



Texas AFT

A Union of Professionals

www.texasaft.org

3000 S. IH-35
Suite 175
Austin, TX 78704

(512) 448-0130
(800) 222-3827
f: (512)448-0678

TESTIMONY OF ERIC HARTMAN
ON BEHALF OF TEXAS AFT
REGARDING SENATE BILL 4
BEFORE THE SENATE EDUCATION COMMITTEE
MARCH 29, 2011

Teachers are the first to tell you that evaluation practices need improvement. Drive-by evaluation, based on nothing more than momentary observation, seldom produces meaningful guidance to improve teaching.

But neither does test-driven evaluation based on error-prone analysis of standardized achievement-test results and used as the basis for high-stakes decisions regarding teachers. Unfortunately, that is precisely the error-prone system SB 4 would establish. We already have an educational system that places extreme, excessive emphasis on students' scores on standardized tests. SB 4 as introduced would reinforce that undue emphasis.

SB 4 says (in SECTION 6): "The recommended appraisal system must provide that at least 50 percent of the appraisal is based on a measurement of the teacher's effectiveness and the annual improvement of the teacher's students' achievement." It goes on: "the recommended appraisal process of teacher with primary instructional responsibilities in a subject and grade for which assessment is administered under [the state accountability system] must provide that the 50 percent measurement is based on standardized achievement data of the teacher's students...."

A July 2010 study by Mathematica researchers for the Institute of Education Sciences at the U.S. Department of Education found that use of three years of test-score data to measure teacher performance in math and reading is highly susceptible to error. They found that "in a typical performance measurement system, 1 in 4 teachers will be erroneously identified" as a substandard performer. With just one year of data, the error rate climbed to 35 percent.

These Mathematica researchers said their analysis actually understated error rates, because the predominant source of variation in student gain scores "is due to the variation in *student-level* factors that are not under the control of the teacher." They cautioned against basing high-stakes decisions regarding teachers on such shaky ground. The evidence led them to conclude it would be more appropriate to use test-score gains and value-added methodologies as no more than a preliminary indicator that could prompt closer inspection of a teacher's performance.

This recent study reflects a strong scholarly consensus among educational researchers against use of value-added methodologies for high-stakes decisions. For example:

--Rand Education researchers in "The Promise and Peril of Using Value-Added Modeling to Measure Teacher Effectiveness" (2004): "The research base is currently insufficient for us to recommend the use of value-added methodology for high-stakes decisions regarding employee evaluation and compensation." The Rand research team comprehensively examined the potential of various value-added models to serve as a basis for measuring school and teacher effectiveness. They found that no current value-added methodologies can account for the many variables beyond the current-year teacher's performance that influence student achievement: for example, controlling for student backgrounds, and the criteria for linking particular students to particular teachers.

--Professor Audrey Amrein-Beardsley, Arizona State University (2008): Analyzing the value-added methodology of William Sanders, the agricultural statistician who has helped design the Houston ISD value-added scheme, Professor Amrein-Beardsley found that Sanders' methodology is opaque and has not been independently tested enough to justify its widespread use. "The problem," Amrein-Beardsley says, "is that when these things are being sold to superintendents, they don't know any better."

--Professor Jesse Rothstein, Princeton University (2009): Professor Rothstein's study of the value-added methodology used in North Carolina found that it, like most value-added models, is based on a false assumption: namely, that students are randomly assigned to teachers. The reality is that a teacher's apparent success or failure as measured by value-added methodology can be decisively influenced by anything-but-random assignment of particular students to particular teachers. Rothstein also found that "accountability policies that rely on measures of short-term value added would do an extremely poor job of rewarding the teachers who are best for students' longer-run outcomes."

--National Academy of Sciences, Board on Testing and Assessment (2009): In a letter expressing strong concerns about the direction of federal policy under the so-called "Race to the Top" grant program, scholars at the National Academy of Sciences said the federal proposal "places too much emphasis on measures of growth in student achievement (1) that have not yet been adequately studied for the purposes of evaluating teachers and principals and (2) that face substantial practical barriers to being successfully deployed in an operational personnel system that is fair, reliable, and valid."

The National Academy of Sciences testing experts added: "The term 'value-added model' (VAM) has been applied to a range of approaches, varying in their data requirements and statistical complexity. Although the idea has intuitive appeal, a great deal is unknown about the potential and the limitations of alternative statistical models for evaluating teachers' value-added contributions to student learning. [The Board on Testing and Assessment] agrees with other experts who have urged the need for caution and for further research prior to any large-scale, high-stakes reliance on these approaches."

In light of all this evidence, Texas AFT therefore opposes any top-down mandate from the state to put any substantial weight on test-score gain data and value-added methodologies in teacher

evaluation. But we do believe that the state can be the catalyst for improvement in teacher evaluation and teacher effectiveness.

In fact, we take our cue on this point from Dr. Lauren Resnick, a scholar from the University of Pittsburgh that you invited to testify before this committee during the interim. Dr. Resnick advised against a premature, top-down state mandate of a new evaluation system. We need first to figure out what works, she said. The best way forward would be for the state to set some guidelines and encourage local trials of alternative approaches to the evaluation and improvement of teacher effectiveness.

The state could provide incentives for this work, and a crucial part of the deal with local districts would be for them to agree to (a) deep involvement of teachers and other stakeholders in the design of the local plan and (b) rigorous, impartial evaluation of the results of their local trials to determine what works and what does not.

We strongly encourage the committee to proceed in this direction. For a closer look at what we are suggesting, we recommend that you take a look at HB 3058 by Rep. Mike Villarreal of San Antonio, which was partly inspired by the Symposium on Teacher Effectiveness that we held here in Austin last month. (We are grateful, incidentally, that a number of you were represented by your education policy staffers at that meeting.)

Another witness today from the ranks of Texas AFT will tell you about the pioneering work of our national affiliate, the American Federation of Teachers, and many local affiliates around the country in collaboration with school districts on exactly the type of local trials of alternative approaches to teacher evaluation and improvement that we are proposing. We also ask you to take note of the comments attached to this written testimony by Andy Dewey, a teacher from Houston ISD who chairs the Texas AFT committee on professional standards (Committee on Quality Educational Standards in Teaching, or QuEST).

We want to close by acknowledging that we see some promising elements in SB 4. The provision on building the “science of reading” component into the EC/4 certification exam looks to us like a good idea. Emphasizing individualized professional development is another good idea, but here we would want to see a collaboration between teacher and principal on the creation of the plan, and we also would note that this should be part of a continuous process of teacher development with necessary supports. We also assure you we stand ready and willing to continue working with the chair and members of the committee on productive approaches to this important task of improving teacher evaluation and effectiveness.

Comments Regarding SB 4 by Andy Dewey, Houston ISD Teacher and Chair of Texas AFT Committee on Quality Educational Standards in Teaching (QuEST), March 29, 2011:

When designing a teacher evaluation system it is important to know beforehand the purposes of that evaluation. An effective teacher evaluation can serve multiple purposes. It can be used to:

- Improve the overall quality of the teacher workforce by identifying and building upon individual and collective teacher strengths, and by improving instruction and other teacher practices to improve student learning;
- Identify exemplary teachers who might serve as mentor and/or master teachers;
- Identify ineffective teachers and develop a system of support to remediate their skills;
- Ensure fair and valid employment decisions, including decisions about hiring, dismissal, and career paths.

The evaluation instrument itself must contain some core components based on valid and reliable professional teaching standards. These components must include:

- **Instructional Practices:** Observable classroom performance. Is the teacher utilizing methodologies and approaches that can be reasonably expected to result in increased student learning?
- **Professional Expectations:** Observable. Does the teacher do the things that an effective teacher is expected to do? These things may include parental contact, lesson planning, professional development, or teacher mentoring.
- **Student learning:** Data based. Multiple measures of student learning outcomes should be included in the evaluation. These may include teacher generated assessments, department or school generated assessments, or district generated assessments. Standardized tests, if used, must be 100% curriculum aligned. The teacher must know what will be tested and for what he or she will be held accountable. Standardized tests that are currently in use have never been validated as a measure of teacher effectiveness and should never be used as a predominant or even a significant portion of student learning data. Certainly 50% of the entire evaluation as mandated by SB 4 places way to much weight on this measure. Value added data is susceptible to error due to the inherent limitations of tests for capturing the complexity and breadth of student learning and the limitations of value-added methodologies. There is solid published research stating that value added scores are susceptible to a 35% error rate in a single year, lowering to a 25% error rate over three years. This means that in any given year value added data will tell us that a teacher can be anywhere in a continuum from ineffective to effective but it cannot be reliably more precise than that. Because of its unreliability and inaccuracy value added data should never be used in teacher evaluation.
- **Professional Development:** Observable. In order for a teacher evaluation to be meaningful it not only must identify areas in need of improvement it must direct

that teacher towards meaningful professional development in order to remedy deficiencies or to improve upon what is already being done well. Professional growth must be viewed as a standard of teaching and not as a due process requirement on the way towards eventual termination. Professional development should be geared towards improving instruction and thus student learning. This is why the use of standardized tests and value added data provides such a challenge in developing professional development plans. There is no way to specifically identify which instructional practices improve test scores. A teacher may be evaluated as a great teacher based on observation but test scores may be low. Conversely students of a teacher with low observable ratings may do well on tests. Unless the evaluation can specifically identify which instructional practices directly improve test scores and which practices do not work, no meaningful professional development plan can be followed. In short the evaluator must be able to say: "Your student scores were low for this specific reason, this is what you must do to raise the scores." The evaluator would then monitor the progress of the professional development or instructional strategies to ensure success.

SB 4 attempts to solve a complex problem in a relatively simplistic way. By placing such a strong emphasis on standardized tests to define effective teaching, the proposed legislation undervalues the art of teaching: the daily interaction between the student and teacher, the delivery of instruction and the constant, almost minute-by-minute assessment of the effectiveness of that instruction that teachers must do.

Testing also does not recognize the context of the classroom, the school, or the community. Does a teacher have opportunities to meet the needs of his or her students, given school schedules, noninstructional duties, paperwork and availability (or inaccessibility) of structured avenues for collaboration with colleagues such as common planning time, lesson study and professional learning communities? Do teachers have access to facilities and resources, meaning the people, materials and tools they need to support their teaching?

Context also refers to the extent to which their schools are well-maintained and have adequate environmental conditions such as space, lighting and ventilation. Is the school climate physically safe? Is the professional climate safe or do teachers feel that experimentation, innovation, or in any way deviating from the norm will lead to administrative discipline? These are just some of the things that may affect standardized test scores but do not mean the teacher is in any way inferior.

A comprehensive evaluation system that continuously improves instruction, professional development, and student learning is a goal of all educators. A system that places too much emphasis on instruments with very narrow objectives and measurement tools will not meet the needs of children, schools, teachers, or the public as a whole.