

Teacher Evaluation

Policy Brief

by

Gene V Glass

Arizona State University

Education Policy Research Unit (EPRU)

Education Policy Studies Laboratory

College of Education

Division of Educational Leadership and Policy Studies

Box 872411

Arizona State University

Tempe, AZ 85287-2411

February 2004

EPSL | EDUCATION POLICY STUDIES LABORATORY
Education Policy Research Unit

EPSL-0401-112-EPRU

<http://edpolicylab.org>

Education Policy Studies Laboratory

Division of Educational Leadership and Policy Studies

College of Education, Arizona State University

P.O. Box 872411, Tempe, AZ 85287-2411

Telephone: (480) 965-1886

Fax: (480) 965-0303

E-mail: epsl@asu.edu

<http://edpolicylab.org>

This document is for internal distribution only. Do not quote. Editorial changes may be made in this document as it is prepared for web posting. Only the version of this document posted on the Education Policy Studies Laboratory website is the final draft.

Teacher Evaluation

Gene V Glass

Arizona State University

This document is for internal distribution only. Do not quote. Editorial changes may be made in this document as it is prepared for web posting. Only the version of this document posted on the Education Policy Studies Laboratory website is the final draft.

Executive Summary

Traditional forms of evaluating teachers (e.g., inspection of credentials, supervisor and peer observation and rating) for purposes of hiring, promotion, and salary increases have served the profession of teaching well for decades and should receive continued support in policy and practice.

Newer forms of evaluation—primarily paper-and-pencil tests for initial and re-certification, and “value-added” techniques such as the Tennessee Value Added Assessment System (TVAAS) that attempt to attribute students’ standardized achievement test score gains to the efforts and expertise of their current teacher—have serious shortcomings. Paper-and-pencil tests of candidates’ knowledge of teaching practices and even subject matter tests are of dubious validity and fail to meet ordinary standards of predictive validity.

Several recommendations at the state-wide policy level can be derived from the above consideration of the issues surrounding teacher evaluation in the State of Florida.

1. Any attempt to substitute test performance for college degree requirements in the teacher certification process should be opposed. Movements in this direction can be discerned in the legislatures in several states. Such policies would surely result in a less skilled and less professional teaching corps. Furthermore, the questionable validity of paper-and-pencil tests can not support such practices.

Certification standards for out-of-state teachers are currently less stringent than for graduates of approved in-state programs of teacher preparation. On account of reciprocity agreements with other states and the issuance of

temporary teaching certificates to graduates of out-of-state teacher preparation programs, in-state graduates face a more daunting row of hurdles to certification (because of an additional entrance examination—the College Level Academic Skills Test—required to *enter* an approved preparation program) than out-of-state graduates. Holders of temporary certificates have three years in which to pass the FTCE tests.

2. Value-added teacher evaluation methods, which attempt to evaluate teachers in terms of the standardized achievement test score gains of their students, are of uncertain validity, have drawn heavy criticism from measurement experts, and raise serious concerns about fairness. They should be opposed in their various forms. References in current statutes (K-20 Education Code: 1012.34 “Assessment procedures and criteria”) such as “The assessment procedure for instructional personnel and school administrators must be primarily based on the performance of students assigned to their classrooms or schools” should be removed from legislation because no method of validly and fairly attributing student test performance to individual teachers or administrators is presently available.

Teacher Evaluation

Section 1: The Issue

Traditional means of evaluating teachers for purposes of hiring, promotion or salary increases have included supervisor (mainly building principal) observation, less often peer observation, credentials review (crediting teachers for professional development activities such as post-graduate education), and much less frequently, student ratings or other forms of evaluative feedback. K-12 schools have decades of experience with these methods; they have been the object of study by researchers for generations, and by and large they are unproblematic and do not arise as hot button policy issues in current political debates.¹

Two methods of teacher evaluation do lie at the center of contemporary policy debates, however: *testing of teachers* and *using students' test scores to evaluate teachers*. The discussion and analysis of these two approaches to teacher evaluation form the substance of this brief.

Section 2: Background

Testing of Teachers

Florida administers the Florida Teacher Certification Examinations to candidates for a teaching certificate in the state of Florida. The FTCE comprises three separate tests: Professional Education, General Knowledge, and Subject Area Exams. Depending on a candidate's background, he or she may be required to take one, two, or all three of these tests. The *Professional Education Test* is multiple-choice test that assesses general knowledge of pedagogy and professional practices and is made up of about 120 items.

The *General Knowledge Test* is a basic skills achievement test made up of four subtests: three multiple-choice tests (Mathematics, Reading, English Language Skills), and an Essay examination. *Subject Area Examinations* measure content area knowledge, usually by means of multiple-choice items. They are intended for certification of secondary school teachers in specific subjects. The tests cover, among other areas, English Grades 6-12, English Grades 5-9, French Grades K-12, German Grades K-12, and Spanish Grades K-12.

Only graduates of Florida state-approved teacher preparation programs who have passed all three portions of the Florida Teacher Certification Examination will qualify for a Professional Florida Educator's Certificate. Those graduates of approved programs who have failed one or more of the three portions of the FTCE will receive a Temporary Certificate, which is valid for three school years. Graduates of approved out-of-state teacher preparation programs can obtain a Temporary Certificate which gives them three years in which to pass the FTCE.

A fee of \$25 is normally charged for taking each of the three FTCE examinations.

Using Students' Test Scores to Evaluate Teachers

Using students' scores on standardized achievement tests to evaluate their teacher is the new and troubling innovation in the accountability movement. In this method of evaluation, the beginning-of-year to end-of-year gain for students on a standardized achievement test is attributed solely to the efforts and ability of the students' teacher. Often a target gain is set, for example, 1.0 Grade Equivalent year's increase across the course of the academic year, and teachers are rewarded with merit pay increases for meeting or exceeding the targeted gain or punished in various ways for failing to meet the

target. This logic is very appealing to politicians or a general public that knows little about the complexities of teaching, learning and measurement of achievement. And indeed, it has found its way into Florida statutes.

The idea that a test score gain can be attributed to a particular teacher's efforts and abilities is often referred to as the "value-added" approach to teacher evaluation: what value does this teacher add to the learning of the students in his or her class? The principal purveyor of services in the area of value added teacher evaluation is the Tennessee Value Added Assessment System (TVAAS) Center at the University of Tennessee under the direction of Professor William L. Sanders. Sanders, who holds an earned doctorate in biostatistics and quantitative genetics and who worked at the Oak Ridge National Laboratory before taking over a statistical analysis center for agricultural research at the University of Tennessee, is the originator of a measurement and statistical analysis system that promises to measure validly and reliably the value that teachers add to the performance of the students in their charge. The TVAAS has been adopted or is being experimented with in twenty states across the U. S. including Colorado, Ohio, and Pennsylvania. The developers of the TVAAS claim that the quantitative measure that their technique produces is not confounded with the students' general level of aptitude, nor the contribution to their current learning of other teachers' efforts in prior years, the efforts of parents guiding the learning of their children outside of school, and many other factors that common sense suggests influence children's performance on tests.

Sanders made trips to Florida in the late 1990s to promote his system of teacher evaluation. In an interview with the Heartland Institute, Sanders remarked, "Several states are discussing it [the TVAAS model]. The state of Florida has enacted legislation,

as I understand it, to move to a value-added or ‘gain’ model in about 2001.”² Highly placed politicians found his logic persuasive. "I think you're going to see more interest in this," said Sen. Anna Cowin, R-Leesburg, chair of the Florida Senate's education committee, who had heard Sanders speak. “Accountability is so important. And to take it down to the individual teacher level -- it's very exciting.”³

The thinking behind the TVAAS system eventually made its way into the Florida State Statutes (K-20 Education Code: 1012.34 Assessment procedures and criteria) in the following form:

- (3) The assessment procedure for instructional personnel and school administrators must be primarily based on the performance of students assigned to their classrooms or schools, as appropriate. The procedures must comply with, but are not limited to, the following requirements:
 - (a) An assessment must be conducted for each employee at least once a year. The assessment must be based upon sound educational principles and contemporary research in effective educational practices. The assessment must primarily use data and indicators of improvement in student performance assessed annually as specified in s. 1008.22 and may consider results of peer reviews in evaluating the employee’s performance. Student performance must be measured by state assessments required under s. 1008.22 and by local assessments for subjects and grade levels not measured by the state assessment program. The assessment criteria must include, but are not limited to, indicators that relate to the following:
 1. Performance of students.

2. Ability to maintain appropriate discipline.
3. Knowledge of subject matter. The district school board shall make special provisions for evaluating teachers who are assigned to teach out-of-field.
4. Ability to plan and deliver instruction, including the use of technology in the classroom.
5. Ability to evaluate instructional needs.
6. Ability to establish and maintain a positive collaborative relationship with students' families to increase student achievement.
7. Other professional competencies, responsibilities, and requirements as established by rules of the State Board of Education and policies of the district school board.

Florida teachers have generally reacted negatively to the plan to evaluate them based in large part on their students' test performance: Jade Moore, executive director of the Pinellas Classroom Teachers Association, remarked, "It's a bad pay system based on a bad set of criteria." Moore was making reference to the use of students' FCAT scores to evaluate their teachers' performance. An article in the *St. Petersburg Times* for April 3, 2003, went on to report more teachers' reactions: "Despite the general resistance, some teachers are participating. 'I don't support the concept, but I have signed up for it,' said Missy Keller, president-elect of the teachers union in Hernando County Keller considers the program something of a gimmick."⁴

As will be shown below, expert opinion on the validity of the TVAAS value-added approach is substantially at variance with the claims made by its backers.

Section 3: Available Data

Teacher testing

Requiring candidates to take paper-and-pencil tests in the subject they teach or in general teaching methods is increasingly popular in state legislation for initial certification and re-certification. Performance tests—as opposed to paper-and-pencil tests—of teaching ability are sometimes talked about but virtually unheard of in state-mandated certification requirements. The cost is simply too great. Performance tests are a part of the National Board Certification procedure for teachers, but this approach is so time-consuming and expensive that few teachers can afford to take the test.

NCS (now known as Pearson Educational Measurement since being acquired by the publishing and consulting firm of Pearson Education⁵) is the big contractor in the area of paper-and-pencil teacher testing. The major concern with this approach is that of *test validity*. Just as they have questioned the National Teacher Examination (NTE) which was created and administered by the Educational Testing Service, experts in measurement and testing have questioned the validity of paper-and-pencil tests of teaching ability. Does the paper-and-pencil test score correlate with or predict teaching performance? Doubts and claims can be heard from both sides of the debate, yet solid, believable validity studies are infrequent.

Validity investigations of teachers' performance on the subject matter tests of the National Teacher Examinations (NTE) have failed to discover any consistent relationship between these tests of subject matter knowledge and teacher performance in terms of student achievement or supervisors' ratings. Most studies report statistically insignificant relationships, both positive and negative.⁶ Ashton and Crocker⁷ reported that five of 14

studies produced a positive correlation between measures of subject matter knowledge and teacher performance as measured by supervisors' ratings and student achievement. Madaus and Mehrens, two measurement experts both strongly inclined to support the use of tests in many areas of education, summarized their discussion of the limitations of paper-and-pencil tests for teacher certification: "...passing a multiple-choice test does not ensure that one will be a good teacher—or necessarily even a minimally competent one."⁸

In spite of attempts to remove "racial or ethnic group bias" from these tests, these tests still show substantial differences among ethnic groups with minority teachers scoring lower than white majority teachers. The panels which claim removal of test bias are little more than small groups of teachers acting as judges and nominating tiny numbers of test questions as being offensive. Such approaches fail to address the fundamental problem: ethnic minorities score much lower on paper-and-pencil tests than they would on peer or supervisor evaluations of their teaching performance.

Paper-and-pencil teacher testing has one other significant drawback. Any such selection test must have what is called a "cut-score," i.e., the score on the test that separates those who are selected from those who are rejected.⁹ Experience has shown that such cut-scores can not be determined non-arbitrarily—nor with adequate agreement among those experts whose judgments are collected in the process of setting the cut-score. The result is potential serious embarrassment if disgruntled test takers dig behind the test development documentation and discover this serious deficiency; lawsuits are certain to result. The testing companies and education agencies that take the responsibility of setting these cut-scores, in fact, refuse to release data that reveal the wide disagreement among judges

charged with the task of setting the pass scores. They act as if there is something to hide in this process, and they are correct.

Using Students' Test Scores to Evaluate Teachers

Several shortcomings of such approaches are clear:

1. Standardized achievement tests in many subjects are non-existent at both the elementary and secondary school levels: history, many of the sciences—not to mention a long list of subjects such as graphic arts, vocational education, physical education, music, and the like. How are teachers of these subjects to be evaluated by the “value-added” schemes?
2. Attributing gains in achievement made by a group of students solely to the efforts and skill of a single teacher or even the teacher who currently has these students in class ignores the reality of schools and classrooms. Secondary school students, for example, have many teachers, and students learn mathematics in their physics course and writing in their history course. At the elementary school level, a student’s progress in grade 3 may very well have a lot to do with the teaching of that student’s second grade teacher.
3. Teacher evaluation approaches that focus so heavily on standardized testing are in jeopardy of elevating a paper-and-pencil test to the level of the entire curriculum itself. Value-added methods of teacher evaluation are a form of high-stakes testing which has been shown to overemphasize not just the content but even the style of particular standardized tests to the detriment of a comprehensive and exemplary curriculum.¹⁰

Research has shown that when these value-added methods of teacher evaluation are implemented, certain consequences tend to ensue¹¹:

1. Evaluation is immediately shifted from the individual teacher to all teachers in the school building because of the absence of achievement tests in many subject areas and the interdependence of many teachers' efforts in the education of the students. Consequently, achievement gain targets are set for schools as a whole, not for individual teachers. Nonetheless, teachers of basic academic subjects (reading, writing and math at the elementary school level) end up carrying the load for the entire school.
2. Curriculum beyond the "basic skills" is given short shrift; teaching in science, social studies, not to mention music, art, health and the like, is shortened or eliminated entirely from the school day.
3. Teachers and administrators both are apt to succumb to the pressure of a system they view as illegitimate and engage in distortion or outright dishonesty in their attempts to cope with such a system.

Complete treatments of the TVAAS methods in the published literature are difficult to come by. In spite of the vigorous marketing of this method to state education agencies and the enthusiastic reception it has received by politicians and policy makers, some twenty years after its introduction, only two expositions of the statistical assumptions and techniques can be found in peer-reviewed academic journals.¹²

Recently, Haggai Kupermintz, a statistician and educational measurement expert at the University of Haifa in Israel, published a penetrating critique of the TVAAS.¹³ Kupermintz pointed to several logical and empirical weaknesses in the TVAAS system

and underscored the need for validity studies of the system that are currently lacking. For example, Kupermintz pointed out that Sanders' own attempt to report a "validity" study of the TVAAS was, in fact, based on a circular definition of teacher effectiveness and provided no independent evidence of the validity of the system at all. Kupermintz also points out how TVAAS estimated *teacher effects* (the technical name for the value added by a teacher) are constrained to add up to a fixed constant within a school system. Consequently, a teacher whose students make much bigger gains in a very high achieving school system will receive a lower value-added score than a teacher whose students learned less across the course of the school year but who teaches in a low-achieving school system. An issue of fundamental fairness thus arises.

Kupermintz also criticized the TVAAS approach for ignoring the interdependence of teaching in the typical school:

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... Attempts to disentangle such complex, interwoven contributions of the science teacher, [and] the math teacher ... into independent "effects" are not only methodologically intractable but also conceptually misguided.¹⁴

When questioned about the capability of the TVAAS system to control for differing levels of student “inputs,” such as intelligence, a key member of the TVAAS Center staff evidenced surprising naiveté concerning the psychology of individual differences. The following hypothetical was posed to the staff member: “Imagine two third-grade classes of 25 pupils each being taught by identical twin teachers who are alike in every respect; imagine that these two teachers teach the entire year in identical ways; but further imagine that all 25 children in one class have a measured intelligence of 130 and that all 25 pupils in the other class have a measured intelligence of 85. Does your approach assume that both teachers will receive identical value-added scores at the end of the school year?” The staff member’s answer to this question was, surprisingly, “Yes.”¹⁵ Clearly, the architects of the TVAAS do not understand the workings of individual differences that lie outside the control of teachers and schools. And they fail to appreciate the fact that prior years’ progress on achievement tests is not a pure measure of intellectual ability. TVAAS fails to control for differences among classes in intellectual ability when attributing value added by teachers.

In 1995, Thomas Fisher, Director of the Student Assessment Services Section of the Florida Department of Education, was asked to evaluate the Tennessee Value-Added Assessment System by the Comptroller of the State of Tennessee. His report, submitted in January, 1996, is available from the Office of Education Accountability division of the State Comptroller’s Office.¹⁶ Fisher was candid and highly critical of the TVAAS model. He wrote, “The value-added system cannot make determination of which teacher contributed how much to student’s skill.”¹⁷ He continued, “I do not support use of the value-added system for this purpose. I do not support giving the

teacher-level value-added information to the school superintendent and school board members because of potential for misuse and denial of due process rights to the individual teachers.”¹⁸ Fisher’s conclusion contained an ominous warning:

Last, one must remember that the question of evaluation of teachers is not a matter simply of educational research and statistical methodology. It involves an individual’s protected interests in employment. These are rights that cannot be challenged without due process. ... Ours is a litigious society, and I suspect that teachers will consider legal action if they believe the evaluation system is irrational or arbitrary.¹⁹

The Office of Education Accountability of the Tennessee Comptroller’s Office also contracted with R. Darrel Bock and Richard Wolfe, statistics and measurement experts affiliated with the University of Chicago and the Ontario Institute for Studies in Education, respectively, to evaluate the TVAAS value-added model from a statistical perspective. Bock and Wolfe concluded:

The most unusual aspect of the TVAAS formulation is in the definition of the teacher gains: they do not represent just students’ average gain during the year of the teacher’s instruction, but extend beyond to following years when the students are taught by other teachers. They are coded in the model in a form described as ‘layered.’ In effect, the gain attributed to any given teacher can represent gain from the previous year to the average of the current year and up to three subsequent years. No clear rationale for this convention is given in the description of the methodology.²⁰

Bock and Wolfe continue:

The TVAAS model represents teachers' contributions to gains, not in terms of difference between students' achievement scores the previous year and the teacher's current year, but as difference between the previous year and the teacher's current year and two following years. Inasmuch as the teacher is not directly responsible for student gains in those following two years, we believe this feature is inconsistent with the basic principle of the value-added assessment system.²¹

A report entitled *The Measure of Education: A Review of the Tennessee Value Added Assessment System* by Baker and Xu that is highly critical of the TVAAS system was published by the Comptroller's Office of the State of Tennessee in 1995. Its conclusions led to the commissioning of the reports by Fisher and by Bock and Wolfe. Its findings, however, were based on its own independent investigations since it preceded both the Fisher and the Bock and Wolfe reports.²² Among its conclusions are these:

1. "Because of unexplained variability in national norm gains across grade levels, it is not clear that those scores are the best benchmark by which to judge Tennessee educators."
2. "There are large changes in value-added scores from year to year, and teachers and administrators have been unable to explain those variations. As a result, the model may not help identify superior educational methods to the extent policymakers had hoped."

3. “The factors affecting student academic gain have not been identified, yet the model infers teacher, school, and district effect on student academic gain from the results of the value-added process.”
4. “The ‘high stakes’ nature of the state-administered test may create unintended incentives for both educators and students.”²³

Baker and Xu’s report goes on to describe the case of Scotts Hill School, which just happens to be situated on the county line separating Henderson and Decatur counties. The TVAAS assessment of Scotts Hill School actually measured the school’s “value-added” contribution to students’ achievement twice: once as though it were a school in Henderson County and again as though it were a school in Decatur County. Since the expected gains for a school are based in part on the performance of students in the entire system of which that school is a part, Scotts Hill School received two measures of value added. Surprisingly, the two measures were substantially different. No adequate explanation of this anomaly was advanced by the TVAAS staff.

The Tennessee Comptroller’s Report ended with three recommendations:

1. The report recommends that all components of the TVAAS be evaluated by qualified experts knowledgeable of statistics, educational measurement, and testing. This recommendation led to the Bock, Wolfe and Fisher reports.
2. The Department of Audit should perform an Information Systems Assessment to evaluate VARAC’s [Value Added Research Assessment Center] documentation practices and assess the safety and security of the TVAAS. The state needs assurance that reasonable operational procedures are in place to protect the hardware, software, and data.

3. The State Board of Education and the State Department of Education need to identify unintended incentives for educators and students and consider ways to reduce their likelihood.²⁴

Why, one might reasonably ask, is this brief spending so much time critiquing the Tennessee Value-Added approach to teacher evaluation when that approach has not been purchased by the Florida Department of Education nor any other major school district in Florida, nor is it mandated by K-20 Education Code: 1012.34 “Assessment procedures and criteria,” which merely says seemingly innocuously that “the assessment procedure for instructional personnel and school administrators must be primarily based on the performance of students...”? The answer lies in the relationship between the TVAAS approach and simpler methods of attempting to attribute student achievement gains to their teachers. Less complex, and often used, methods of measuring teachers’ impact on students’ achievement employ simple gain scores (June performance minus September performance on standardized tests) or worse (deviations in grade-equivalent scores between the average performance of a class and the grade level expectation, for example). The TVAAS value-added technique with its three-year data streams and complex statistical corrections is substantially better than these crude measures of teachers’ effect, and yet *it is clearly inadequate*. So much the worse for more simple techniques. In fact, all of the shortcomings and more that are now coming to light with respect to the measurement of Adequate Yearly Progress as mandated in federal No Child Left Behind legislation are present in the TVAAS system and its simpler alternatives.²⁵

Section 4: Quality of Available Data

Teacher Testing

Much is known about the validity of paper-and-pencil tests for teacher certification. There is no imperative for new research in this area.

Using Students' Test Scores to Evaluate Teachers

Much research is needed concerning the properties of value-added assessment techniques. Unfortunately, those in the best position to share data with the research community that would illuminate many of the issues surrounding this approach have proved to be uncooperative. As Kupermintz pointed out in his critique of TVAAS, "In order to enable a proper validity investigation, TVAAS data must be made available to interested, qualified researchers. To date, numerous requests by the author for access to the TVAAS data have been met with blanket 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 ... and organizations such as the Carnegie Foundation have requested data directly from Sanders only to be turned down or stalled."²⁶ Such actions on the part of scholars and employees of public institutions are inconsistent with the values of and standards for responsible professional practice.

Section 5: Recommendations

Several recommendations at the state-wide policy level can be derived from the above consideration of the issues surrounding teacher evaluation in the State of Florida.

1. Any attempt to substitute test performance for college degree requirements in the teacher certification process should be opposed. Movements in this direction can be discerned in the legislatures in several states. Such policies would surely result in a less skilled and less professional teaching corps. Furthermore, the questionable validity of paper-and-pencil tests can not support such practices.

Certification standards for out-of-state teachers are currently less stringent than for graduates of approved in-state programs of teacher preparation. On account of reciprocity agreements with other states and the issuance of temporary teaching certificates to graduates of out-of-state teacher preparation programs, in-state graduates face a more daunting row of hurdles to certification (because of an additional entrance examination—the College Level Academic Skills Test—required to *enter* an approved preparation program) than out-of-state graduates. Holders of temporary certificates have three years in which to pass the FTCE tests.

2. Value-added teacher evaluation methods, which attempt to evaluate teachers in terms of the standardized achievement test score gains of their students, are of uncertain validity, have drawn heavy criticism from measurement experts, and raise serious concerns about fairness. They should be opposed in their various forms. References in current statutes (K-20 Education Code: 1012.34 “Assessment procedures and criteria”) such as “The assessment procedure for instructional personnel and school administrators must be primarily based on the performance of students assigned to their classrooms or schools” should

be removed from legislation because no method of validly and fairly attributing student test performance to individual teachers or administrators is presently available.

Notes and References

¹ The authoritative reference on the methods, practices and policy concerning teacher evaluation is Millman, J. & Darling-Hammond, L. (Eds.) *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: SAGE Publications. Of particular relevance to the issues discussed in this brief are the following chapters:

Chapter 4. Sykes, G. Licensure and certification of teachers: An appraisal.

Chapter 5. Scriven, M. Teacher selection.

Chapter 12. Good, T. L. & Mulryan, C. Teacher ratings: A call for teacher control and self-evaluation.

Chapter 14. Glass, G. V. Using student test scores to evaluate teachers.

Chapter 16. Madaus, G. & Mehrens, W. A. Conventional tests for licensure.

Chapter 18. Jaeger, R. M. Setting standards on teacher certification tests.

² The interview from 1999 with William Sanders by George Clowes of the Heartland Institute, a think tank located in Chicago, IL, is available at <http://www.heartland.org/archives/education/nov99/sanders.htm>.

³ Hegarty, S. (1999, January 17). Schools grading plan uses new tack: A Tennessee professor of statistics says his system examines students' improvement over time. *St. Petersburg Times*. Retrieved February 1, 2004, from http://www.shearonforschools.com/st_petersburg_01171999.htm

⁴ Hegarty, S. (2003, April 3). Teachers not buying state's performance bonus program: Some may find the program divisive. Others think teachers simply should be paid more. *St. Petersburg Times*. Retrieved February 1, 2004 from http://www.sptimes.com/2003/04/03/State/Teachers_not_buying_s.shtml

⁵ See the company's website at <http://www.pearsonedmeasurement.com/>.

⁶ The following authors provide consistent evidence of the lack of validity of paper-and-pencil tests for predicting teachers' success as seen by peers and supervisors.

Andrews, J.W., Blackmon, C.R., and Mackey, J.A. (1980). Preservice performance and the National Teacher Examinations. *Phi Delta Kappan*, 61(5), 358-359.

Ayers, J.B., and Qualls, G.S. (Nov/Dec 1979). Concurrent and predictive validity of the National Teacher Examinations. *Journal of Educational Research*, 73 (2), 86-92.

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: testing teachers for the improvement of American education. Pp. 169-238 in E.Z. Rothkopf (Ed.) *Review of Research in Education*, Vol. 14. Washington, D.C.: American Educational Research Association.

Quirk, T.J., Witten, B.J., & Weinberg, S.F. (1973). Review of studies of concurrent and predictive validity of the National Teacher Examinations. *Review of Educational Research*, 43, 89-114.

Summers, A.A., & Wolfe, B.L. (1975, February). Which School Resources Help Learning? Efficiency and Equality in Philadelphia Public Schools. Philadelphia, PA: ERIC Document ED 102 716.

For an excellent summary of this entire line of research, see Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. Education Policy Analysis Archives, 8(1). Retrieved February 1, 2004 from <http://epaa.asu.edu/epaa/v8n1/>.

⁷ Ashton, P. & Crocker, L. (1987, May-June). Systematic study of planned variations: The essential focus of teacher education reform. *Journal of Teacher Education*, 38, 2-8.

⁸ Page 260 in Madaus, G. & Mehrens, W. A. (1990). Conventional tests for licensure. Pp. 257-77 in Millman, J. & Darling-Hammond, L. (Eds.) *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: SAGE Publications.

⁹ On the controversy surrounding the setting of cut-scores on all kinds of paper-and-pencil tests see the following references:

Jaeger, R. M. (1990). Setting standards on teacher certification tests. Pp. 295-321 in Millman, J. & Darling-Hammond, L. (Eds.) *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: SAGE Publications.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-61. Also available online under the title "Standards and criteria Redux" at <http://glass.ed.asu.edu/gene/papers/standards/>.

Glass, G. V. (2003). Cut-Scores: Where Do They Come From? Chapter 5, pp. 145-162 in Boston, C.; Rudner, L. M.; Walker, L. J.; & Crouch, L. (Eds.) *What Reporters Need To Know About Test Scores*. Washington, D. C.: Education Writers Association.

¹⁰ McNeil, L. M. (2000). Contradictions of school reform: Educational costs of standardized testing. New York: Routledge.

Amrein, A.L. & Berliner, D.C. (2002, March 28). High-stakes testing, uncertainty, and student learning Education Policy Analysis Archives, 10(18). Retrieved February 1, 2004 from <http://epaa.asu.edu/epaa/v10n18/>.

¹¹ Glass, G. V. (1990). *Op cit*.

¹² Sanders, W. L. & Horn, S. P. (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), 247-56.

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

¹³ Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, 25(3), 287-298. Also see Kupermintz, H. (2002). Value-Added Assessment of Teachers, Chapter 11 in

Molnar, A. (Ed.) *School Reform Proposals: The Research Evidence*. Greenwich, CT: Information Age Publishing, Inc.

¹⁴ Kupermintz, H. (2003). Op cit. P. 290.

¹⁵ Personal communication, S. P. Horn, July 17, 1998.

¹⁶ Fisher, T. H. (1996). A review and analysis of the Tennessee Value-Added Assessment System, Part 2. Office of Education Accountability, Comptroller of the Treasury, State of Tennessee. Retrieved February 1, 2004, from <http://www.comptroller.state.tn.us/orea/reports/tvaascp2.pdf> .

¹⁷ *Ibid.* P. 46.

¹⁸ *Ibid.* P. 46

¹⁹ *Ibid.* P. 47.

²⁰ Bock, R. D. & Wolfe, R. (1996). A review and analysis of the Tennessee Value-Added Assessment System, Part 1. Office of Education Accountability, Comptroller of the Treasury, State of Tennessee. Retrieved February 1, 2004, from <http://www.comptroller.state.tn.us/orea/reports/tvaascp1.pdf> . P. 70.

²¹ *Ibid.* P. 70.

²² Baker, A. P. & Xu, D. (1995). The measure of education: A review of the Tennessee Value Added Assessment System. Nashville, TN: Office of Education Accountability, Comptroller of the Treasury, State of Tennessee. Retrieved February 1, 2004, from <http://www.comptroller.state.tn.us/orea/reports/tvaas.pdf> .

²³ *Ibid.* Pp. i-iii.

²⁴ *Ibid.* P. iv.

²⁵ See Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 32(7), 3-13.

²⁶ Kupermintz, H. (2003). Op cit. P. 297.