INTRODUCTION

Teachers in the US have long been paid the same regardless of the effectiveness of the teacher. Indeed, two teachers with the same amount of experience are paid about the same amount of money regardless of how effective or ineffective the teachers. This remains true today despite a large body of evidence and years of parents' experiences in schools that tell us that some teachers are simply better than other teachers.
We know that in many professions, employees are paid based on their productivity and effectiveness, not necessarily on their years of experience. Since we know some teachers are more effective than others, common sense seems to dictate three policies to address this situation:

1) Grant tenure only to those teachers who are effective in improving student achievement; and,

2) Allow only effective teachers to remain tenured;

Who could argue with such common sense proposals? They seem as American as apple pie. Yet, there are two over-arching problems with these proposals.

First, we don't know how to measure teacher effectiveness very well at all; and,

Second, when we provide monetary incentive to teachers--such as remaining in a teaching position--to teachers to improve student achievement, such systems simply don't help to improve student achievement and often hurt it.

Each of these issues is addressed in more detail below.

MEASURURING TEACHER EFFECTIVENESS USING STUDENT ACHIEVEMENT

This part of the report is divided into 4 sections: immediate problems for the 2011-12 school year, evaluating teacher effectiveness using student achievement without value added measures, evaluating teacher effectiveness using student achievement with value added measures, and measuring teacher effectiveness through observations.

Immediate Problems for the 2011-12 School Year

One immediate problem is that measuring teacher effectiveness will start in the 2011-12 school year in SB 4(denoted as 2012 from now on). Yet, 2012 will be the first year of the STAAR test. Thus, the first problem encountered by TEA and districts in developing an evaluation process is that ***there will be no growth measure available for STAAR in 2011-12.***

Thus, neither TEA nor districts will have any type of growth measure available unless a sophisticated statistical value-added analysis is performed. TEA will likely be unable to do this at the teacher level because they will not have data that matches teachers to students for classrooms throughout the state. Even if they did have access to this data, the accuracy of the

data would vary from district to district and likely not be accurate enough to effectively measure teacher effectiveness. Further, only a handful of districts have the capacity to do this (e.g., Houston, Dallas, and Austin) and the current value-added models in these districts leave much room for improvement.

Options for Evaluating Teacher Effectiveness Using Student Achievement without Value-Added Measures

If no value-added estimates are available at the classroom level, districts will likely turn to one of four options in evaluating teacher effectiveness for teachers with standardized test scores:

(a) Average proficiency level;

(b) Average scale scores;

(c) Changes in proficiency levels; and/or,

(d) Changes in scale scores (growth).

Proficiency Level and Scale Scores

We know that using status measures (data at one point in time) such as average proficiency levels or average scale scores are incredibly inaccurate since passing rates and average scores are almost entirely dependent on the ability level and characteristics of the students. Indeed, any student passing rate or test score is highly correlated with the percentage of students in poverty and moderately correlated with the percentage of non-White students. If such measures are used to evaluate teachers, then teachers (especially provisional/beginning teachers) will have a large incentive to seek positions teaching relatively affluent and non-minority students.

With respect to ability level, every analysis of student test scores find that prior achievement is the best predictor of future achievement. Thus, teachers who are lucky enough to teach high achievers will appear more effective than they may actually be while teachers of low-achievers will appear less effective than they might actually be.

For example, suppose we have two teachers, teacher A and teacher B and that 80% of teacher A's students did not pass TAKS in 2010-11 while 90% of teacher B's students achieved commended status on the TAKS. Which teacher has a better chance at having a STAAR passing rate greater than the school, district, or state passing percentage? Teacher B, of course! But does that mean teacher B is more effective than teacher A? Certainly not!

Changes in Proficiency Levels

Many schools, districts, and states use changes in proficiency levels as a measure of student achievement. However, as Dr. Daniel Koretz (a highly regarded expert in testing and achievement) notes in his book *Measuring Up* (2008), ***changes in proficiency levels always provides an inaccurate indicator of student achievement***. Indeed, he clearly shows that using ***changes in proficiency levels as a measure of achievement and effectiveness will always result***

in the mis-identification of effective teachers, schools, districts, and states.
Yet, a perusal of district websites and press releases, TEA's website and press releases, and even legislators' press releases shows that this is the most commonly used metric to identify student achievement and "growth." Yet, the distribution of student scores around the cut score largely determines changes in the percentage of students passing.

For example, suppose 70% of students in both teacher C's class and teacher D's class have passed the test last year. This year, 80% of teacher C's students passed the test, but only 75% of teacher D's students passed the test. Many schools and districts would claim that teacher C is the more effective teacher. However, when examining the individual student scores, all of the 30% of the students not passing had scores within two questions of passing the test while all of the 30% of students not passing the test in teacher D's class had prior scores that were more than 5 questions below the passing cut score. In fact, teacher D's students made greater growth than teacher C's students, but they simply did not make enough growth to move over the passing cut score. This is precisely why schools and districts should examine the changes in the average scale scores or, better yet, a statistically computed growth measure.

Changes in Scale Scores/Growth Measure

Of the four options mentioned above, this option is clearly the best. Changes in the average scores are not as strongly correlated with student background characteristics as scores for one year, although there is still a fairly strong correlation. Further, the change in scale scores includes much more detailed information than the binary information provided by the percentage of students passing. Thus, the changes in average scale scores do not encounter the same problem as the changes in the average proficiency levels.

However, both changes in the average scale scores and growth measure are still correlated with student background characteristics (such as poverty status, parental level of education, disability status, and English Language Learner status, among others), classroom characteristics (percentage of high achieving or low achieving students, percentage of poor students, percentage of special education students, percentage of ELL students, classroom culture, etc), and school characteristics (percentage of high achieving or low achieving students, percentage of poor students, percentage of special education students, percentage of ELL students, school culture, etc). The remedy for these drawbacks is the use of value-added measures which are discussed below.

Problems Measuring Teacher Effectiveness Using Value-Added

There is no doubt that the use of value-added methodologies is substantially fairer and more accurate than relying on the results of the methodologies described above. Yet, there are still some very serious problems with using value-added methodologies when making high-stakes decisions. These are described in some detail below.

Lack of Random Assignment of Students

One primary purpose of this bill is to estimate a teacher's effectiveness on her/his student achievement. This requires making a causal inference about the teacher. This is not possible unless students are randomly assigned to teachers. But, students are NOT assigned randomly. When students are not randomly assigned, we cannot statistically control for the unequal characteristics of students across classrooms and culture of a classroom that is not under the control of the teacher.

Inconsistent Teacher Effectiveness Ratings

Even with close attention paid to creating the best possible models and ensuring the most accurate data, teacher ratings are highly inconsistent from year to year, and have very high rates of misclassification. One recent major study found that using one year of data—which is what would be used for first-year teachers--there is a 35% chance of identifying an average teacher as highly ineffective. More importantly, using three years of data—the precise time frame used for provisional teachers—there is a 25% chance that an average teacher would be mis-identified as a highly ineffective teacher. Thus, obtaining a standard certificate will literally be a roll of the dice for beginning teachers. Frighteningly, 1 out of every 4 beginning teachers who remain in the classroom for three years will be denied tenure simply because statistical models are not very accurate.

Accuracy of Tests in Assessing Learning

Value-added rests on the assumption that the tests used I value-added are good measures of student learning and provide an accurate barometer of student knowledge and skills. However, there are many reasons why this may not be true such as poorly constructed tests and teaching to the test which causes test score pollution.

If we rate the same teacher with the same students, but with two different tests in the same subject, we get very different results. Recent research by economist Jesse Rothstein using the highly touted Measuring Effective Teaching (MET) study funded by the Gates foundation compared value-added results for the same set of teachers and students using two different tests in the same subject area. He found that more than 40% of teachers who placed in the bottom quarter on one test (the state mandated test) were in the top 50% when using the other test (alternative). That is, teacher ratings based on the state assessment were only slightly better than a coin toss for identifying which teachers did well using the alternative assessment. Similar research in Houston ISD by Sean Corcoran found that a substantial percentage of the teachers

rated effective on the TAKS were considered ineffective on the Stanford tests and vice versa.

Student Peer and Background Effects

The entire point of employing value-added models is to control for student background characteristics and peer effects. However, at this point, we cannot fully tease out these effects. No-matter how hard researchers 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 such as peer effects (race, ethnicity, poverty, and parental level of education of peer group). Compounding this effort is the highly segregated nature of classrooms in schools in Texas. As many researchers have noted, Texas schools are highly segregated and have become more segregated over the last 25 years. This hampers our ability to make valid comparisons across teachers who work in vastly different settings. Statistical models attempt to adjust away these differences, but at present, simply cannot do so in an effective way.

Lack of a Pre- and Post-Test Combination

A vast amount of research has documented that kids learn over the summer and that different kids learn different amounts. Indeed, research consistently finds that higher income students learn more over the summer than lower income kids. Annual testing significantly over-estimates the effectiveness of teachers with high-income students and significantly under-estimates the effectiveness of teachers with low-income students. Consequently, annual testing data are not very useful for identifying student learning gains, thus value-added estimates based on annual testing provide inaccurate measures of teacher effectiveness.

Inclusion of Teachers

Any value-added measure would only include teachers in grades four through eight who teach reading or mathematics. This is approximately 15% of teachers in the state. Value-added methodologies simply would not be appropriate at other grade levels and subject areas because (1) tests are not administered in that grade level or subject area, (2) tests are not administered in consecutive years, or (3) the tests are not vertically scaled.

This, the vast majority of the teachers will NOT be assessed using value-added methodologies. Thus, all the problems associated with other measures of effectiveness become paramount.

Test Characteristics

To use value-added assessments as part of the evaluation of teachers, then the tests must have certain properties, one of which is that the gains at the low- and high-end of the distribution represent the same amount of learning. More specifically, a 10 point gain at the low end of the test score distribution—say 20 to 30--indicates the same amount of learning as a 10 point gain at the top end of the distribution—say 80 to 90. Most tests are not designed this way. The tests that are designed this way typically only work this way at one grade level, but not across grade levels. We might be able to make comparisons between two 7th grade reading teachers, but certainly not between a 7th grade reading and a 7th grade math teacher or even a 6th grade math

teacher and a 7th grade math teacher. Thus, for the system to be accurate, only teachers assigned to teach the same grade level and subject area could be compared. This will cause sample size problems which greatly reduce the accuracy of value-added estimates.

Sample Size

Most teachers in grades four and five teach less than 30 students. Current value-added estimates are very, very poor estimators of teacher effectiveness when the sample size of students is very low. In particular, the value-added estimates are incredibly unstable when the sample size is this small. For example, one study showed that only about 20% of teachers in the top quintile one year were in the top quintile the next according to the results of the value-added estimates. This means one of two things:

1. The estimates of teacher effects are not much more accurate than drawing random numbers; or;
2. Teacher quality changes on a yearly basis and a teacher can be terrible in one year and good the next while another teacher might be good one year and terrible the next.

If number 1 is the correct interpretation, then value-added estimates are far too inaccurate to use when estimating teacher effectiveness,

If number two is the correct interpretation, then the underlying idea behind evaluating teacher effectiveness—namely that teacher effectiveness is stable over time—is incorrect.

Regardless of whether conclusion 1 or 2 is correct, value-added estimates are not very useful

Current belief is that the problem is number 1 and we must try creating more stability in the estimates by increasing the sample size. Four possible ways to do this are:

(a) Increase class size to at least 100;
(b) Collapse groups of students across time;
(c) Collapse groups of students across teachers (school effectiveness measures or subject area group effectiveness measures; and, or
(d) Some new statistical technique.

Option (a) is not really a viable option as neither teachers nor parents would find it acceptable. . Option (c) would not allow for the evaluation of individual teachers since students would be pooled across teachers. Researchers are currently working on option (d), but little progress has been made. Option (b) is the one chosen by most researchers, but requires larges investments to create longitudinal data systems and also becomes increasingly sensitive to missing data.

Missing Data

Missing data can seriously affect the accuracy of value-added models. Missing data include test scores data, connections between teachers and students, pull-out teacher-student connections, student entry and exit dates, and student characteristics. If teachers are to be assessed using

multiple years of student data, then the potential for missing data increases dramatically.

High student mobility increases the likelihood of missing data and poor, urban schools have higher mobility than suburban schools. Thus, missing data tends to be correlated with specific student and school characteristics. Further, for STAAR, it appears that TAKS-alternative and TAKS-modified tests will have different scale scores as they do now under the vertical scale score system. This greatly exacerbates the problems of missing data and increases the correlation with poor, minority, and urban schools. Thus, we cannot just ignore missing data. Yet, most value-added systems do just that.

Model Choice

There has not been enough research on value-added models to determine a "best" model. Interestingly, a recent RAND report found that 50% of the teacher effect can be explained by differences in the models used.

Gaming the System

Teachers will game the system by teaching to the test and even cheating. There is an extremely strong research base in this area and these effects are generally attributed to Campbell's Law, which states:

> ***"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."***

Teaching to the test is considered many to be the entire point of testing and accountability systems. Indeed, what could be wrong with teaching students what will be on the test and then measuring how well students can answer questions about the content? This certainly works within teachers' classrooms when there is a 1-to1 relationship between content and what is on the test. The problem arises when the test cannot cover the entire curriculum which is the case with TAKS and STAAR. When teachers teach to the test rather than the broader curriculum, the test scores no longer accurately reflect students' knowledge and skills with respect to the broader curriculum. Given that the STAAR will focus on even a smaller portion of the curriculum than TAKS, this danger is even more pronounced than with TAKS.

MEASURING TEACHER EFFECTIVENESS THROUGH OBSERVATIONS

There is good evidence that teacher observations can be quite accurate in identifying teacher effectiveness. This is especially true when those observing are well-trained and have adequate time to conduct the observations. Unfortunately, not all of our observers are well-trained and evidence strongly suggests that principals simply do not have the time to provide high-quality instructional leadership already and increasing the number of observations and evaluations will only exacerbate this problem. This is why many districts have teacher facilitators and teacher evaluators employed in the central offices. However, these are the very positions that legislators have called upon districts to eliminate instead of eliminating teachers. In the upcoming school year, districts will likely be forced to reduce central office employees that provide instructional leadership and perform classroom observations as well as a reduce school assistant principals and counselors. All of these reductions will place a greater burden on principals who are clearly already over-burdened as evidenced by the extremely high turnover of principals, especially in low-performing schools and in all high schools.

Thus, this bill would lead to either greater principal turnover, lower quality observations, or an unfunded mandate to employ additional individuals qualified to complete the increased number of teacher evaluations. Funding to employ these additional individuals is most definitely money well-spent. However, given the current budget situation, it is unlikely that observers employed at central offices will remain employed.

CONCLUSION

There is a large and growing body of evidence that teacher quality and effectiveness is the most important school-level factor associated with student achievement. Yet, research has also consistently found that a student's home life is a far more powerful predictor of student achievement than teacher quality or effectiveness. Indeed, even a mediocre teacher can appear highly effective if teaching a classroom full of highly motivated students whose parents provided a loving, nurturing foundation and ensure full student engagement in the learning process at home and at school. Likewise, a truly effective teacher may appear only mediocre when faced with large class sizes filled with students who come to school hungry, unmotivated, and feeling the effects of poor health, abuse/neglect, and lack of overall parental support.

Valued-added methodologies have helped us take a step forward in disentangling all these various factors in our attempt to identify effective and ineffective teachers. Yet, we have sill not created value-added models that are accurate and stable enough to be used in making high-stakes decisions.