standards of excellence for comparing teachers; that they provide useful diagnostic information to guide instructional practice; and that student test scores adequately capture desired outcomes of teaching. Our analyses of the TVAAS model highlight potential weaknesses and identify gaps in the current record of empirical evidence bearing on its validity.

The Tennessee Value-Added Assessment System (TVAAS) is a statistical methodology designed to evaluate the influence of school systems, schools, and individual teachers on student learning. It is arguably the most prominent example of the "value-added" approach in state accountability systems. The statistical machinery behind TVAAS, developed by Dr. William Sanders at the University of Tennessee, implements a mixed-effects model, applied to longitudinal standardized test score data across several subject areas, to estimate the effects of schools and individual teachers on student achievement progress. Estimates of teacher effects claim to be objective, fair, dependable, and accurate indicators of teacher effectiveness. Moreover, these estimates are supposedly independent of potential competing determinants of student learning, most notably race, SES, general ability, and prior achievement in the tested subjects. This paper examines these claims from a construct validation perspective.

An Overview of TVAAS

TVAAS is the centerpiece of an ambitious educational reform effort implemented by the Tennessee Education Improvement Act (Education

[1] I wish to thank Bob Linn and Lorrie Shepard for helpful comments, and Kelly Alanis and Andrew Brodsky for their assistance.

Improvement Act, 1992). Inequalities in school funding, followed by a lawsuit brought against the state by a coalition of small rural districts, have led to a comprehensive reform of the Tennessee educational system. Under pressure from business, the legislature adopted a strong accountability model that required concrete evidence for satisfactory year-to-year improvements down to the classroom level. Based on encouraging pilot studies with the value-added model conducted by Sanders and his colleagues during the 1980s, the Tennessee legislature has embraced the model as the methodology of choice to generate the desired evidence on the performance of students, teachers, schools, and school systems. The legislation describes TVAAS as follows:

> (1) A statistical system for educational outcome assessment which uses measures of student learning to enable the estimation of teacher, school, and school district statistical distributions; and
>
> (2) The statistical system will use available and appropriate data as input to account for differences in prior student attainment, such that the impact which the teacher, school and school district have on the educational progress of students may be estimated on a student attainment constant basis. The impact which a teacher, school, or school district has on the progress, or lack of progress, in educational advancement or learning of a student is referred to hereafter as the “effect” of the teacher, school, or school district on the educational progress of students.
>
> (b) The statistical system shall have the capability of providing mixed model methodologies which provide for best linear unbiased prediction for the teacher, school and school district effects on the educational progress of students. It must have the capability of adequately providing these estimates for the traditional classroom (one (1) teacher teaching multiple subjects to the same group of students), as well as team taught groups of students or other teaching situations, as appropriate.
>
> (c) The metrics chosen to measure student learning must be linear scales covering the total range of topics covered in the approved curriculum to minimize ceiling and floor effects. These metrics should have strong relationship to the core curriculum for the applicable grade level and subject. (Education Improvement Act § 49-1-603 [1992])

For details of the TVAAS methodology and the estimation of system, school, and teacher effects see Sanders, Saxton, and Horn (1997). Using annual data from the norm-referenced tests comprising the Tennessee Comprehensive Assessment Program (TCAP), schools and school systems are expected to demonstrate progress, at the level of the national norm gain, in five academic subjects. Beginning in 1993, reports have been issued to educators and the public on the effectiveness of every school and school system. Teacher reports are not part of the public record; rather, value-added assessment of teacher effectiveness has been provided only to teachers and their administrators. Following is an examination of some aspects of the validity of the TVAAS teacher estimates of effectiveness.

Validity Considerations

Validity is the most fundamental consideration in the evaluation of the uses and interpretations of any assessment. Since validity is specific to particular uses and interpretations, clearly it is not appropriate to make an unqualified statement about an assessment's validity. An assessment that has a high degree of validity for a particular use may have little or no validity if used for a different purpose. For this reason, the Test Standards admonish the developers and users of assessments to start by providing a rationale "for each recommended interpretation and use" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 17).

This paper discusses specific inferences from estimates of teacher effects that have been promoted by TVAAS developers as reflected in the legislation's language, and examines empirical evidence bearing on these inferences. Specifically, it addresses the following questions:

- Do teacher effects adequately capture teachers' unique contributions to student learning?
- Do teacher effects reflect equal standards of excellence for all teachers?
- Do teacher effects reflect desirable or objectionable instructional practices?
- Are student test scores adequate measures of desired outcomes of teaching?

Unique Contribution to Student Learning

Student learning and development of academic proficiencies is a highly complex process, shaped and influenced by a multitude of factors: personal characteristics (both cognitive and non-cognitive), physical and mental maturation, home environment, cultural sensitivities, institutional and informal community resources, and, of course, the formal process of schooling. Even when we confine our attention to schooling alone as a major determinant of student learning, complexity abounds. School culture and climate, teacher qualifications, curriculum frameworks, and instructional approaches all interact jointly to produce measurable growth in student academic skills and knowledge. This complexity and the dynamic and interactive nature of the learning process have consistently defied simple explanations and have created monumental conceptual and methodological challenges for researchers and practitioners who attempted to disentangle and isolate specific, direct effects on student achievement and growth. Two factors seem to be especially prohibitive: (a) the dynamic, interactive nature of the learning process, and (b) the inevitable confounding of many of the formal and informal influences on the process.

The second factor deserves special attention. Because of structural and functional features of the U.S. educational system, learning environments present themselves as "syndromes" or amalgams rather than as additive clusters of independently accrued conditions. Low SES students, for example, in addition to contending with impoverished home environment, typically face inadequate facilities, a less qualified teaching force, diminished curricula and uninspiring instructional methods, and explicit or implicit segregation along racial and ethnic lines. Consequently, these students consistently lag behind their more privileged peers in academic achievement and progress. TVAAS developers have made the bold claim that their system adequately accounts for all the potent influences on learning (thereby allowing the isolation of teacher direct effects) by employing the experimental design principal of "blocking," using each student's prior achievement as the only control or "proxy" for all such influences: "[E]ach child can be thought of as a 'blocking factor' that enables the estimation of school system, school, and teacher effects free of the socio-economic confoundings that historically have rendered unfair any attempt to compare districts and schools based on the inappropriate comparison of group means" (Sanders et al., 1997, p. 138).

In the design and analysis of controlled experiments, blocking is an extremely powerful tool for partialling out "contaminating" variability to improve the precision of estimation of treatment effects. Such benefits are realized through careful design and deployment of blocking factors, using well-established routines for randomization and balancing, without which causal inferences regarding treatment effects become highly suspect. Unfortunately, uncontrolled observational studies can never hope to ensure an adequate blocking regime. Consequently, the TVAAS strategy of using students' prior achievement as a sole blocking factor raises two serious concerns.

Incomplete Control

First, it is unclear to what extent prior achievement captures all the important confounders that ought to be controlled for. Typically, various proxy indicators poorly measure variables like socioeconomic status, home environment, and others mentioned above as potentially important in promoting student learning. In addition, such factors are correlated only weakly or moderately with prior student achievement (especially when only linear relationships are considered). As a result, important influences on learning may remain unaccounted for, leading to potentially biased results. While the TVAAS model can be expanded to accommodate more covariates, this has been deemed unnecessary based on Sanders' team's secondary, ex post facto analyses that showed that school effects are uncorrelated with variables such as the percentage of students receiving free and reduced lunches in the school, the racial composition of the student body, the location of the building (urban, suburban or rural), or the mean achievement level of the school.

Unfortunately the technical and substantive specifications of these analyses have never been published (except as general descriptions of results; see, e.g., Sanders & Horn, 1998), making it hard to evaluate the above conclusions. Such details are important because TVAAS calculates system, school, and teacher effects separately in each school system. A multi-level analysis, for example, may reveal different within- and between-system patterns for the above correlations. In addition, in a recent study using data from 58 elementary schools, Hu (2000) documented a correlation of .39 for per pupil expenditure and average TVAAS value-added scores in both math and reading. Percent minority was correlated .42 with math and .28 with reading (the corresponding correlations for percent of reduced-price/free lunch were .49 and .27, respectively). Taken together, these variables explained between 19% and 28% of the variability in the value-added 3-year averages. Hu's findings, therefore, argue against the TVAAS claim of sufficient control afforded by taking into account only prior achievement.

Block-Treatment Confounding

The second and more serious potential limitation of using student prior achievement as a blocking factor in the TVAAS model is the potential confounding of student achievement and teacher effectiveness. The usefulness of blocking depends on random assignment or careful systematic allocation of treatment conditions among the experimental blocks. Therefore, in the educational data analyzed by TVAAS, teacher effectiveness (treatment) should be at least statistically independent from student prior achievement (block). Figure 1 presents data from a study that examined the relationships between teacher effectiveness and fifth-grade achievement in math (Sanders & Rivers, 1996). For each prior achievement student group, it shows the proportions of least and most effective teachers assigned to these students.

In the lowest prior achievement groups, slightly more than 10% of the students were assigned to highly effective teachers, while almost 30% were assigned to the least effective teachers. In contrast, in the highest prior achievement group, slightly more than 5% of the students were assigned to ineffective teachers and more than half were assigned to highly effective teachers. It is unclear whether these results reflect systematic inequalities in the allocation of teachers to students or a possible misattribution of teacher effects. In either case, these patterns suggest that the manner in which TVAAS accounts for exogenous influences on student learning runs the risk of introducing systematic biases in the estimation of the magnitude of the contribution to student learning directly attributable to teachers.

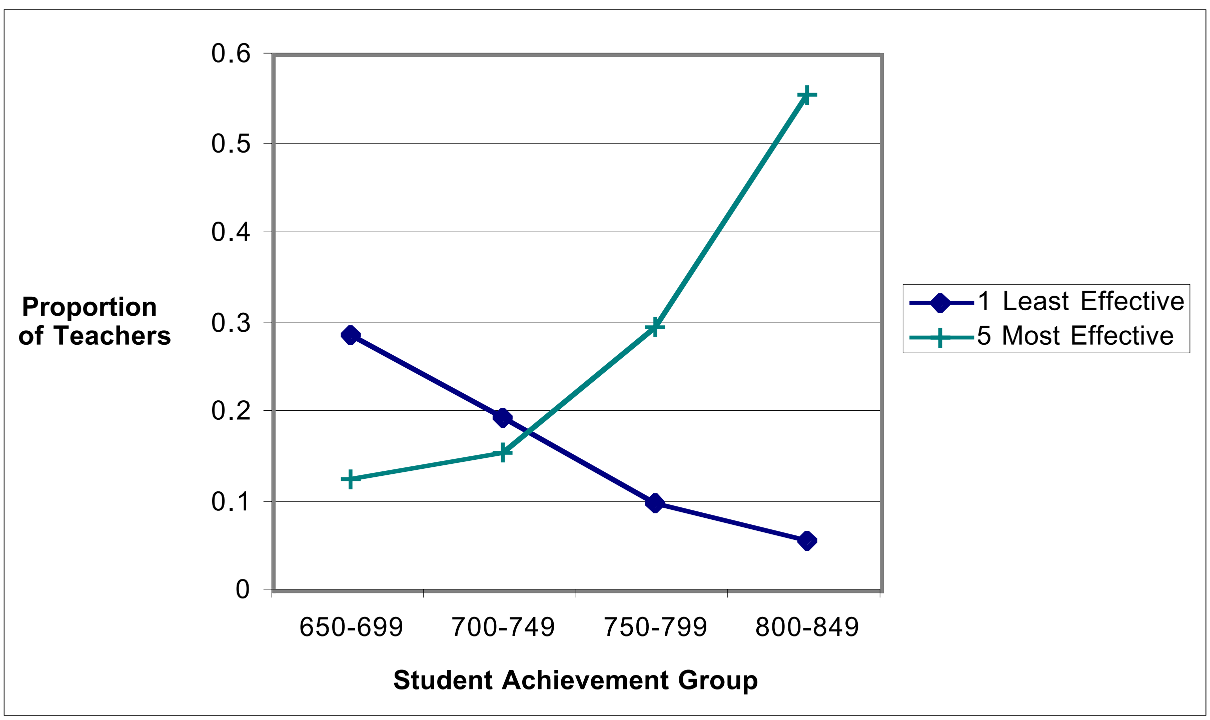

Figure 1. Teacher effectiveness by student achievement.

The following results from a small-scale simulation demonstrate the impact on teacher effects of artificially confounding teacher true effects and the average independent gains of their students. SAS Proc MIXED was used to obtain estimates of teacher effects similar to those produced by the full TVAAS model for different configurations of student and teacher contributions to gains in test scores. In this simulation, student and teacher true contributions are independent of each other. Table 1 shows the results for four hypothetical teachers, each with five students, under three different simulation conditions. Overall gain is the summation of student and teacher true effects (plus a small amount of random noise), and teacher estimates show the effects attributed by the model to teachers.

In simulation I, teacher true contributions to gains are all zero, yet the estimates of teacher effects are non-zero and reflect the relative contributions of their students. Simulation II demonstrates that when effective teachers are systematically assigned weak students and vice versa, teacher and student contributions operate in different directions to produce null estimates for teachers (these results reflect the fact that teacher effects sum up to zero in the model; more on this feature of the model will be covered later). Simulation III again shows that student independent contributions to

Table 1

Teacher Estimates as a Function of Student and Teacher Effects

		True effects		
	Gain	Student	Teacher	Teacher estimate
Simulation I				
Teacher 1	5.5	5	0	-5.17
Teacher 2	5.7	5	0	-4.97
Teacher 3	15.5	15	0	5.07
Teacher 4	15.5	15	0	5.07
Simulation II				
Teacher 1	20.8	5	15	0.04
Teacher 2	20.4	5	15	-0.03
Teacher 3	20.5	15	5	-0.01
Teacher 4	20.5	15	5	0.00
Simulation III				
Teacher 1	25.4	5	20	1.02
Teacher 2	20.3	5	15	-1.68
Teacher 3	25.6	15	10	1.70
Teacher 4	20.5	15	5	-1.04

gains may distort the estimates of teacher contributions. It must be noted that these demonstrations are highly artificial and do not represent adequately the TVAAS model; yet, they are instructive in dramatizing the potential biases in teacher estimates due to systematic confounding of independent teacher and student contributions to score gains.

Examining the correlations between students' average score levels and their average gains in a sample of the Tennessee data, Bock and Wolfe (1996) have commented:

> Although the magnitude of all of the correlations is less than 0.3, a good number of them are large enough to have implications for the comparison of gains between teachers whose students differ in average achievement level . . . [A]djustments for expected gain as a function of student score level should be included when the magnitude of the correlation exceeds, say 0.15. (p. 27)

Standards of Excellence

When statistical estimates become a part of the procedure for summative evaluation of teachers, fairness is a key consideration. In the TVAAS model, teacher effects are "shrunken" estimates: When not enough student data are available, a teacher is assumed to perform at the level of his or her school system mean. The fewer students a teacher has, the stronger the pull toward the overall system mean. "A very important consequence is that it is nearly impossible for individual teachers with small quantities of student data to have estimates measurably different from their system means" (Sanders et al., 1997, p. 143).

An equally important consequence of this estimation approach is that the model treats individual teachers and schools unevenly. For example, an outstanding teacher with complete data will be identified as outstanding, whereas an equally remarkable teacher with more transient students would not be identified as exemplary. In contrast, a poor teacher whose students are transient would be saved from detection by unreliability in the data. Another implication of this strategy is that teachers in different school systems will be pulled toward different means. Equally effective teachers with the same amount of data will be judged differently when average performance in their respective school systems differs. While anecdotal results have been brought to bear on this issue, no systematic study has examined the rates of false positive and false negative classifications associated with the application of shrunken estimates to teacher effects. Darling-Hammond (1997) has summarized pointedly: “No person should be evaluated for high-stakes decisions based on statistical assumptions rather than on actual information” (p. 255). Yet, when not enough data are available, statistical assumptions underlying the use of shrunken estimates in TVAAS govern the evaluation of teacher effectiveness.

In addition to the sensitivity of teacher estimates to their school system context (via the system’s average performance), the accuracy of these estimates varies as a function of the amount of available data. Teachers with less student data are evaluated with less precision. The degree of uncertainty in teacher effects is expressed by the magnitude of the estimates’ standard errors. Bock and Wolfe (1996) have recommended that teacher estimates should be reported in ways that make the magnitude of the standard errors evident, for example, by graphical displays that show confidence intervals for the teacher gains. This would make it obvious—as shown in the example Bock and Wolfe provide in their report (p. 66)—that some teachers with gains in the middle range may be indistinguishable from other teachers with gains in the high or low categories.

A more subtle and potentially harmful problem may also exist. An important assumption of the mixed-model methodology as implemented in the TVAAS model is that random effects are normally distributed around a zero mean. The implication is that the estimation of teacher effects is a “zero-sum game.” Thus, the estimate of each individual teacher critically depends on the performance of all other teachers in the school system. The assumption of a symmetric distribution of teacher effects within each school system is questionable at best. Moreover, it ignores an entire line of research documenting strong contextual effects operating at the collective rather than the individual teacher level (e.g., Talbert & McLaughlin, 1993). It is also interesting to note that while the prevailing accountability message to students is “every child can and should succeed,” the peculiarities of the statistical model preclude this eventuality when teachers are concerned. The fact that the estimation of teacher effects is carried out separately in each school system may exacerbate the problem and render problematic the comparison of teacher effects across school systems.

Effectiveness and Instructional Practices

The definition of teacher effectiveness exclusively in terms of student gains on standardized tests makes the TVAAS model a black box mechanism. It does not offer any insight into what makes a teacher successful in promoting or hindering students' learning. Sanders and Horn (1995) have argued that this nonprescriptive approach is in fact advantageous:

> Assessment should be a tool for educational improvement, providing information that allows educators to determine which practices result in desired outcomes and which do not. TVAAS is an outcomes-based assessment system. By focusing on outcomes rather than the processes by which they are achieved, teachers and schools are free to use whatever methods prove practical in achieving student academic progress. TVAAS does not assume a "perfect teacher" or a "best way to teach." Rather, the assumption is that effective teaching, whatever form it assumes, will lead to student gains.

In contrast to Sanders and Horn's (1995) neutrality, a great deal of attention has been directed lately to identifying the prominent characteristics of quality teaching (see, e.g., Darling-Hammond, 2000; Wenglinsky, 2000). The TVAAS model's narrow and mechanistic definition of effectiveness may in fact discourage efforts to establish strong research-based programs for improving teaching practices. By equating teacher effectiveness with student performance gains, educators and policymakers may be misled because of the tautological nature of such a definition. The risk is that the origin of the definition will be forgotten and teacher effects will be treated as if they were independent indicators of effectiveness, a possibility considered next.

A widely cited conclusion from the Sanders and Rivers (1996) study states: "Based upon these results, students benefiting from regular yearly assignment to more effective teachers (even if by chance) have an extreme advantage in terms of attaining higher levels of achievement" (p. 7). Sanders and Rivers have reached their conclusion after examining the consequences for student performance of teacher assignments over a 3-year period, showing dramatic differences in performance for students who were consistently assigned during that period to effective or ineffective teachers. But these results are insightful only if we ignore the fact that teacher effectiveness is *defined* in terms of their students' performance gains. Figure 2 demonstrates that the patterns observed in the longitudinal analysis (spanning 3 years) can be predicted by examining the distribution of teacher effects in the baseline year alone.

Sanders and Rivers (1996) have divided the distribution of teacher effects in the baseline year into quintiles to form five effectiveness groups. From their Table 1 it is possible to calculate the average teacher effect in each group—that is, the average

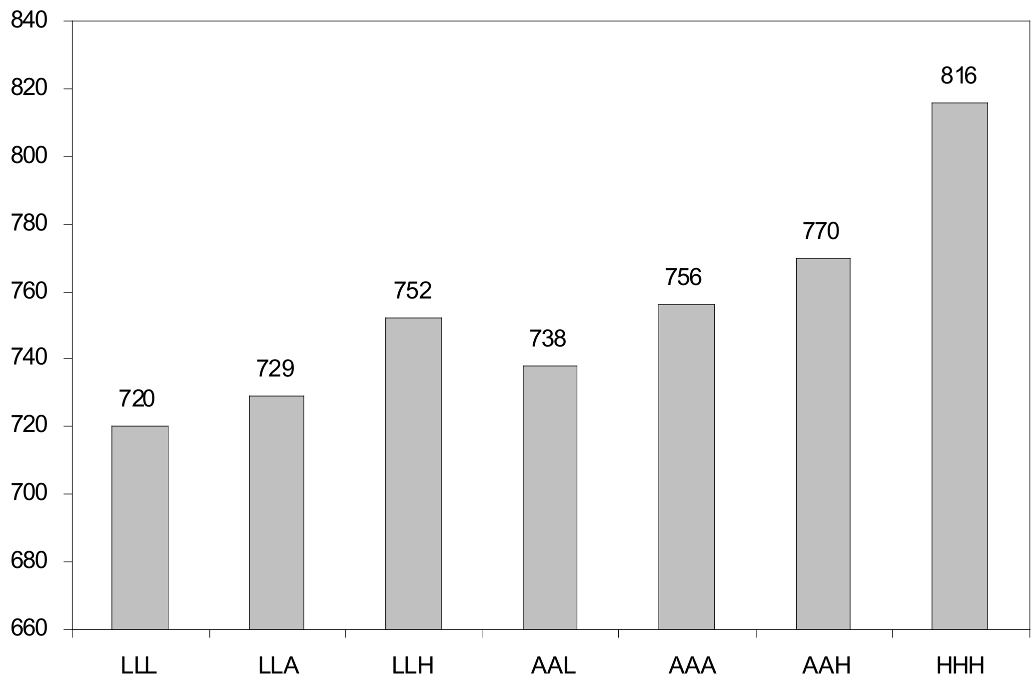

Figure 2. Cumulative teacher effects.

student achievement attributed to each particular teacher—to show that teachers in the middle quintile group have students who gain on average about 9 points more than students of teachers in the low quintile group. Similarly, we find a differential of 32 points between the typical performance of students of teachers in the highest and lowest quintile groups. If we assume that these differentials are consistent across years, we can forecast the terminal expected score for students with different sequences of teachers in a 3-year period. Figure 2 presents such predictions for the sequences shown in Sanders and River's Figure 1. The resemblance of their empirical results to our forecasts is clear. It is arguable, therefore, that 3-year cumulative effects are a reflection of the sum of the effects estimated for high-, medium-, and low-effective teachers in the baseline year. Students of teachers who are defined as effective based on their students' elevated gains indeed do gain more. Stronger interpretations run the risk of overstating the case by dramatizing the inherent tautology of teacher effectiveness defined in terms of student score gains, and inserting a distorted causal interpretation of the pattern of cumulative effects.

Use of Standardized Test Scores

Much has been written about the usefulness and limitations of standardized test scores. Despite heroic efforts to diversify the arsenal of large-scale educational assessment instruments (most notably in California and Kentucky) in the 1990s, most statewide testing programs currently rely primarily on conventional multiple-choice tests. Low cost, ease and consistency of scoring, and a mature industry of testing companies offering a comprehensive menu of services for administering, processing, scoring, analyzing, and reporting test results ensure the privileged status of multiple-choice tests.

According to Sanders and Horn (1995), "any reliable linear measure of academic growth with a strong relationship to the curriculum could be used as input into the [TVAAS] process." "Strong relationship to the curriculum" is taken to mean that the assessment instrument is *aligned* with the curriculum underlying teaching and learning, as explicitly expressed in state and local content standards that specify what students should know and be able to do. The evaluation of the alignment of tests with content standards is often much too superficial. If asked whether their tests are aligned with the content standards of a state, any test publisher will give an affirmative answer. But the answer is unlikely to stand up to close scrutiny. No test or assessment is likely to cover the full domain of a set of content standards. Even those aspects covered will vary in degree and depth of coverage. Hence, an adequate evaluation of alignment must make it clear which aspects of the content standards are left uncovered by the test, which are covered only lightly, and which receive the greatest emphasis. Such an analysis provides a basis for judging the degree to which generalizations from the assessment to the broader domain of the content standards are defensible. If only aspects of the domain that are relatively easy to measure will be assessed, a narrowing of and distortion of instructional priorities may follow.

The use of off-the-shelf tests for high-stakes accountability often leads to practices that undermine the validity of inferences about the achievement domains that the tests are intended to assess. The use of "scoring high" materials closely tailored to particular standardized tests is designed to raise scores. But increased scores do not necessarily mean that improvements would generalize to a domain of content broader than the test. In particular, when teaching effectiveness is equated with student gains, it becomes impossible to distinguish between instructional practices that narrowly teach to the test and those that genuinely promote student skill and knowledge in the broad domains reflected in the curriculum.

Gains in scores on state assessments generally are interpreted as improvement in student achievement, and by implication, the quality of education. The reasonableness of such an interpretation depends on the degree to which generalizations beyond the specific assessment administered by the state to the broader domains of achievement defined by the content standards are justified. A variety of factors, such as teaching that is narrowly focused on the specifics of the assessments rather than on the content standards the assessments are intended to measure, may undermine the validity of desired generalizations. Hence, it is important to evaluate the degree to which generalizations of gains on assessments to broader domains of achievement are justified. One practical and relatively powerful way of investigating generalizability is to compare trends for state assessments with trends for the state on the National Assessment of Educational Progress (NAEP). A systematic study comparing TVAAS and NEAP results would be highly instructive.

Conclusion

The idea of evaluating schools and teachers on the basis of “value added” to students’ education each year has wide appeal for policymakers. Instead of ranking schools from best to worst, the intention is to monitor the amount of gain in student achievement from one grade to the next. This approach has obvious advantages over traditional alternatives when coupled with a sophisticated statistical modeling apparatus capable of handling massive cumulative longitudinal data. Technical and methodological sophistication, however, are only part of the full array of considerations that form a comprehensive evaluative judgment. Ultimately, the value of proposed use of any methodology and the information it produces heavily depends on the soundness of claims made by the system’s advocates. A validity argument assembles and organizes the empirical evidence, as well as the logical line of reasoning linking the evidence to favored inferences and conclusions. Haertel (1999) has pointed out two weaknesses of the typical validation inquiry: a “checklist fashion” for amassing supporting evidence, and “a powerful built-in bias toward looking for supporting evidence, not disconfirming evidence” (p. 6). Both symptoms are evident when we examine the case for using TVAAS teacher effects as indicators of teacher effectiveness.

This paper points to some of the considerations that deserve closer attention when evaluating the soundness of inferences drawn from the TVAAS estimates of teacher effectiveness. Evidence and arguments were presented to call for more systematic studies of the system. Specifically, such studies need to address the potential confounding of teacher effects and other independent factors contributing to student academic progress, the dependency of estimates of teacher effects on model assumptions and on the context of their school systems, the explicit links between student score gains and instructional practices, and the generalizability of multiple-choice test results as indicators of instructional impact on student progress toward desirable educational goals. Such studies need to combine re-analyses of the TVAAS database, sensitivity analyses employing simulations, surveys and focus groups of teachers and administrators, intensive content analyses of the match between the TCAP and the state content standards, and small-scale randomized teaching experiments. The complexity of the TVAAS model and the nature of the Tennessee accountability system based on this model require no less in order to ground the proposed interpretations of estimates of schools and teachers on student learning in sound scientific evidence.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bock, R. D., & Wolfe, R. (1996). *Audit and review of the Tennessee Value-Added Assessment System (TVAAS): Final report*. Nashville, TN: Tennessee Comptroller of the Treasury, Office of Education Accountability.

Darling-Hammond, L. (1997). Toward what end: The evaluation of student learning for the improvement of teaching. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid measure?* (pp. 248–263). Thousand Oaks, CA: Corwin Press.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, *8*(1). Retrieved May 15, 2002, from http://olam.ed.asu.edu/epaa/v8n1

Education Improvement Act, 9 Ten Stat. Ann. §§ 49-1-603-608 (1992).

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5–9.

Hu, D. (2000). *The relationship of school spending and student academic achievement when achievement is measured by value-added scores*. Unpublished doctoral dissertation, Vanderbilt University, Nashville, TN.

Sanders, W. L., & Horn, S. (1995). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. *Education Policy Analysis Archives*, *3*(6). Retrieved May 10, 2002, from http://olam.ed.asu.edu/epaa/v3n6.html

Sanders, W. L., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*, 247–256.

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: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Talbert, J. E., & McLaughlin, M. W. (1993). Understanding teaching in context. In D. K. Cohen, M. W. McLaughlin, & J. E. Talbert (Eds.), *Teaching for understanding: Challenges for policy and practice* (pp. 167–206). San Francisco: Jossey-Bass.

Wenglinsky, H. (2000). *How teaching matters: Bringing the classroom back into discussions of teacher quality*. Princeton, NJ: Educational Testing Service.