

TEI Letter to the Dallas Board of Trustees from  
Michael A. Dryden, LLC.  
April 2014

Evidence based on Dallas ISD students suggests that the Teacher Excellence Initiative (TEI) will not produce valid and stable results to tier teachers into a bona fide merit wage system
---

**Background:**

Dallas ISD is about to embark on a bold experiment to shift from a seniority wage system where teacher compensation is based on years of experience and educational attainment, to a merit wage system where compensation will be based on performance. This differs from recent “incentive” models where the district maintained the seniority wage system but paid extra compensation to teachers whose students had outstanding student performance. If approved, the TEI will be one of the few merit wage systems for public school teachers in the world.

The TEI is designed to have three performance criteria, student performance, classroom performance, and student perceptions of the classroom learning environment. Many of the mechanisms for measuring these criteria have not been made public or at least I do not know about them. I am writing this letter to inform the BoT that past student history and recent new analyses strongly suggest that further research needs to be performed to validate the TEI before the BoT gives approval. Almost every component of the TEI has a history of lack of validity and reliability with Dallas students for establishing performance tiers. I will take each of the three components, one at a time, to inform the BoT of recent research based on Dallas ISD students questioning the TEI, followed by an analysis of the logic of triangulation of these three components and consequences of approving a possibly invalid system.

**Student performance:**

I am not clear about actual student performance calculations. I keep hearing that teachers will have a choice between absolute performance like percent passing, or two value-added oriented growth models, either CEI scores or a version of the median student growth percentile (SGP). It would be absurd to let the teacher select the first method because percent passing is not solely a function of teaching. By having this option, teachers of students from more affluent families could be falsely rated as effective replacing truly effective teachers where fixed numbers will be allocated for each payment tier. The only real options are the student growth measures.

Let us start with the familiar CEI scores. CEIs were never designed to tier teachers into 7 to 9 bands and they are unstable in the middle of their distribution. CEI scores were designed to be within campus indicators of possible teacher performance issues. At the district level, with the exception of the upper and lower 20th percentiles, the CEI scores over time almost mimic a random number generator. Given the complexity of the CEI calculations, I wrote a paper on the visual understanding of the CEI scores that is enclosed. In that paper I hope the BoT sees that

the power of the CEI scores is at the campus level because the errors associated with CEI scores tend to be similar within a campus but not across campuses. For example, a powerful predictor of achievement, the emphasis of education in the home, is never measured and for peer groups within a campus that unmeasured trait will be similar. At another campus, it may be very different than the first campus. If the district decided to tier teachers within a school based on CEI scores, instead of across all schools, then that would be equivalent to a failed procedure in industry called stack ranking. In other words, each campus would have a fixed number of teachers within each tier.

At least the CEIs attempt to eliminate biases due to factors outside the teacher's control such as poverty, overcrowding, level of English language acquisition, and multiple measures of prior achievement. The Median Student Percentile Growth Model (SPG), which I will explain shortly, does not adjust for factors outside the classroom. To illustrate how CEI scores cannot be used to tier teachers the figures below, which are based on actual CEI performance data, show that the middle 60% of the district-level CEI distribution over three years mimics a random number generator. That is hardly conducive to ranking teachers for wages based on performance. In the diagrams below, each dot represents the math CEI percentile rank of each, the middle axes represent the 50<sup>th</sup> percentile, and the ends represent 1st and 99th percentile rank. If the CEI scores were stable for placing teachers in payment tiers we would expect more teachers to align along the diagonal. I will send a data file for those BoT members who like to analyze with Excel. It has a nifty 3-d rotation for BoT members to see the actual stability of the CEI scores.



Figure 1. CEI Percentile rank positions (1-99) of teachers over time

The CEIs are limited in their use for TEI. They do not include non-core and certain core subject teachers and more importantly, they take time to calculate. It would be almost impossible to

complete the calculations before the new year starts in mid-August for teachers. The alternative looks like it will be a modification of the Colorado Growth Model or Median Student Growth Percentile, SGP. It is a very popular measure used in many states and some districts but it fails to adjust for fairness variables and ignores errors in measurement. Thus, SGP relies on extremely high quality assessments such as professionally developed state tests. The district developed ACP tests do not have the level of precision of well designed, piloted, and field tested state assessments. There is a high probability that SGP measures are more random over time than even the CEI scores because fairness variables have not been partitioned out of the system and measurement errors will be much higher than CEI scores which are based on professionally developed assessments.

As a simple explanation of SGP consider a student, Johnny, who scored in the 45th percentile on a pretest in Chemistry. Now let's define Johnny's peers as all those who scored between the 40th and 50th percentile. On the posttest in Chemistry AMONG HIS PEERS Johnny ranked at the 70th percentile. Maria scored at the 92nd percentile on the pretest and her peer group is the 90-99 percentile group. On the posttest in Chemistry, she scored at the 57th percentile AMONG HER PEERS even though she scored higher on the posttest. Now Ms. Smith, who has 120 chemistry students with these SGP scores, finds the median SGP of all her students and that is her effectiveness measure. Actually, Dallas ISD may be taking the mean and not the median.

SGP sounds simple except for a few major details. Nobody has tested the accuracy of the ACPs for placing students accurately into peer groups. The well-established core subject ACPs have a history of problematic items. The newly designed ACP measures in other subjects are assumed to be of highly reliable but this has not been verified and history is not in the favor of the ACPs. While I know there are personnel in DISD capable of using industry standard Rasch modeling in developing assessments for growth, this has never been done.

Last year when I was helping the chemistry and Algebra I teachers at Conrad High School I did a very deep analysis of various test scores, including ACP tests, to develop a narrative profile of the students where the teacher selected the student and up popped a description of the student in terms of issues with achievement performance. To do this narrative I had to analyze the ACP tests. I signed an agreement not to discuss specific items of ACP tests but I can tell you they were not designed to measure growth. Worse, they had multiple conceptual issues in math and science that bordered on embarrassment, a wrong answer key, and some dual answer items. The plot below maps the ACP chemistry responses with a technique new to education called multiple correspondence analysis. Each point is a unique item response like 1a, 3b. All the red points (distractor responses) should be on the left and all the green points (correct answer) on the right. To be appropriate as a growth measure the plots should be flatter along the horizontal axis, it is called unidimensionality.

The BoT should realize all the recent ACP testing is to develop both a pretest and a posttest in almost all subjects to run a SGP model of growth for the TEI. Being very concerned about the ability of the ACP tests to tier teachers, I had a freedom of information request made to quality control the ACP data and test the TEI. Unlike their release of the CEI data, the district refused to release the ACP data and that request is now with the Attorney General's office for a ruling.

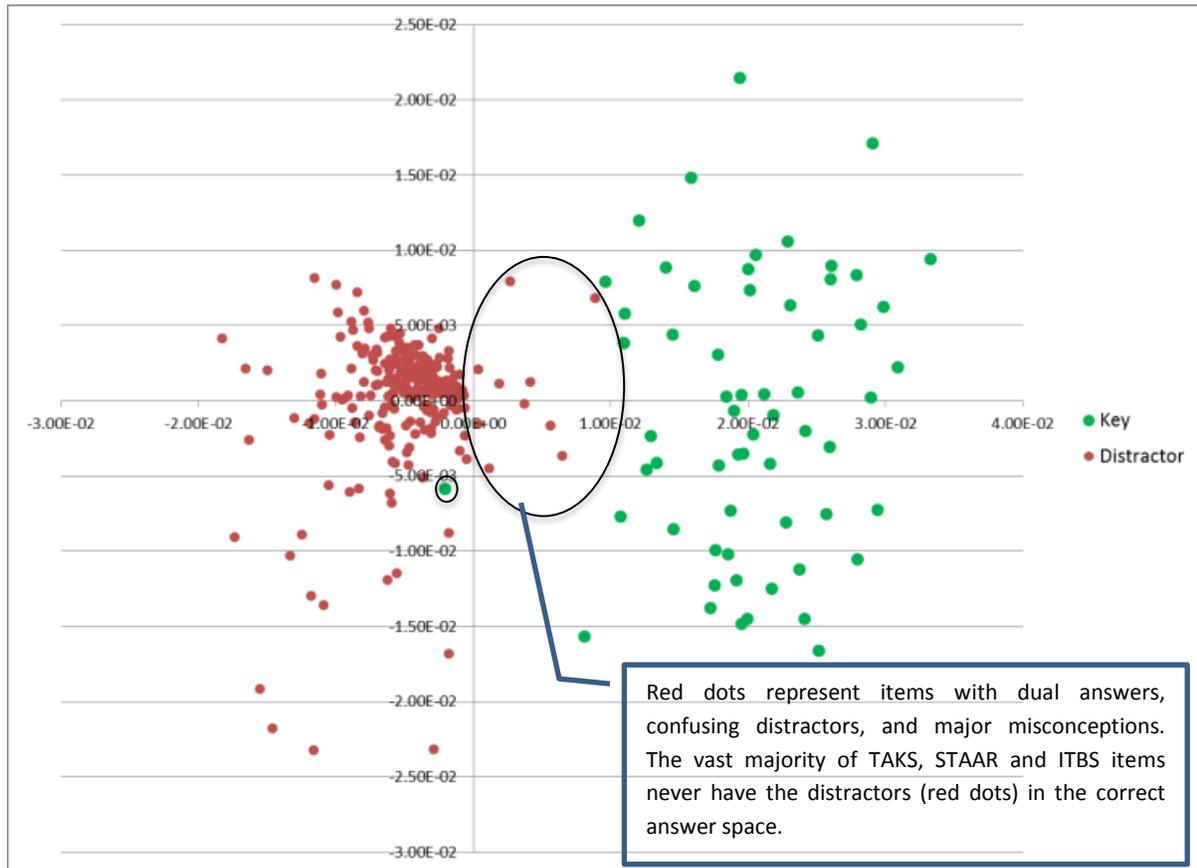


Figure 2. MCA of PreAP Chemistry ACP item responses in first two dimensions. (Fall 2012)

Another SGP issue is the assumption that the pretest is an accurate metric for a starting point and that everyone has the same growth expectation. In essence, the CEIs adjust for this by statistically having difference buckets. With CEIs a limited English female student on free lunch and a pretest at the 40th percentile has a different growth expectation than a white male student not on free lunch and a pretest at the 40th percentile. It looks like the Dallas SGP model will put them in the same peer group with the same expectations of achievement growth. It is difficult to see how this is a fair system if the peer group is defined solely based on one pretest score. If there are many peer groups based on the many fairness variable then the SGP buckets would have too small a number of students and other techniques like Bayesian statistics, which is embedded in the HLM model of the CEIs, would need to be used. The Median Student Growth Percentile model is not so simple anymore.

Had I been given access to the ACP longitudinal data I would have tested the stability of the SGP over time. I am confident it would be less accurate than the CEI technique and not applicable for placing teachers on tiers for a merit wage system.

**Student Perception of Classroom Learning Environment Surveys (Tripod):**

The second part of TEI is the student survey on the classroom learning environment. I am thrilled that the district might be using Prof. Ron Ferguson’s Tripod survey to understand the learning environment of the students. Almost twenty years ago, I got together with one of the top learning environment researchers in the world, Prof. Barry Fraser of Australia to administer one of the largest system-wide surveys focused on the classroom learning environment, not just school climate. One of the instruments we used was the My Class Inventory, MCI. Our research mirrored the Tripod survey results. There was considerable overlap between the student’s perceptions of their learning environment and achievement growth. In our example, the learning index was the CEI score at the student level. However, in Dallas the climate-achievement growth correlation seems to be defined by the upper end of the ability spectrum and the upper end of the learning environment spectrum. In other words, highly positive learning environments are related to highly effective teachers (or perhaps the reverse) but little else can be said among average or low performing students and the average or low perceptions of the classroom learning environment. The figure below illustrates this concept. Realize this was the mid-1990s when the district administered the ITBS test it could better discriminate the high achieving students. It is unknown whether the STAAR test will have the same results.

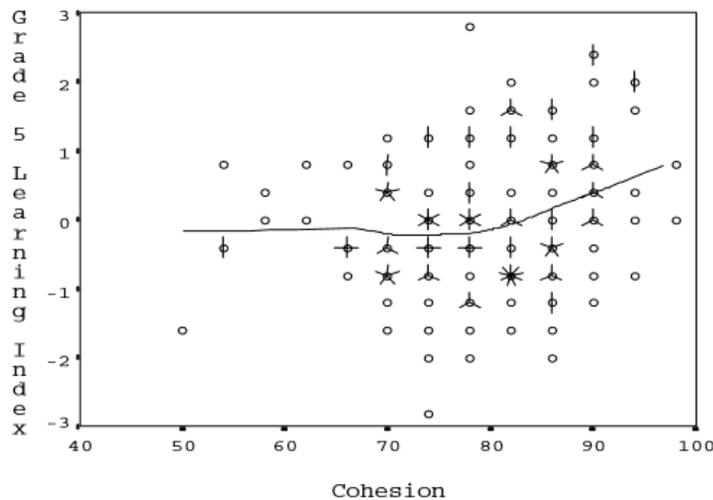


Figure 5. Relationship between cohesion and learning efficiency, grade 5.

Trend: Low cohesion has little relationship to learning but high cohesion is directly related to learning.

Like the CEI scores, the student surveys may be applicable for identifying exemplary teachers. The student survey results may not be applicable to place teachers in lower tiers. Thus without further research we should not conclude the student surveys are sensitive enough to tier teachers across the entire teacher effectiveness spectrum, especially for a merit wage system. Prof. Ron Ferguson has dedicated his career to improving the performance of minority and traditionally underserved students. The Tripods surveys were designed to help identify classroom learning environments for improvement. I cannot find evidence they were designed to tier teachers for the purpose of a wage structure.

On a recent DISD survey on college readiness, 2012, I used a new technique called thermometer coding to analyze the data and it clearly indicated that a very large portion of the students failed to read the items correctly. The Tripod pilot survey needs to be analyzed using thermometer coding to see if lower reading ability students in Dallas actually understand the questions or take them seriously. If 40% of the students do not read the items carefully or take them seriously, it would be impossible to place teachers in a merit pay system without severely misrepresenting many teacher rankings. I suspect the reason I got flat lines twenty years ago on the My Class Inventory was that the lower reading ability students either had comprehension issues or failed to take the survey seriously.

There is a more serious issue related to using young children to rate teachers for the purpose of teacher pay. Like a bad divorce proceeding, many young children simply cannot separate their world from others and may think they are the source of the problem. It may not be wise for the district place young minds in a situation where they believe they had a part in the possible demotion or termination of teachers.

### **Classroom Performance:**

The third part of TEI is classroom performance or spot observations. Teachers are saying the spot observations are rigged with early observations being falsely low and final observations being tied to student achievement scores, perhaps SGP scores. Without examining the data, I have no way to know if this is true. If this is true then obviously spot observation growth will be highly correlated to achievement growth and the district might use this manipulation as a false sign of validity for the TEI methodology. This suspected manipulation is actually testable if there is enough variance in the rubric scores. If raters give similar final observation scores for different reasons on the same teachers then internal consistency has to be challenged. There are sophisticated statistical techniques to see if raters are consistent not only in their overall assessment of a teacher but the underlying traits the teacher exhibits. This alleged rigging would collapse in the second year anyway. Achievement scores from year to year are not perfectly stable and there would be natural variation. Rigged the spot observations would then be correlated to noise, a highly unnatural outcome. Nobody has nor will perform validity tests on these observations. In addition, there is no evidence that the district trained raters to

calibrate observation ratings to some acceptable standard. The BOT should demand that all spot observations be finished and reported to the teachers before test results come back, whether they approve the TEI, or not.

**Accumulation of evidence.**

Finally, the district seems to be touting some strange triangulation logic that if three imprecise methods show a trend then that trend must be highly accurate. That logic does not work if the methods lack accuracy. If we want to find out how fast someone is running a 100 yard dash it would be difficult to draw an accurate conclusion if our timing instruments were an hour glass, a sun dial and a pendulum.

All of the above concerns are testable and the Board of Trustees should use an independent group to test the validity of the TEI before they vote on it. The consequences are just too severe. The TEI is an attempt to be a bona fide merit wage system, not an incentive bonus on top of a seniority wage system. Given past analyses, there is a high probability that the TEI will define a bogus merit wage system making the district highly susceptible to protected class litigation. There is a high probability that TEI tier system will falsely place many teachers in the wrong tier. Poor teachers falsely placed in high tiers will not leave DISD but good teachers falsely placed in low paying tiers will quickly leave the district. The evidence strongly points to the real possibility that the implementation of the TEI will accelerate the current teacher exodus. Passing the Teacher Excellence Initiative could be catastrophic for the students.

Commentary:

What I presented above is not new. Top educational researchers have been questioning the validity of value-added measures for years, including SGP. The MET Project is increasingly under scrutiny by leading educational researchers. Using student responses to inform teachers on how to improve the classroom-learning environment could be very beneficial but taking that same information to place teachers into wage-oriented tiers is uncharted territory. I strongly recommend that the model TEI model proposed by the district be vetted with real Dallas student data that becomes available this summer and that the scheduled vote on it next month is delayed until December 2014.

Michael A Dryden, LLC