Educational Evaluation and Policy Analysis
Fall 2003, Vol. 25, No. 3, pp. 287–298

Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System

Haggai Kupermintz
University of Haifa, Israel

This article addresses the validity of teacher evaluation measures produced by the Tennessee Value Added Assessment System (TVAAS). The system analyzes student test score data and estimates the effects of individual teachers on score gains. These effects are used to construct teacher value-added measures of teaching effectiveness. We describe the process of generating teacher effectiveness estimates in TVAAS and discuss policy implications of using these estimates for accountability purposes. Specifically, the article examines the TVAAS definition of teacher effectiveness, the mechanism employed in calculating numerical estimates of teacher effectiveness, and the relationships between these estimates and student ability and socioeconomic background characteristics. Our validity analyses point to several logical and empirical weaknesses of the system, and underscore the need for a strong validation research program on TVAAS.

Keywords: *learning gains, student achievement, teacher effectiveness, teacher evaluation, TVAAS, validity, value-added assessment*

Student test score gains have been proposed recently as a measure of the educational "value-added" contributed by teachers and schools to student learning. Recent educational reform efforts seek to employ standardized test score gains as a key policy instrument for holding educators and school systems accountable. The attainment of adequate yearly progress is therefore a cornerstone of the No Child Left Behind legislation (Linn, 2003). Previous efforts to develop systems of teacher evaluation based on student performance have long frustrated education leaders and policy makers. Shrinkfield and Stuffelbeam (1995) argued that "there is no topic on which opinion varies so markedly as that of the validity of basing teacher effectiveness on student learning," and Millman and Schalock (1997) commented that persistent substantive and methodological shortcomings have contributed to "teacher skepticism and growing criticism of attempts to link learning gains to teacher work" (p. 7).

Nevertheless, renewed interest in testing is influenced by the business metaphor of contemporary accountability discourse that views test scores as accurate measures of educational "value." This has provided the impetus for the development of new teacher evaluation systems, utilizing longitudinal analyses of test data. By modeling student progress over time, the argument goes, value added analyses provide accurate and trustworthy quantitative measures of student learning. These measures, in turn, can be directly attributed to the professional efforts of individual educators and schools, thereby mitigating "many problems in assessment and measurement" (Sanders, 2000, p. 331). For a review of value-added indicators and their potential use in appraising school and teacher performance, see Meyer (1996).

Advances in testing practices, psychometric and statistical modeling, as well as in longitudinal data collection, management, and maintenance have ushered a new generation of value-added models, specifically designed to support accountability systems by providing information on educational productivity. A pioneering

effort, and currently the most influential value-added model is the Tennessee Value Added Assessment System (TVAAS). The system has been developed in the late 1980s by Dr. William L. Sanders at the University of Tennessee, as the keystone of the Tennessee Education Improvement Act in 1992. TVAAS has generated tremendous attention among policy makers, administrators, and educators and is currently available commercially from the SAS statistical software company's SAS InScool department. Twenty one states, including Colorado, Ohio, and Pennsylvania, are experimenting or using the TVAAS model (Olson, 2002).

Some commentators hailed TVAAS as "an accountability revolution . . . that can provide an objective answer to questions of teacher effectiveness" (Stone, 1999, p. 240). This article examines the soundness of such assertions. We direct attention to the manner by which estimates of teacher effectiveness are defined and calculated in TVAAS and their validity for purposes of setting educational policy and making personnel decisions.

An Overview of TVASS

TVAAS is the centerpiece of an ambitious educational reform effort implemented by the Tennessee Education Improvement Act (1992). Inequalities in school funding led to a lawsuit brought against the state by a coalition of small rural districts. Under pressure from the business sector to reform the system, a strong accountability model was adopted by the legislature. Concrete evidence was to be provided for satisfactory year-to-year improvements down to the classroom level. Based on pilot studies with the value-added model conducted by Sanders and his colleagues during the 1980s, the Tennessee legislature embraced the model as the methodological backbone of the new accountability system. 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, 1992, §49-1-603)

Under the Tennessee accountability system, schools and school systems are expected to demonstrate progress at the level of the national norm in five academic subjects, as measured annually by scores on a battery of standardized tests comprising the Tennessee Comprehensive Assessment Program (TCAP). Beginning in 1993, value added reports have been issued to educators and the public on every school and school system. Teacher reports are not part of the public record; rather, value-added assessment of teachers has been provided since 1996 only to teachers and their administrators. We will describe in some detail certain technical and substantive features of the system that are especially relevant for appreciation of policy implications, but will not attempt a comprehensive review. For further details on the TVAAS methodology, see Sanders, Saxton, and Horn (1997).

Validity Considerations

Validity is the essential consideration in the evaluation of the uses and interpretations of any assessment. The logical and evidential bases for claims and inferences about scores obtained from any testing procedure are captured by a validity argument. The case for proposed interpretations

and inferences offered to support a suggested use, rests on favorable empirical findings in light of theoretical propositions regarding the nature of the construct purported to be measured. A key condition for establishing a compelling case is the demonstration that competing explanations or "rival hypotheses" are less consistent with the facts. This article describes and evaluates the case for using TVAAS as a teacher evaluation tool. We examine the arguments and evidence offered for interpreting TVAAS teacher effects as indicators of teacher quality and comment on their use for guiding educational policy. Specifically, the article looks at the TVAAS definition of teacher effectiveness, the mechanism employed in calculating numerical estimates of teacher effectiveness, and the relationships between these estimates and student ability and socioeconomic background characteristics. Policy implications of our validity investigation are discussed for each of these issues. We begin by examining the definition and measurement of teacher effectiveness.

The Definition of "Teacher Effectiveness"

A typical interpretation for the between-teacher variability in TVAAS teacher effects is that "the single largest factor affecting academic growth of populations of students is differences in effectiveness of individual classroom teachers" (Sanders, 1998). The conclusion is based on the empirical "finding" that "differential teacher effectiveness is a strong determinant of differences in student learning" (Darling-Hammond, 2000). The statement appears to imply that there are two distinct variables—teacher effectiveness and differences in student learning—and that the former causes the latter. Unfortunately, such causal interpretation is faulty because teacher effectiveness is *defined and measured* by the magnitude of student gains. In other words, differences in student learning *determines*—by definition—teacher effectiveness: a teacher whose students achieve larger gains is the "effective teacher." TVAAS divides teachers into five "effectiveness" groups according to their ranking among their peers in terms of average student gains. To turn full circle and claim that teacher effectiveness is the *cause* of student score gains is at best a necessary, trivial truth similar to the observation that "all bachelors are unmarried."

Before gains in student test scores can be interpreted as a measure of teacher effectiveness, real evidence must be offered. The proponent must demonstrate that other hypotheses are less plausible explanations of the observed between-teacher difference in student performance. Do student characteristics or socioeconomic backgrounds account for this variability? Do other school or community context variables systematically covary with teacher effects? Such questions should drive rigorous, systematic validity investigation. In the meanwhile, a more careful interpretation of the TVAAS findings on teacher effects is called for. Teacher effects document between-teacher variability in the average test score gains of their students. This variability may arise for different reasons, some of which directly associated with teacher effectiveness, but others may reflect the context in which teaching occurs or the qualities of the specific group of students being taught. Policy makers and administrators who wish to use the TVAAS value-added information must consider these alternative explanations when contemplating the likely consequences, intended and unintended, of any policy move. Any systematic differences among teachers or students that correlate with value-added scores may offer insight. Further research should collect and analyze convergent and discriminant evidence by using independent criterion measures of teaching effectiveness as well as student and school variables. In subsequent sections, we present findings that demonstrate potential confounding of teacher effects with student and classroom characteristics. These findings suggest that a simplistic interpretation of value-added data is unwarranted. But first, additional questions emerge as we consider in the next section the actual calculation of teacher effects.

Construction and Calculation of Teacher Effects

The calculation of teacher effects in TVAAS is a complex process that blends the estimation of the average performance gains in each school system,[1] and the average performance of each teacher's students, relative to the system performance. In order to understand the process we must first recognize that the TVAAS is comprised of three different statistical models: (a) a *system model* estimating average performance a particular school system, for each year, grade, and academic subject, (b) a *school model* estimating average performance for a particular school within a system, and (c) a *teacher model* estimating the

average student performance associated with a particular teacher in the system. Only the system and teacher models will concern us here.

An example of calculating teacher gains is given in Sanders, Saxston, and Horn, 1997 (pp. 156–160). Imagine a specific school system in Tennessee. In 1993, the average reading scores of the system's 2nd-grade students was estimated by the system model to be 662.9; in 1994, the estimated average of the same student cohort, now in third grade, was 688.1. Consequently, the average system gain in reading between second to third grade is a simple difference between average scores: 688.1 – 662.9 = 25.2. Now, imagine a particular 4th-grade teacher whose estimated deviation from the system average—the teacher effect—in 1994 in reading was 1.6—the estimated readings scores of this teacher's 4th grade students in 1994 was on average 1.6 points above the system's. It is important to note that the teacher model constrains teacher effects to average to zero within each school system. Using a statistical mechanism called "estimable functions," a combined estimate of teacher gains is computed by adding the teacher effect to the system average gain: 25.2 + 1.6 = 26.8. "Thus estimable functions that add the teacher effect to the system gain translate the teacher effect into a measurement of gain" (Sanders, Saxston, & Horn, 1997, p. 156). In order to make the statistical computations manageable, data are processed and effects are estimated separately within each system (or county, in cases of multiple systems within a county). Individual teacher reports are based on a 3-year average of teacher estimated gains, calculated as described above.

It is clear from the algorithm that each individual teacher estimate depends on the performance of all other teachers in the system. In other words, TVAAS teacher effects are norm-reference measures that rank teachers within each school system. Criterion-reference interpretations of teacher effects or comparisons of teacher scores across system are unwarranted. Questions about fairness and equity must be raised if personnel decisions employ normative information that imply in practice different standards or benchmarks in different school systems. Teacher effects can not be interpreted in terms of absolute performance standards. For example, a weak teacher in relatively weak school system may obtain a more favorable evaluation in comparison with a similarly weak teacher in a strong system. School systems in Tennessee differ widely on value added measure. In 2002, value-added estimates ranged from 71% of the national norm gain to 130% in math, and from 74% to 145% in reading (Value added information is routinely posted on the Tennessee department of education web site. The 2002 data are available at http://www.state.tn.us/education/tstvaas2.htm). Due to the substantial variability in performance between systems, teachers in low- and high-performing systems will be judged against very different evaluative criteria.

Other school or system differences, beyond test score performance, may further erode the validity of teacher effects as indicators of the quality of individual efforts exerted by particular teachers. The TVAAS model represents teacher effects as independent, additive, and linear. Educational communities that value collaborations, team teaching, interdisciplinary curricula, and promote student autonomy and active participation in educational decisions may find little use for such information. A model that regards teachers as isolated, independent actors and students as passive recipients of teacher "effects" may not be adequate in some contexts. When the fit between the model and the phenomenon it seeks to represent is poor, validity is threatened, as the example below demonstrates.

When a science teacher emphasizes the computational aspects of the curriculum and requires his students to engage in intensive mathematical explorations, increased student mathematical proficiency should be expected. When the math teacher collaborates or coordinates her efforts with the science teacher to help students meet the elevated demands of the science curriculum, further facilitation of students' math ability may be realized. The availability of high quality, technology-rich learning environments at the school (for example, cognitive tutors, see Anderson, Corbett, Koedinger, & Pelletier, 1995) introduce additional opportunities for learning and teaching. Attempts to disentangle such complex, interwoven contributions of the science teacher, the math teacher, and the computerized learning environment into isolated, independent "effects" are not only methodologically intractable but also conceptually misguided. Teaching and learning are aspects of a synergistic phenomenon whereby dynamic forces continuously interact to

produce accumulating changes in student knowledge structures, repertoire of problem solving strategies, meta-cognitive capacity, as well as attitudes, affects, and volition. An additive, linear depiction of this highly complex phenomenon is clearly inadequate in many cases. Teacher evaluations, to be valid, must take into account the context of teaching (Talbert & McLaughlin, 1993), or at least a reasonable approximation.

The Accuracy of Teacher Effects

When statistical estimates are used for formal evaluation of teachers, accuracy becomes a key consideration. In TVASS, the accuracy of estimated teacher effects depends on the amount of data available for each teacher—estimates for teachers with less data (i.e., less students taught in a particular year) are less accurate than those of teachers with more data. Furthermore, teacher effects are "shrunken" towards the system's average—when student data are scarce, a teacher is assumed by the model to perform at the average level of his or her school system. The fewer the students, the stronger the pull towards 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, Saxston, & Horn, 1997, p. 143). From a statistical point of view, this strategy makes the most efficient use of all existing data, but we must still ask whether this strategy provides adequate protection against potential bias. An analogy may clarify the dilemma: Suppose students who miss certain tests (and therefore have less data on which to base an evaluation) are assumed to perform at the average classroom or school level, and their scores are subsequently compared to those of students who were tested more thoroughly. Would we consider such an approach to student evaluation fair if substantial consequences, graduation from high school for example, depended on these scores?

An outstanding teacher who taught more students will be correctly identified, whereas an equally remarkable teacher serving a more transient student population would appear less exemplary because her performance score will be pulled towards the system average. Similarly, a poor teacher may evade detection. Moreover, teachers in different school systems will be pulled towards different means—equally effective teachers with the same amount of data will be judged differently due to differences in the average performance of their respective school systems. The degree to which system average will mask true teacher differences is determined by how many students each teacher taught, rather than by the absolute quality teaching.

According to Sanders (personal communication, February 9, 2000), TVAAS is designed to reduce the likelihood of incorrectly identifying low-performing teachers. However, no systematic study has examined the rates of false positive and false negative classifications, and whether certain teachers are more likely to be adversely affected by virtue of the student population they serve. Darling-Hammond (1997) warned that, "no person should be evaluated for high-stakes decisions based on statistical assumptions rather than on actual information" (p. 255). Policy makers should consider whether minimum requirements for statistical accuracy should be set before information can be employed in personnel and policy decisions.

Attribution of Gains: Students or Teachers

Let us return to the question of interpreting teacher effects in light of conflicting explanations. There is a growing recognition that "[e]ffective instruction begins with what learners bring to the setting" (Bransford, Brown & Cocking, 1999, p. xvi). Students enter the classroom with powerful ideas, knowledge and skills, as well as learning habits and practices that provide the necessary resources for further learning. Effective instruction will allow each student to progress to his or her maximum potential. In one case progress to the full potential may be rapid and robust, while in another only small tentative steps are possible even with the most talented and motivated teacher.

The concept of aptitude (see Corno et al., 2002) is useful when thinking about potential growth in response to instruction. Individuals vary greatly in their readiness to profit from instruction or aptitude. It is therefore necessary to take student aptitude into account when evaluating teachers. Equally competent teachers will produce different results with groups of students that differ appreciably in cognitive, affective, and motivational aptitude profiles. Moreover, student aptitude profiles may reflect social and cultural influences because family and community resources contribute considerably to the development and

realization of student potential. When such resources vary substantially among groups of students served by different teachers, an evaluation system must provide adequate safeguards. A fair assessment for students must rely on appropriate opportunities to learn. Similarly, teachers must be afforded appropriate opportunities to teach to be evaluated fairly. Consistent assignment to teaching more challenging students will clearly hamper a teacher's chances of showing strong measurable progress in terms of student gains on test scores.

The TVAAS model takes student aptitude into account by treating student prior achievement (i.e., previous years' test scores) as a "blocking" variable intended to statistically adjust for differences in preparedness for instruction. Students' progress is then evaluated relative to their own prior achievement. The rationale is that "the child serves as his or her own 'control.' This enables the partitioning of school system, school, and teacher effects free of the exogenous factors that influence academic achievement and that are consistently present with each child over time" (Sanders & Horn, 1998, p. 249). The statistical blocking adjustments were developed in the context of controlled experiments, the design of which is under the investigator's control, and their effectiveness depends on two conditions: 1) random assignment of students to teachers, or 2) a careful, systematically balanced allocation of students to teachers. The application of statistical adjustments in observational studies may prove problematic when the necessary experimental conditions cannot be met or assumed (Weisberg, 1979), as demonstrated in the following example.

A reanalysis of data presented by Sanders & Rivers (1996) in an unpublished but well-cited report shows a strong association between teacher effects and student prior achievement. Such pattern calls into the question the trustworthiness of blocking as a strategy for "leveling the playing field." Figure 1 presents data from a study that examined 5th-grade achievement in math (Sanders & Rivers, 1996). Students are divided into four prior achievement groups, from low to high, and the proportions of teachers in each group evaluated as the least or most effective is presented.

In the lowest prior achievement group, slightly more than 10% of the teachers were evaluated as highly effective, while almost 30% of the teachers were evaluated as least effective. In contrast, in the highest prior achievement group, slightly

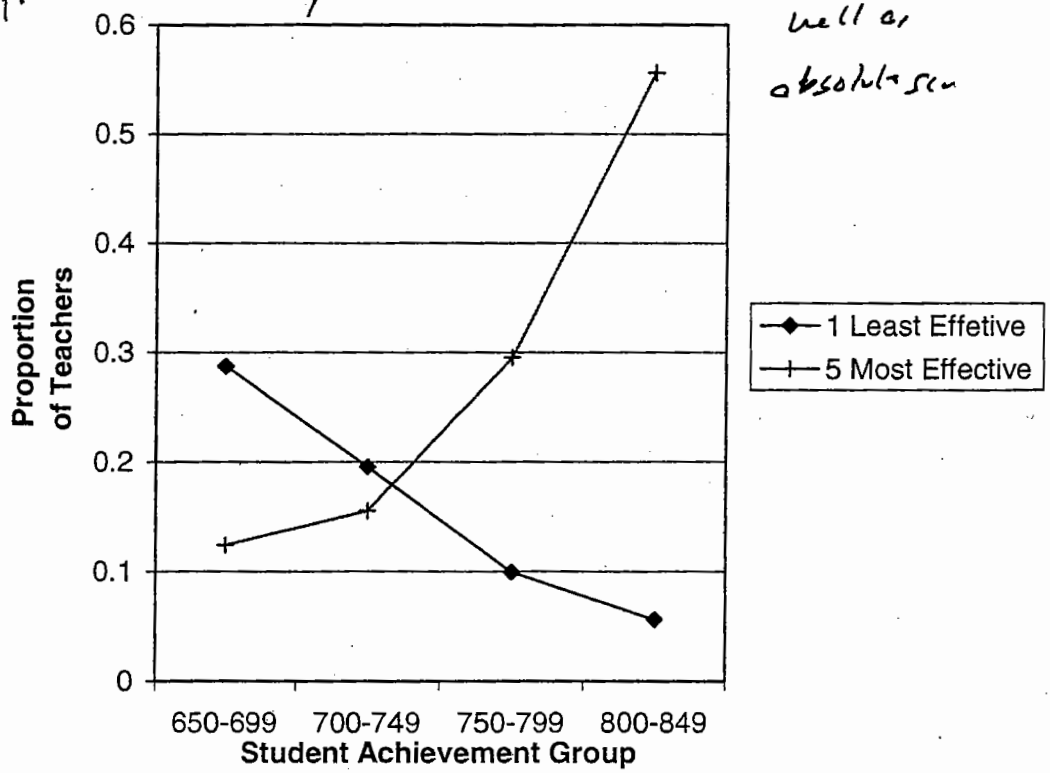

FIGURE 1. *Teacher effectiveness by student achievement.*

more than 5% of the teachers were evaluated as ineffective, and more than half were evaluated as highly effective teachers. It is unclear whether these results reflect systematic inequalities in the assignment of students to teachers or a possible misattribution of effects to teachers rather than to student aptitude (as captured by prior achievement). In either case, difficulties arise when we try to disentangle statistically student and teacher responsibility for the observed gains. If the "blocking" strategy was successful we would expect to find comparable distributions of teacher effectiveness estimates across student groups with different prior achievement.

The potential misattribution of teacher effects is demonstrated further by using simulation data, which reflect more systematically different configurations of student and teacher contributions to gains. In response to a query by the author, Sanders (personal communication, December 12, 1999) provided an analysis to demonstrate that TVAAS "does indeed consider prior achievement of students during the estimation process." The same analysis was used with different data. Student and teacher true and independent contributions to gains were simulated and model estimates were examined against true teacher effects. Table 1 shows the results for five hypothetical teachers, each with five students, under three different scenarios. Overall gain is the summation of true student and teacher contributions to gains—the teacher and student effects generated by the simulation and therefore their true values are known to us; teacher estimates show the effects estimated by Sanders' analysis. A successful analysis should recover the true teacher effects accurately.

In Simulation I, teacher simulated contributions to gains were all set to zero, yet the estimates of teacher effects produced by the TVAAS analysis are nonzero and reflect the relative rank order of contributions to gains made by students. Simulation II shows that when effective teachers (those who produced stronger gains by our design) are systematically assigned weak students and vice versa, teacher and student contributions operate in opposite directions to cancel each other and produce zero TVAAS estimates for teachers. Simulation III again shows that student independent contributions to gains may distort the estimates of teacher contributions. We hasten to comment that these demonstrations are highly contrived and do not adequately represent the full TVAAS model; yet, these simulations are instructive because they draw attention to the potential biases inherent in teacher estimated effects due to the confounding of independent teacher and student contributions to score gains. Coupled with the empirical evidence of a sizable correlation between TVAAS teacher effects and students prior achievement, these analyses highlight the need for systematic research on the issue.

TABLE 1
Teacher Estimates as a Function of Student and Teacher True Effects

	Overall gain	True Effect		Teacher Estimate
		Student	Teacher	
Simulation I				
Teacher 1	5	5	0	−5.17
Teacher 2	5	5	0	−4.97
Teacher 3	15	15	0	5.07
Teacher 4	15	15	0	5.07
Simulation II				
Teacher 1	20	5	15	0.04
Teacher 2	20	5	15	−0.03
Teacher 3	20	15	5	−0.01
Teacher 4	20	15	5	0.00
Simulation III				
Teacher 1	25	5	20	1.02
Teacher 2	20	5	15	−1.68
Teacher 3	25	15	10	1.70
Teacher 4	20	15	5	−1.04

The question of responsibility for student learning is central in attempts to construct and implement teacher accountability systems. Failure to achieve proper isolation of teacher direct effects on learning may result in perverse policy decisions, benefiting teachers who are routinely assigned to students likely to make stronger gains, regardless of their teachers. Teachers with a more problematic student "clientele" are likely to be evaluated more harshly by the system. Furthermore, the confusion of student and teacher effects may unintentionally entice teachers to seek the strongest students at the expense of students who require more investment but can offer only small gains in return. As stakes increase, the incentive for teachers to target "high yield" students will intensify. Attaching tangible rewards or sanctions to value-added information is likely to encourage the development of a cynical calculus of the worth of different students to maximizing teachers' return on the investment. Linking student test score gains to teacher financial gains may appeal to some policy makers, but our analyses should give them pause.

The exclusive attribution of gains to teachers conceals potentially harmful practices whereby teachers are effective with certain students but not with others. Because student variability in gains *within a classroom* is averaged during the statistical calculations to produce a total teacher effect, teachers who tend to concentrate efforts on students who are likely to demonstrate more robust gains at the expense of other, more challenging, students goes unnoticed. The negative long-term consequences of transforming student test score gains into the ultimate goal for teachers will probably be felt strongest by those students whom the new education legislation promised not to leave behind. Few teachers or schools would be able to afford ignoring the calculations of optimal gain in favor of pure pedagogical considerations.

The Role of School and Community Context

Finally, we turn our attention to the full range of potential influences on student learning: personal propensities and resources (both cognitive and noncognitive), physical and mental maturation, home environment, cultural heritage, institutional and informal community resources. It is within this context that the effects of formal, institutional schooling must be understood. In other words, formal education does not work in a vac- uum. Furthermore, even if we confine our attention to formal schooling alone, complexity abounds. School culture and climate, teacher qualifica- tions, curriculum frameworks and instructional approaches, and a myriad of other factors, interact synergistically to produce growth in student academic skills and knowledge. This complexity has consistently defied simple explanations.

The problem is further complicated because as Darling-Hammond & Post (2000) commented, "few Americans realize that the U.S. educational system is one of the most unequal in the industrialized world, and students routinely receive dramatically different learning opportunities based on their social status." (p. 127). Consequently, learning environments in the United States present themselves as "syndromes" or amalgams rather than as additive clusters of independently accrued conditions. Low SES students, for example, experience scholastically impoverished home environments, go to schools where facilities are lacking or inadequate, learn with less qualified teachers who tend to use unchallenging curricula and uninspiring instructional methods, and are routinely subjected to explicit or implicit segregation along ethnic lines. These students consistently lag behind their more privileged peers in academic achievement and progress. Vast inequalities have been amply documented and discussed intensively among educational researchers, educators, and policy makers (see Kozol, 1991).

TVAAS developers have made the bold claim that by using student prior achievement as a covariate (or blocking factor, as explained above) the model adequately accounts for all the potent external influences on student learning, thereby allowing the proper isolation of teacher direct effects on learning. Thus, student prior achievement plays in TVAAS the dual role of a measure of student aptitude as well as an omnipotent "proxy" for a wide range of social and cultural background factors: "Each child can be thought of as a 'blocking factor' that enables the estimation of school system, school, and teacher effects free of the socioeconomic confoundings that historically have rendered unfair any attempt to compare districts and schools based on the inappropriate comparison of group means" (Sanders, Saxton, & Horn, 1997, p. 138).

Differences in socioeconomic factors correlate with prior achievement but the magnitude of the

...ical correlation cannot justify using the achieve-...ent variable as the sole proxy for background ...ariables. Moreover, consistent evidence confirms ...e intuitive conjecture that socioeconomic factors ...are correlated with student *gains* even after prior ...achievement has been accounted for. In a recent ...example, using a large longitudinally-matched ...data set of 5th-grade students in North Carolina, ...adadd & Walsh (2002) reported sizable correla-...tions between socioeconomic status and school ...value-added scores. "Use of either the South ...Carolina or the North Carolina model for calcu-...lating a school's value added would still yield a ...measure of school effectiveness that is posi-...tively correlated with average performance and ...negatively correlated with the percent of stu-...dents eligible for subsidized lunches or the per-...cent black . . . Schools that have a dispropor-...tionately high intake of high income and White ...students continue to look better than those serv-...ing students from more economically or racially ...disadvantaged backgrounds." (p. 11). Berk (1988) ...summarized the empirical evidence document-...ing numerous student and school characteristics ...that correlate with test score gains. Recognizing ...the importance of background variables to under-...standing student progress, the Dallas value-added ...accountability system (which shares many of the ...statistical techniques underlying TVAAS) includes ...such "fairness variables" in the calculation of ...school and teacher productivity measures because ..."clearly, certain student populations are more dif-...ficult to educate; if we are to hold all schools and ...teachers accountable, then we have to create a *level ...playing field* for making these judgments" (Thum ...& Bryk, 1997, pp. 102–103, italics in original).

In contrast, Sanders & Horn (1998) reported that "the cumulative gains for schools across the entire state have been found to be unrelated to the racial composition of schools, the percentage of students receiving free and reduced-price lunches, or the mean achievement level of the school" (p. 249). No specific study was cited as support; it seems that the source of the contention are data displays included in an unpublished report circulated by the University of Tennessee Value-Added Research and Assessment Center (1997). The document, whose authors are unidentified, displays scatter plots of the percentage of minority students in each of some 1,000 Tennessee schools against the 3-year estimated cumulative average gains in each school. The report does not provide any formal statistical analyses of these patterns, leaving the reader to evaluate its conclusions by eyeballing the graphical displays. The report concludes that, "the graphs show that the effectiveness of a school cannot be predicted from a knowledge of the racial composition."

A closer inspection of the graphs reveals that schools with more than 90% minority enrollment tend to exhibit lower cumulative average gains. For example, about 70% of the schools with high-minority enrollment showed gains in math that were below the national norm; comparable patterns can be observed for reading, language, and social studies. Similar graphs for school systems reveal an even stronger relations between average gains and the percentage of students eligible for free or reduced-price lunch. An additional inspection of value-added data reported for Tennessee schools in the 1999–2000 school year shows, for example, that schools in the bottom quartile of the distribution of student participation in the free or reduced-price lunch achieved on average around 103% of the national norm gain, while schools in the upper quartile achieved only around 95% of the norm. To date, no systematic study has documented the extent to which various external variables correlate with TVAAS value-added scores.

Sanders and Rivers (1996) provide further evidence for the role family background factors may play in influencing student progress. Table 3 in Sanders and Rivers gives the frequency with which White and Black students were assigned to teachers in each effectiveness level. Generally, White students were more often associated with more effective teachers than were Black students. 15.9% of the teachers of white 3rd-grade students were identified as ineffective, compared with 26.7% for the Black students. In contrast, 22.4% of the White students, and 14.4% of the Black students, respectively, were associated with teachers in the highest effectiveness level.

The inclusion of demographic variables in TVAAS analyses may lead to changes in teacher classification into effectiveness levels. Ballou (2002) reports that including demographic controls in the TVAAS model did not change the numerical scores for teacher effectiveness by a great amount. These changes, nevertheless, had noticeable effects on whether a teacher succeeded or failed to reach a preset performance standard—"more than one third of the teachers who ranked

in the top 10% when our assessments included socioeconomic and demographic controls no longer belonged to that category when these controls were omitted from the analysis." (Ballou, 2002, p. 14). In other words, about a third of the teachers who deserved to be rewarded for superior performance could be denied recognition because the calculation of their effects did not take into account factors beyond their control and that potently affected their students' achievement. The prospect of incorporating potentially biased teacher evaluations in policy decisions suggests that the lack of independent, reliable, and rigorous research on TVAAS is a serious gap in need of quick remediation, before firm policies are put in place that attach stronger consequences to TVAAS scores.

The Need for a Strong Validation Program

The use of standardized test scores for the purposes of teacher evaluation has been criticized elsewhere. Koretz (2002) concluded that, "what is needed is an active program of research focused on both the development and the evaluation of alternative methods of holding educators accountable." Given the increasing popularity and use of value-added methodology for accountability purposes, the paucity of published research findings from TVAAS that specifically pertain to teacher effectiveness is puzzling. TVAAS findings on teacher effects have been discussed in only three peer-reviewed journal articles, two book chapters, and three unpublished research reports, all of which authored by TVASS staff. Moreover, only one journal article and two unpublished reports actually present findings from original empirical studies (none of which used the full TVAAS model in its analyses!) Two unpublished dissertation studies, one by a former TVASS staff member and the other by one of the authors of a 1995 evaluation of TVAAS, provide additional analyses. Other publications, as well as numerous presentations and newspaper interviews with Sanders and other TVAAS staff, typically repeat these findings and their implications or provide general descriptions of the statistical methodology, program operations, and the variety of reports produced by the system, but do not provide additional empirical findings on the issues raised in this article.

In light of the potential threats to the validity of TVAAS teacher evaluation information, a serious research program is urgently needed. We have seen that the definition of teacher effectiveness in TVAAS may be misleading, that the normative aspect of calculating teacher effects may erode the interpretation and comparability of these effects, that the volatile accuracy of teacher effects may result in misclassification of teachers, and that the classroom composition of ability and social or ethnic background characteristics may influence the magnitude of teacher effects, regardless of teacher quality.

As part of a strong validation research program for investigating the seriousness and magnitude of these threats to validity, several studies can be recommended. The correlations of TVASS teacher effects with a variety of teacher, classroom and student variables will provide essential convergent and discriminant validity information. Teacher variables may include independent teacher evaluations from peers and supervisors as well as indicators of teacher preparation and certification. Good candidates are variables, which were found to be strongly correlated with student achievement in reading and mathematics, as reported by Darling-Hammond (2000). Classroom and school variables should reflect the range of social and cultural factors that may be argued to affect student learning and growth, independent of teacher quality. These analyses should highlight both correlational measures and estimates of misclassification of teachers when they are assigned into discrete effectiveness levels. Special attention should be given to the study of the relationship between teacher effects and student prior performance. In addition to prior achievement test scores, more general aptitude measures should be included in the analyses. Lastly, a survey of teachers and administrators should provide information on the manner in which TVAAS scores are being used for supporting personnel and policy decisions. A recent survey conducted by the Tennessee Comptroller office (Morgan, 2002) concluded that lack of understanding has resulted in inappropriate or low use of TVAAS information. Consequently, teacher and administrator survey should also ask teachers and administrators how well they understand the basic assessment strategies employed by TVAAS and their strengths and limitations.

In order to enable a proper validity investigation, TVAAS data must be made available to interested, qualified researchers. To date, nu-

merous requests by the author for access to the TVAAS data have been met with blank refusals, offering no other reason than a concern that "the data may be misused." The Tennessee Comptroller's report concluded that "Tennessee, not Educational Value Added Assessment Services, owns the TVAAS data. Therefore, the state should make decisions on who has access to the information." Education researchers, such as Robert L. Linn from the University of Colorado, Boulder, and organizations such as the Carnegie Foundation have requested data directly from Sanders only to be turned down or stalled. Even officials within the state government have had trouble securing access to Sanders' data. Many of these organizations could provide excellent reviews of the TVAAS system and assist the state–free of charge—in analyzing data." (Morgan, 2002, p. 39).

Conclusion

The idea of evaluating schools and teachers on the basis of the "value-added" to students' education each year has wide appeal for policy makers. 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 technical advantages over the traditional alternatives when coupled with a sophisticated statistical modeling apparatus capable of handling massive cumulative longitudinal data. Technical and methodological sophistication, however, is only part of the full array of considerations that form a comprehensive evaluative judgment on the system's merit.

A validity argument assembles and organizes the empirical evidence as well as the logical line of reasoning linking supporting evidence to proposed inferences and conclusions. Haertel (1999) pointed out two weaknesses of the typical validation inquiry in the context of student testing: a "checklist fashion" for amassing supporting evidence, and "a powerful build-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 article underscored some of the considerations that deserve closer attention when evaluating the soundness of inferences drawn from the TVAAS value added scores. The complexity of the TVAAS model and the nature of the Tennessee accountability system require a comprehensive validity inquiry in order to ground the proposed interpretations of estimates of schools and teachers on student learning in sound scientific evidence. Important policy decisions must be supported by firm evidence as to the quality of information they rely upon.

Mixed models such as those adapted by TVAAS for the analysis of educational data are very useful in analyzing agricultural or other experimental data. The adaptor of such methods to a new domain must provide a compelling validity argument to convince the relevant stakeholders that the transaction is appropriate. Presently, the TVAAS methodology is supported mainly by general statistical theories but is only loosely aligned with relevant theories in education. It is the responsibility of the educational research community to provide the scientific, evidence-based assessment of the effectiveness of the system. Stakeholders, and especially teachers, students, and parents need such information in order to hold decision makers accountable. Until a more complete case for TVAAS and the value added methodology has been developed, policy makers will be prudent to adhere to Ballou's (2002) admonition: ". . . those who look to value-added assessment as the solution to the problem of educational accountability are likely to be disappointed. There are too many uncertainties and inequities to rely on such measures for high-stakes personnel decisions."

Notes

[1] In Tennessee, school districts are called "systems"; we will use this term for compatibility with the terminology used by TVAAS

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*(2),167–207.

Ballou, D. (2002). Sizing up test scores. *Education Next, Summer 2002,* 10–15.

Berk, R. A. (1988). Fifty reasons why student gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1,* 345–363.

Bransford, J. D., Brown, A. L., & Cocking, R. (Eds.), (2000). *How people learn: Brain, mind, experience and school.* Washington, DC: National Academy Press.

Corno, L., Cronbach, L. J. (Ed.), Kupermintz, H., Lohman, D., Mandinach, E. B., Porteus, A. W., &

Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of R. E. Snow.* Mahwah, NJ: Erlbaum.

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, Inc.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved February 28, 2003 from http://epaa.asu.edu/epaa/v8n1/.

Darling-Hammond, L., & Post, L. (2000). Inequality in teaching and schooling: Supporting high-quality teaching and leadership in low-income schools. In R. D. Kahlenberg (Ed.), *A notion at risk: Preserving public education as an engine for social mobility.* New York: Century Foundation.

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.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity, *Journal of Human Resources, 37*(4), 752–777.

Kozol, J. (1991). *Savage inequalities.* New York: Crown Publishers.

Ladd, H. F, & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review, 21,* 1–17.

Linn, R. L., (2003, Winter). *Requirements for measuring adequate yearly progress.* CRESST Policy Brief 6. National Center for Research on Evaluation, Standards, and Student Testing [CRESST], University of California, Los Angeles.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek, and D. W. Jorgenson, (Eds.), *Improving America's schools: The role of incentives,* (pp. 197–223). Washington, DC: National Academy Press.

Millman, J., & Schalock, H. D. (1997). Beginnings and introduction. In J. Millman (Ed.). *Grading teachers, grading schools: Is student achievement a valid measure?* (pp. 3–10). Thousand Oaks, CA: Corwin Press, Inc.

Morgan, J. P. (2002). Multiple choices: Testing students in Tennessee. TN: Comptroller of the Treasury, Office of Education Accountability.

Olson, L. (2002, November 20). Education scholars finding new 'value' in student test data. *Education Week, 22*(12), 1–14.

Sanders, W. L. (1998). Value-added assessment. *The School Administrator, 55*(11).

Sanders, W. L. (2000). Value-added assessment from student achievement data: opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329–339.

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*(3), 247–256.

Sanders, W. L. & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Research Progress Report. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added 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, Inc.

Shinkfield, A. J., & Stufflebeam, D. L. (1995). *Teacher evaluation: guide to effective practice.* Boston, MA: Kluwer Academic Publishers.

Stone, J. E. (1999). Value added assessment: An accountability revolution. In M. Kanstoroom and C. E. Finn, Jr. (Eds.), *Better teachers, better schools* (pp. 239–249). Washington, DC: Thomas B. Fordham Foundation.

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

Thum, Y. M., & Bryk, A. S. (1997). Value-added productivity indicators: The Dallas system. In Jason Millman (Ed.) *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 100–119). Thousand Oaks, CA: Corwin.

TVAAS. (1997). Graphical summary of educational findings from The Tennessee Value-added Assessment System. University of Tennessee Value-Added Research and Assessment Center.

Weisberg, H. I. (1979). Statistical adjustments and uncontrolled studies. *Psychological Bulletin, 86*(5), 1149–1164.

Author

HAGGAI KUPERMINTZ is Assistant Professor, Faculty of Education, University of Haifa, Haifa 31905, Israel; kuperh@construct.haifa.ac.il. His areas of specialization are assessment, methodology, and cognitive science.

Manuscript Received June 14, 2001
Revision Received March 6, 2003
Accepted July 8, 2003