Texas Classroom Teachers Association

P.O. Box 1489
Austin, TX 78767

512-477-9415
1-888-879-8282
Fax 512-469-9527
http://www.tcta.org/

tcta.

**Testimony to the Senate Education Committee
Regarding SB 4
By Holly Eaton, Director of Professional Development and Advocacy
March 29, 2011**

Although SB 4 is a comprehensive teacher quality bill with components that we have no objections to and some we even support, we must regretfully oppose SB 4 based on its requirement that at least 50 percent of teachers' appraisals must be based on standardized achievement data including standardized state tests.

- Specifically, the bill requires that for teachers who teach state-assessed subjects, at least 50 percent of the teacher's appraisal must be based on standardized achievement data of the teacher's students including the state assessment. We have long been on record in opposition to basing an individual teacher's evaluation on standardized student test performance. **The evidence continues to mount that it is inappropriate and unsupported by the research to base teacher evaluation, compensation, or other high-stakes consequences on student test scores.**

- The bill essentially requires two different appraisal criteria for different teachers, which is likely to result in the state and local districts incurring significant additional costs in developing and administering such a system. **Given the fact, in reality, only those teachers responsible for reading and math in grades 3 to 8 are in a situation in which assessments are vertically aligned such that student growth can be shown from year to year (estimated at around only 20% of teachers), developing a whole system assigning evaluation ratings based on student standardized test performance seems a largely futile and expensive endeavor,** especially in this budgetary climate.

- The bill eliminates current provisions which allow teachers to be appraised less frequently if the teacher agrees in writing and the teacher's most recent evaluation rated the teacher as at least proficient and did not identify any area of deficiency. Instead, the bill requires all teachers to be evaluated at least twice each school year. Especially given tight budgetary times, and the resulting demand to focus resources where needed most, it seems prudent to retain current language which ensures that more focus is given to teachers who are new or who received a less than proficient evaluation in the prior year.

- Although we appreciate the direction the Senate is taking to try to fund the underlying support systems that are currently in place in order to help students succeed on the state assessment system (including the Student Success Initiative, textbooks, and related professional development), **the likelihood that the final budget will adequately fund these support systems is in serious question. That being the**

case, and faced with a new more rigorous assessment system, we do not believe that now is the time to *up* the stakes (for teachers or anyone else) based on standardized test results.

- Given the fact that so many things are in flux right now due to the budget situation, and the fact that districts already have the ability to modify the current state appraisal system to fit their needs under current law, we seriously question the timing of legislation requiring the state to implement a new state teacher appraisal system this legislative session.

Regarding the bill's requirement that at least 50% of a teacher's appraisal for a teacher teaching state-assessed subjects must be based on their students' standardized achievement including on state assessments, our concerns stem from the fact that as one expert put it, **no-matter how hard statisticians try, and no matter how good the data and statistical model, it is very difficult to separate a teacher's effect on student learning gains from other classroom effects, like peer effect (race and poverty of peer group).** Statistical models attempt to adjust away these differences, but usually come up short. Bruce Baker, associate professor in the Department of Educational Theory, Policy and Administration at Rutgers University Graduate School of Education in New Brunswick
http://www.northjersey.com/news/education/evaluation_031311.html?c=y&page=1

For example, basing teacher evaluations on student test scores does not accurately distinguish more from less effective teachers because even relatively sophisticated approaches cannot adequately address the full range of statistical problems that arise in estimating a teacher's effectiveness. Some of these statistical problems arise because of the influence of student socioeconomic advantage or disadvantage on learning, measurement error and instability, the nonrandom sorting of teachers across schools and of students to teachers in classrooms within schools, and the difficulty of disentangling the contributions of multiple teachers over time to students' learning. **As a result, reliance on student test scores for evaluating teachers is likely to misidentify many teachers as either poor or successful.** Even with the best models and data, teacher ratings are highly inconsistent from year to year, and have very high rates of misclassification. *Evidence about the use of test scores to evaluate teachers,*
http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf

A good illustration of this is a recent major study which found that there is **a 35% chance of identifying an average teacher as poor, given one year of data, and a 25% chance given three years**. *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains* http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Then you have issues about the fact that it's not possible to equate the difficulty of moving a group of children 5 points (or rank and percentile positions) at one end of a test scale to moving children 5 points at the other end. In such a system, the only fair way to compare one teacher to another would be to ensure that each has a randomly assigned group of children whose initial achievement is spread similarly across the testing scale. Real schools and districts don't work that way. Bruce Baker, associate professor in the Department of Educational Theory, Policy and Administration at Rutgers University Graduate School of Education in New Brunswick
http://www.northjersey.com/news/education/evaluation_031311.html?c=y&page=1

You also have the issue of statistical error with small group size. According to experts, the larger the number of students in a tested group, the smaller the average error will be because

positive errors will tend to cancel out negative errors. **But the sampling error associated with small classes of, say, 20-30 students could well be too large to generate reliable results. Most teachers, particularly those teaching elementary or middle school students, do not teach enough students in any year for average test scores to be highly reliable.** In schools with high mobility, the number of these students with scores at more than one point in time, so that gains can be measured, is smaller still. *Evidence about the use of test scores to evaluate teachers,* http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf.

That's the reason school districts aren't held accountable under our state accountability system for student subgroups of less than 30.

For these and other reasons, the research community has cautioned against the heavy reliance on test scores, even when sophisticated VAM methods are used, for high stakes decisions such as pay, evaluation, or tenure. *Evidence about the use of test scores to evaluate teachers,* http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf[1]

According to one expert, basing teacher evaluations and other consequential decisions on scores that may be influenced by which students a teacher serves provides a substantial disincentive for teachers to serve kids with the greatest needs, disruptive kids, or kids with disruptive family lives. Many of these factors are not, and cannot be captured by variables in the best models. Some have argued that including value-added metrics in teacher evaluation reduces the ability of school administrators to arbitrarily dismiss a teacher. Rather, use of these metrics provides new opportunities to sabotage a teacher's career through creative student assignment practices.

The same expert states that **the reality of an evaluation that includes a single large, or even significant weight, placed on a single quantified factor is that that specific factor necessarily becomes the tipping point, or trigger mechanism.** It may be 45% of the evaluation weight, but it becomes 100% of the decision, because it's a fixed, clearly defined (though poorly estimated) metric. Bruce Baker, associate professor in the Department of Educational Theory, Policy and Administration at Rutgers University Graduate School of Education in New Brunswick http://www.northjersey.com/news/education/evaluation_031311.html?c=y&page=1

In sum, we remain opposed to any linkage of student test scores to individual teachers for high stakes consequences due to a lack of research support for this concept's validity.

---

1. Growth measures implicitly assume, without justification, that students who begin at different achievement levels should be expected to gain at the same rate, and that all gains are due solely to the individual teacher to whom student scores are attached; growth measures do not control for students' socioeconomic advantages or disadvantages that may affect not only their initial levels but their learning rates. Although value-added approaches improve over these other methods, the claim that they can "level the playing field" and provide reliable, valid, and fair comparisons of individual teachers is overstated. **Even when student demographic characteristics are taken into account, the value-added measures are too unstable (i.e., vary widely) across time, across the classes that teachers teach, and across tests that are used to evaluate instruction, to be used for the high-stakes purposes of evaluating teachers.**[16]

16. Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journalof Economics*, 125 (1), January: 175-214; Newton, Xiaoxia, et al. (Linda Darling-Hammond, Edward Haertel, and Ewart Thomas). Forthcoming. *Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts;* Sass, Timothy. 2008. *The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy.* Washington, D.C.: CALDER. http://www.urban.org/uploadedpdf/1001266_stabilityofvalue.pdf; Lockwood, J. R., et al. (Daniel McCaffrey, Laura S. Hamilton, Brian Stetcher, Vi-Nhuan Le, and Felipe Martinez). 2007. The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44 (1), 47 – 67.

For a variety of reasons, analyses of VAM results have led researchers to doubt whether the methodology can accurately identify more and less effective teachers. VAM estimates have proven to be unstable across statistical models, years, and classes that teachers teach. One study found that across five large urban districts, among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40%. Sass 2008, citing Koedel and Betts 2007; McCaffrey et al. 2009. For similar findings, see Newton et al., forthcoming. Sass, Timothy. 2008. *The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy.* Washington, D.C.: CALDER. http://www.urban.org/uploadedpdf/ 1001266_stabilityofvalue.pdf.