The Potential Impact of the Classroom Effectiveness Indices on the Teacher Excellence Initiative: A Visual Examination

By Michael A Dryden, LLC

March 2014

Dallas ISD is about to embark on a bold plan to change the teacher wage structure from a seniority wage system where teachers are paid based on years of experience and educational attainment to a merit-wage system where teachers are paid based on performance. One potential measure of performance being contemplated is the Classroom Effectiveness Indices or CEI. In the past, the interpretation of the CEI scores by the media, board members and even some district administrators has shown gross misunderstanding of the real power of the CEI scores. Few people realize that the underlying data creating the CEI scores are a valuable asset in terms of the district's new push towards personalized learning. At the campus level CEI scores are an invaluable tool for principals and others as long as the limitations are acknowledged and these scores are used with other pieces of information.

However, CEI scores lack sensitivity to tier teachers into a pay for performance merit wage system. At best, they could be used to discriminate consistently very high performing teachers for an exemplary rating. Even then the exemplary rating would not be applicable to all teachers. In an attempt to explain CEI scores to the general public, the rest of this paper will be devoted to a visual non-mathematical explanation of CEI scores. While the presentation that follows is obviously an oversimplification of the actual process, the underlying philosophy it presents will hopefully better inform the public of the power and limitations of the CEI scores, especially if used in the Teacher Excellence Initiative. The reader can find a more thorough technical discussion at https://mydata.dallasisd.org/docs/CEI/SEI_CEI_Research.pdf .

CEI scores are based on a specific post-test assessment. For example, the post-test may be 5th grade STAAR mathematics as predicted by various 4th grade assessments the prior year. The first step of CEI calculations is to determine which part of the pre-tests and post-test are due to school and classroom factors and not outside factors such as race/ethnicity, gender, limited English proficiency status, free/reduced lunch status and any interactions of these variables that help to explain the post-test. See figure 1. The fairness adjusted pretest is further adjusted to take into consideration school level factors such as school mobility, over-crowdedness, census block family education, census block family income, census block poverty index, percent free/reduced lunch, percent minority, percent African American, percent Hispanic and percent limited English proficient.

Next, the part of the post-test that can be explained by the pretest after controlling for student level and school level fairness variables is found. That part which cannot be explained is a residual or remainder and that residual is the basis for the CEI calculations. Literally, that which cannot be explained by known data is the basis for CEI calculations. The residual is believed to embed an estimate of the student's learning as a result of the instruction in the school.

The residual is actually an instructional effect plus other things such as important variables not measured, and measurement errors. Some people call this a value added measurement (VAM).

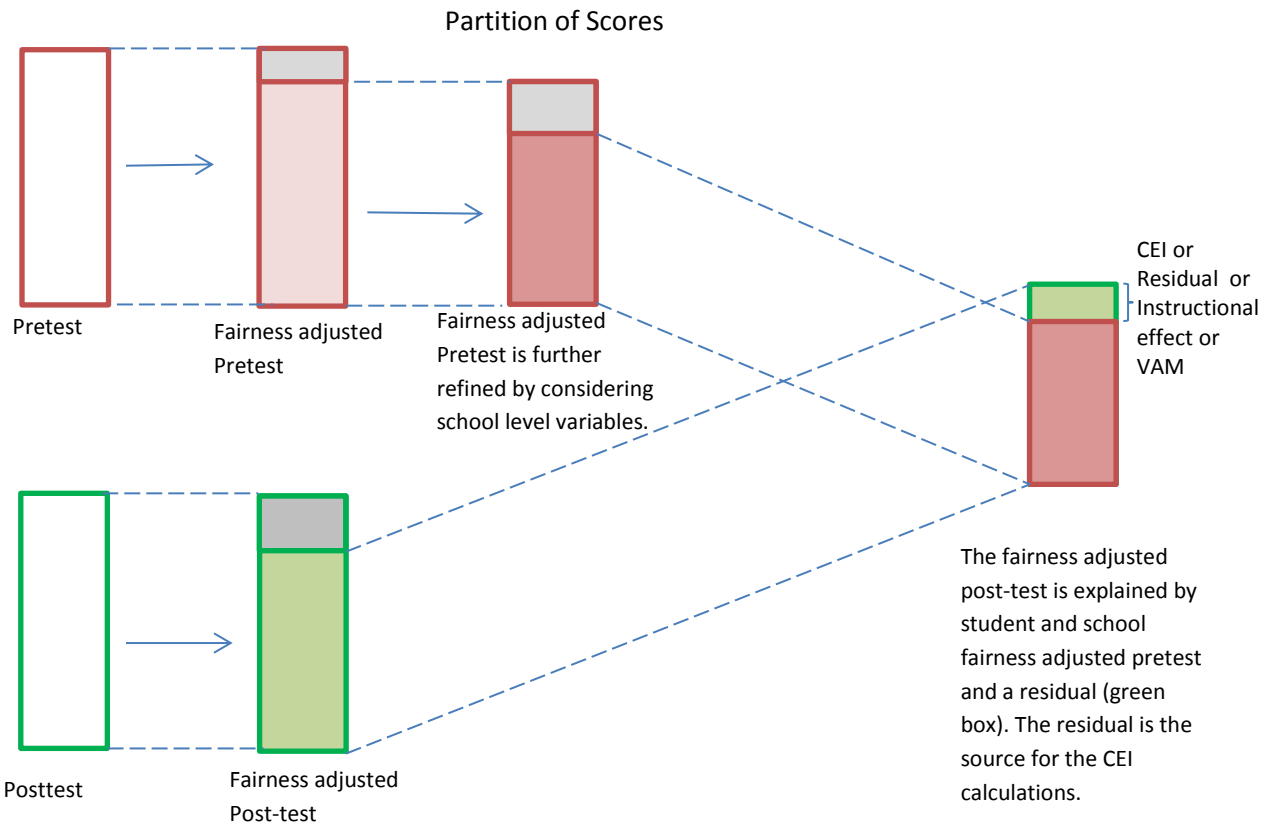

Figure 1. Diagram for development of CEI scores

The green residual in figure 1 is often called a value-added instructional effect on the student and is recalibrated relative to the district by setting the district average instructional effect at 50 with a standard deviation of 10. This means a student with a residual of 55 is half a standard deviation above the district average of similar students. The E&A department then provides a classroom chart to help the teacher. It may look something like figure 2.

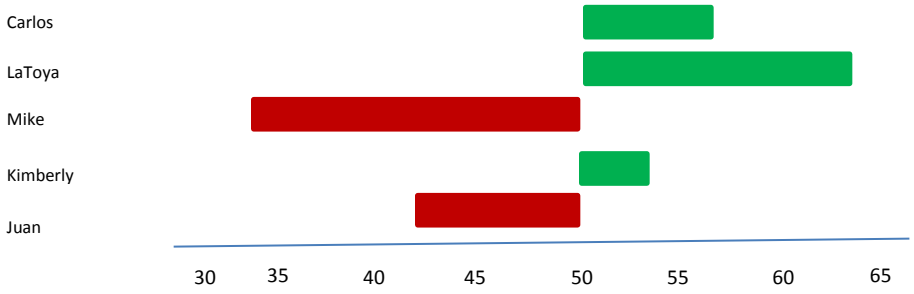

Figure 2. Sample classroom student learning or residual results.

If the district average is taken as the expectation then figure 2 can be interpreted to mean Carlos and LaToya exceeded expectations and value based on instruction relative to the district was added, Kimberly met expectations and Mike and Juan did not meet expectations. The teacher and perhaps the instructional coach, principal or assistant principal might get together to determine why Mike and Juan were not learning as much as their potential suggests.

But the observed residual/instructional effect is really a true instructional effect score plus other things. These other things might be caused by assessment unreliability issues, important variables not measured or even imprecise measures. For example, a STAAR test item may have a correct and a partially correct response. The emphasis of education in the home is very important but not measured and free/reduced lunch or census block income information is not as accurate a measure of poverty compared to actual family income.

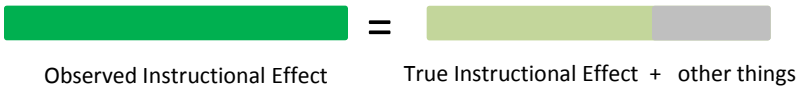

Figure 4 Observed instructional effects as combination of true instructional effects and other things

Within a campus these "other things" such as not measuring emphasis of education in the home, peer influences, or effect of language comprehension on test results, are often similar and the rank order of the residual is similar to the rank order of the true instructional effect.

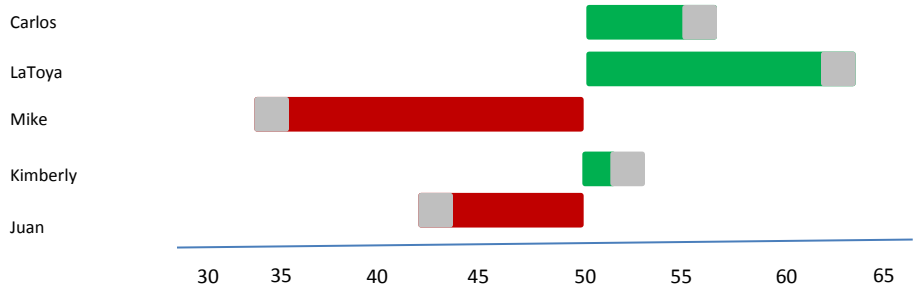

Figure 5 Student summary illustrating error effect as constant.

These residuals or instructional effects scores are actually ranks and not scale variables. As ranks, taking the difference between scores, and thus CEI scores, has little meaning.

Obviously, large residual or instructional effect score differences are suggestive of improvement or loss of effectiveness but attempts to quantify a difference or interpret small differences should not be made. These residual scores are just not that sensitive to be calculated to the nearest tenth of a decimal place or sorted by the nearest 1000th of a decimal place.

Unfortunately a principal cannot examine 200-2000 students each with four to six student bar graphs based on various post-tests. To help principals quickly identify potentially ineffective teachers the E&A

department calculates an average of the instructional effects (residuals) by teacher. A reliability adjusted version of this average is the CEI. Note that it is not exactly a teacher effect since many teachers could contribute to the math learning of a child. For example, a bilingual teacher could raise the language comprehension level of a child to better solve math word problems. People tend to think of CEI scores as a measure of teacher effect but these scores are really an attempt tp measure the school based instructional impact. It is assumed the teacher of the posttest subject was the main contributor to the instructional impact on a child.

The actual power of the CEI score is at the residual or instructional effect level. It would be a powerful piece of information in terms of personalized learning to know if a child is reaching the same potential as other similar students and the district should recognize this information in any application for grants based on personalized learning. Sometimes there are extraordinary factors beyond what is in common with the other students, like a divorce in the student's family, and the principal and teacher will understand that anomaly and make extra efforts the following year to "recover" the learning of that particular child. Other times a principal may notice that the residual scores of a teacher are very low for too many students suggesting further investigation into a potentially ineffective teacher. Under no circumstances should a CEI score, in isolation, be made to judge teachers. In fact, it is against Dallas ISD policy. As the reader can see the student residual as a student learning index can be a great analytical tool for personalized learning and the CEI of the teacher can be a great tool for principals to spot problematic teachers. Given the known insensitivity of the CEI data, E&A categorizes teachers into only five tiers and that may even be a stretch.

Problems occur when Board members, politicians and even naïve central administrators attempt to treat CEI scores as pure values and with an unjustified high degree of accuracy. In the late 2000's the district deemed that teachers with a CEI score of less than 50.0 at low performing campuses must be replaced by a teacher from a non-low performing campus who had a CEI greater than 50.0 and that replacement teacher would earn a substantial bonus if they transferred. That was an absurd practice.

If the district decides to use CEI scores across the district as part of a procedure to tier teachers into a merit pay system, that same misunderstanding could happen with the Teacher Excellence Initiative. Five teachers may have the same TRUE classroom effectiveness score but the OBSERVED classroom effectiveness scores could be vastly different if the unknowns are not the same. The emphasis of education in the homes of TAG students is probably not the same as at Roosevelt High. The peer pressures at T. J. Jefferson may not be the same as at Molina even though the student populations are similar. The language acquisition status of recent refugee immigrant students at McShan Elementary may not be the same as recent immigrants at J. F. Kennedy Elementary. Since only the observed classroom effectiveness scores are known, five teachers with the same TRUE CEI scores may have wildly different OBSERVED CEI scores.

Nobody really knows how much of the observed instructional effect is actually true instructional effect (signal) versus error (noise). If the errors are very small or if they are the same across the district then the observed CEI score could be used to estimate teacher effectiveness across the district.

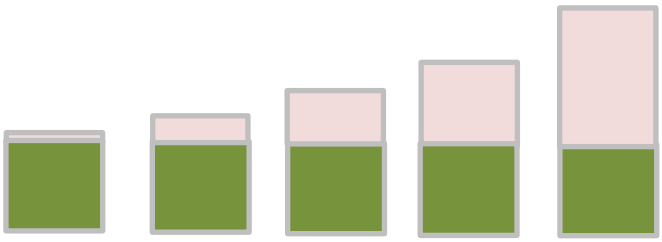

Figure 6. Observed Teacher Effect as Same True Teacher Effect and Various errors

Fortunately, the assumption that either the errors within the observed CEI scores are very small or the same for everyone across the district versus errors that will vary greatly can be tested. The entire premise of a merit system based on student performance is that teacher effectiveness is a stable construct over time. Teachers can improve or decline but wild variation in teacher effects over years would undermine the notion that teachers could be tiered one year and remain locked in that tier with only exceptional performance to slowly jump tiers.

Since the district must assume that a true CEI score is relatively stable, then a teacher will have similar, but not exactly the same, observed CEI percentile rank scores across three years. If we take a cohort of teachers who have CEI percentile rank scores for three years in a row and assume that the errors are small or the errors are constant then the plot of their CEI percentile rank scores should cluster around the diagonal of the cube and might look something like the first plot in figure 8.

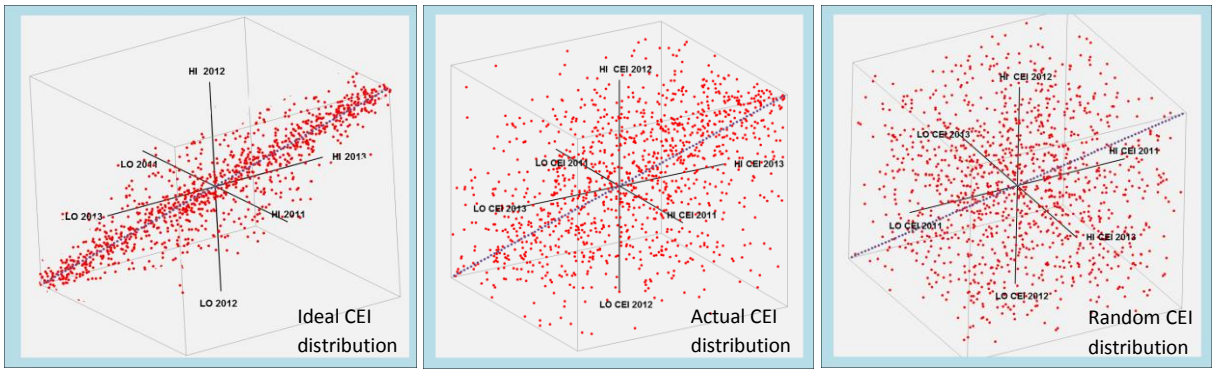

Figure 8. Possible distributions of 3 year CEI percentile rank scores if errors are small or nearly constant.

If the errors are not the same then the observed CEI percentile rank scores could be all over the place. Of the 2,681 teachers of mathematics in 2010-11 with CEI scores only 1,224 continued to have a math CEI in 2011-12 and 2012-13. In the scatterplots above and below, the ranks are the original ranks among all teachers for that year based on the 1,224 math teachers with the three years of CEI scores. The middle chart in figure 8 represents actual math CEI percentile rank scores; the third diagram represents total randomness.

The second scatterplot of figure 8 shows that the actual three year CEI percentile rank distribution almost fills up the cube and the teachers are not clustering along the diagonal based on this picture.

The first chart of figure 9 plots the CEI percentile rank scores of teachers in 2013 versus 2012 for those teachers who had a CEI scores all three years. The distribution looks almost random but upon closer inspection the extreme upper left and lower right quadrants are less dense and the upper right quadrant indicates very high CEI teachers tend to be retained. In the middle circle the distribution seems to mimic randomness.

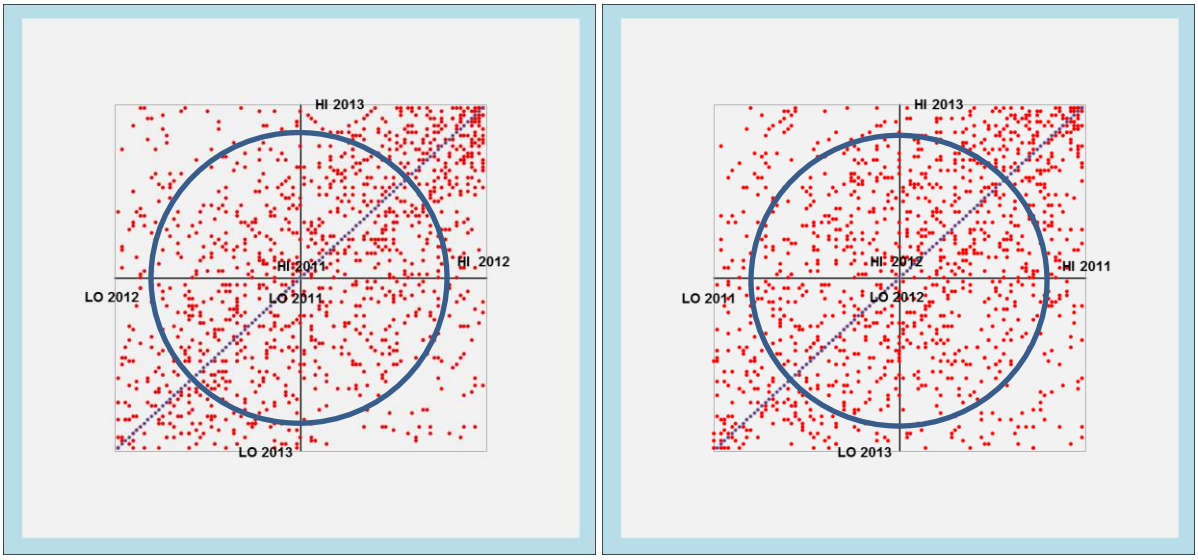

Figure 9. Actual Math CEI values for 3 year cohort of teachers.

The second plot of figure 9 plots the CEI scores of 2010-11 versus 2012-13 and exhibits the same patterns as the first plot. The best visual exploration of signal versus noise would be to look directly down the diagonal of the cube. This would illustrate deviations from the diagonal.

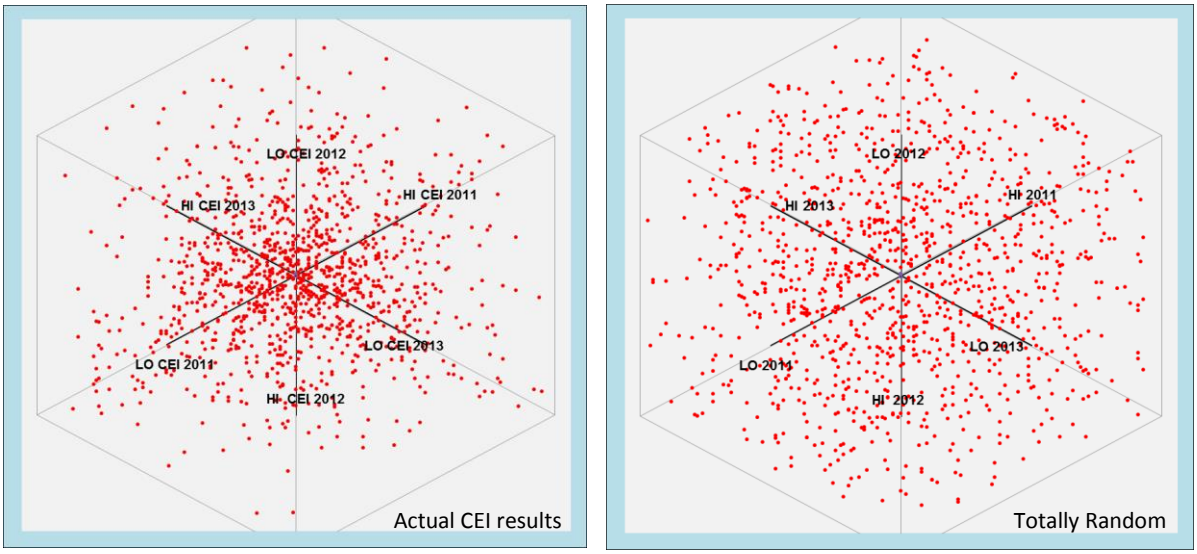

Figure 10. Comparison of Signal Information of 3 year CEI Cohort Values Versus Pure Noise.

The first scatterplot in figure 10 illustrates there is some signal based on examining three years of CEI percentile rank scores and it is not completely random, which is shown in the second plot of figure 10. The precision of CEI scores seem to increase when looking at CEI scores over time but most of the signal is at the extremes. There is a higher probability that a teacher with a very high CEI score three years or more is a very good teacher. A surviving teacher with very low CEI scores three years in a row needs to be highlighted and examined on a case by case basis. There have been situations where teachers consistently follow extremely high CEI teachers and look ineffective.

The point is the CEI scores become more stable over time but the information is at the extremes. The CEI scores do NOT support placing teachers on a tier system unless that tier system had three tiers and based on longitudinal data. Those three tiers might be exemplary high CEI teachers, surviving low CEI teachers and the 70% of teachers in between of which no district-wide judgments can be made based solely on CEI scores.

To put this in perspective, of the 1224 teachers in the cohort, randomly we would estimate 1224*.15*.15*.15 = 4 teachers to have a CEI percentile score of 85th percentile or higher three years in a row. There are 50 actual teachers with CEI percentile ranks greater than or equal to the 85th percentile three years in a row. CEI scores seem excellent for identifying exemplary teachers.

A 46% math teacher retention rate over two years is very unhealthy. Backing up one year and examining the three year CEI retention rate Table 1 shows only 36-37% teachers of reading, math and science had CEI scores in the same subject three years in a row since 2009-10. This is not the same as an actual teacher retention rate since some teachers of small enrollment classes have no CEI and teachers are assigned classes other than mathematics, but these rates will be close to actual retention rates.

Table 1
Three Year CEI Retention Rate of Teachers Based on CEI Scores

	Number of CEI scored teachers				
CEI subject	Base in 2009-10	Retained 2010-11	Retained 2011-12	Retained 2012-13	3 year subject retention rate
Mathematics	2691	2007	1506	997	37%
Science	730	549	406	272	37%
Reading/Language Arts	3048	2278	1623	1100	36%
Computer Science	71	56	45	31	44%
Social Studies	480	404	318	258	54%
Foreign Languages	140	112	79	50	36%

Review of CEI scores:

Here is what the public should realize about CEI scores.

1. CEI scores are NOT perfect measures of teacher effects and have other components associated with them. However, useful and important information can be gathered at the campus level.
2. The CEI residuals at the student level may be a good surrogate for identifying students who are not reaching their learning potential in the classroom. Whether failure to reach potential is due to the teacher or other factors is best dealt with on an individual student basis.
3. CEI scores are aggregate student residuals and act as a quick way for principals to identify potential teacher effectiveness issues. They work because the unmeasured factors within a campus tend to be very similar. However, this assumption that unmeasured factors within a campus tend to be similar needs to be verified.
4. When examining CEI scores across the district and not within a campus the signal to noise ratio is small but does not seem totally random. Stated another way, the CEI scores mimic a random number generator for the middle 70% of teachers if the CEI scores are used across the district.
5. CEI scores across the district based on a three year survivor analysis show very high CEI teachers seems more stable than the teachers in the middle range. Visually and empirically it looks like there is a high probability that a teacher with a CEI of 85 percentile of higher three years in a row is an exemplary teacher. Other than that, no conclusions about teacher effectiveness can be justified.
6. A two year comparison of CEI scores indicates tier placement of teachers based on a one year CEI value would be almost totally arbitrary and not indicative of student performance gains. This would make the TEI tier structure invalid and undermine the validity of the TEI as a bona fide merit pay system. Without a bona fide merit wage system, or a bona fide senior wage system, the district could open itself up to protected class litigation.
7. The low retention rate of math teachers shown in this study places any merit pay system in jeopardy if a core set of higher performing teachers are not being retained.
8. The CEI scores are based on TAKS, STAAR and ITBS scores and not ACP scores. Therefore many teachers do not have CEI scores. For a merit system to work all teachers should be included. The evidence of this report suggests that calculating CEI scores based on ACP test results would be counterproductive for placing teachers in 7 to 9 tiers for salary level discrimination. In fact, given the known lower reliabilities of ACP tests versus TAKS, STAAR and ITBS, the scatterplots should become even more random and the premise that consistently high CEI scores identify exemplary teachers may not hold for ACP based CEI values.