Science & Math CESE Education

The BEACON

News from The Coalition for Excellence in Science and Math Education

Volume XVIII, No. 2 Queries? email M. Kim Johnson (next page)

In this issue: Editor's Message – Kim Johnson. – Part II of the Beacon version of the briefing on NM educational performance and new teacher evaluation protocol (growth based portion) given to the Legislative Education Study Committee and the Legislative Finance Committee Joint Meeting in August 2014

EDITOR's MESSAGE

A LOOK AT CESE'S METHOD FOR SCHOOL IMPROVEMENT AND A DESCRIPTION OF WHAT IS WRONG WITH THE CURRENT NM TEACHER EVALUATION METHODOLOGY

This is the second and last part of a two part Beacon that discusses and analyses key issues in the education of New Mexico's students. The first part covered a summary of the performance of schools in New Mexico over the last seven years and the analysis of the ABCDF Act (school grading) that was signed into law in 2011. Though the Act, or something similar, is necessary to comply with federal requirements, our analysis shows that it is not really reflecting how many schools perform, and it almost certainly misleads many schools in terms of how well they think they are performing compared to how well they really do perform with respect to student achievement as measured by standardized testing.

The federal requirements mentioned, above, required that if New Mexico were not going to suffer penalties, either or both monetary and control of federal Title I (associated with poverty) money as called for in the No Child Left Behind Act (NCLB) of 2003, something had to be done regarding the way NM calculates its schools' achievement. The NCLB Act required that all student score at a "proficient" level in math and reading by 2014, which is clearly an impossible goal, considering that "proficient" can roughly be equated to "average." So, even though the school scoring changed based on a provision in the NCLB act allowing the Federal Department of Education to waive certain requirements of the act, the change in NM did nothing to actually show schools a way to improve. In fact, if anything, it appears that the ABCDF Act actually complicated school evaluation to the point of making it essentially incomprehensible to almost all parties concerned – especially regarding what the results actually mean.

CESE, however, did develop a method that we believe would actually lead to not just some improvement in NM school performance, but potentially to very significant improvement. This method will be covered in some detail in the following pages.

Additionally, one of the requirements by the federal Department of Education is the evaluation of teachers using the growth of their students as measured by standardized tests (with some special exceptions that we do not cover in detail). This has turned into a significant problem, because it can be shown to be ineffective and literally cause inappropriate evaluations for many teachers. Most teachers are well aware of the fact that there are problems, and we will attempt to show specifically what foundation logical problems really are.

To remind readers of some of the details behind our analysis efforts, we repeat, below, a short summary of the previous Beacon's analyses.

New Mexico has, like the rest of the country, been subjected to a significant set of education "reforms". These reforms have been initiated and carried out by our own state's Public Education Department (PED) at the behest of the federal government and from within the PED.

The Beacon is published by the Coalition for Excellence in Science and Math Education (CESE). A 501(c)3 nonprofit corporation, we are incorporated in the State of New Mexico. Visit our web site at http://www.cese.org.

WEBMASTER: Dave Thomas

CESE annual dues are $25 for individual, $35 for family, and $10 for students. Please see last page for membership form. Email Beacon submissions to Editor, M. Kim Johnson, kimber@comcast.net.

Current reform efforts began for New Mexico in 2011. The major parts include a way to evaluate schools, changing the way teachers and principals are evaluated, and making 3rd graders who do not measure as proficient in reading to be held back to repeat the 3rd grade the following year. So far, the last item has not been implemented, but is being debated at the state legislature as this is being written. However, the first two have been implemented, and there are consequences that can already be seen from these two items. In fact, the results require the creation of a number of significant changes to educational procedures in New Mexico.

The questions we ask are: have or are these reforms likely to cause any actual positive improvement in student performance? If the answer to that question is a yes or no, just what impact will or have these changes made and what is a reasonably projected outcome if they continue? **It is important to note that we have asked these questions without preconceived notions as to what the answers may be.**

There are other questions that could be asked and elements of these reforms that are not addressed because of the effort required, lack of good data, and priorities set when we began this analysis.

We presented a briefing covering these issues to a joint session of the Legislative Education Study Committee and the Legislative Finance Committee on August 27, 2014 and in several other venues since then. We believe the data presented in this briefing should serve as a reason to rethink the direction that the school reforms are going, at least for New Mexico, which does not always respond the way other states do to the same situations. That briefing can be found at http://www.cese.org/wp-content/uploads/2014/10/CESE_LESC_MKJ-FinalCommented1.pdf. It has been annotated with notes for better understanding, but the essence of what was presented will follow, below, with some new material added.

This briefing has been well received judging from audience questions. Undoubtedly, some of the viewers were probably not impressed, but nevertheless, the data and its analysis do tell a story if presented in an understandable manner. We hope that we can present these findings, herein, in a manner that does not require a PhD in either education or mathematics to easily interpret. We do realize that most people are not educators and mathematicians yet are constantly bombarded by numbers and graphs, so we took quite a bit of time to try and direct this analysis toward the normal person (a little tongue in cheek, there). We do hope this has been successful, but do welcome questions. My (Kim Johnson's) e-mail address is in the column to the immediate left should there be any questions.

Finally, one must be honest: education improvement is extremely difficult. One cannot expect "Silver Bullets" to work, nor is it easy to gather sufficient data to perform a good analysis. However, what we have appears to be the only data available to the public, and we believe that if there are any errors, they are minor. And please understand that there is nothing personal here and we are not throwing rocks at anyone. We are simply analyzing the data. Kim Johnson

HOW DO WE IMPROVE EDUCATION IN NEW MEXICO?

In the previous Beacon (January, 2015), we presented a pie chart (duplicated later) showing that demographics explains from 60 to 80% of a schools performance on the standardized tests New Mexico has been using (the NMSBA - New Mexico Standards Based Assessment). Also, the demographic factors that contribute the most are minority status and poverty status, **but** it is the combination of the two that overwhelmingly correlate to the average test scores for the state's schools. Other demographic elements such as percent of English language learners, student mobility, etc., are relatively small as to scoring correlation.

We also addressed the school grading system that was put in place in 2011 – the ABCDF Act. We showed that the grades and scoring awarded did moderately track with schools' demographics, but that there were some extremely contradictory grades compared to actual scores received on the tests. That is, two schools scoring approximately the same on the test could be awarded up to a difference of two grades on their PED report cards. Based on the other problems associated with criteria used, weighting, etc., we concluded that the ABCDF Act is often not reflective of students' actual performance on the tests used to score them. This, of course, assumes that the tests, themselves, are reflective of the cognitive levels of understanding they purport to be measuring for the applications they are used for.

All this leads to the question: "How do we help New Mexico Schools to improve their performance regarding the actual cognitive learning as measured by the standardized tests?" We will address the method that CESE has derived and proposes to be used by the PED and the school districts. (We could say much about the use and misuse of standardized tests, but leave that for another day.)

Additionally, realizing that teachers form the first interface for student learning external to the home environment, we examine what has become a rather contentious teacher evaluation method, the impact of which is one of the key elements that determines learning outcomes. This method is based on student growth over the past two years compared to the year the teacher is evaluated. This currently accounts for 50% of a teacher's evaluation. It also assumes that after demographics impact is removed, **all** of a student's learning is attributable to **only** the teacher. This is not specifically addressed in the PED evaluation guide, but becomes implicitly apparent when looking more deeply into the technique used. In fact, it is a bad assumption.

The CESE Proposed Method for School Improvement

CESE initially discovered the fact that there are two "humps" when looking at averaged state achievement test scores versus various demographic factors generally associated with those students demographically disadvantaged compared to other students. These things include poverty, English language learners, minority status, etc. Two "humps" means that if one pictures what is generally called a *normal* distribution, you see one "hump" that is centered in the middle of a curve showing the distribution of test scores. But in New Mexico (and some other states), one graphs the scoring distribution by demographic category (poverty, minority, etc.) and finds that there is a grouping above the middle and a smaller grouping below the middle—two humps – bimodal for the mathematicians). Most of the students in the upper hump are demographically advantaged in comparison to the lower hump. This is not new information, but we could find no evidence that anyone had ever looked at this effect, or precisely what it may mean for New Mexico. In fact, there was a dearth of information at the time we found the effect even at the national level.

What does this mean? Our best hypothesis was something that is quite intuitive, and the same that most people come to when shown the data and given a little hint in the right direction. That is, demographics effect student learning. Well, of course demographics explains much of student learning. That appears to most people to be obvious, especially after the effect is pointed out. Even those who might otherwise deny it, cannot argue when confronted with data that show the effect with absolutely minimum exception in New Mexico.

That led us to the CESE method for improving schools. That is, what if we predicted school performance based on demographics of a school's student body, and if the correlation was strong, we should be able to both determine which schools significantly outperform expectations and which schools do not.

First, we needed to determine which demographic variables really mattered the most. So we looked at correlations of individual demographic elements and standardized test performance. A sample of this is shown in figure 1 on the next page. This is typical for

Continued on page 4

Continued from page 3

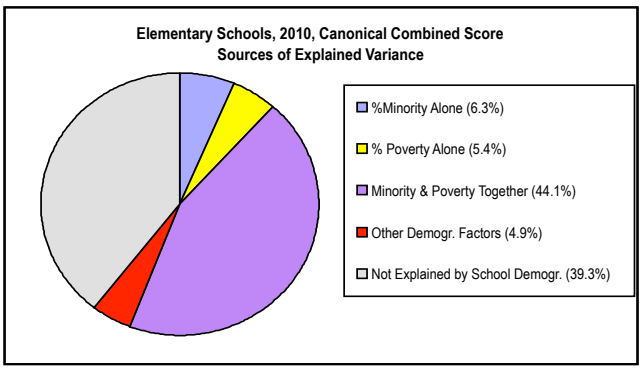

Figure 1. Correlation of various demographic components with test performance

elementary, middle, and high schools. We find that ethnicity and poverty are the largest factors that correlate with performance scores, but that the greatest correlation, by far, is with the combination of ethnicity and poverty, as shown in the pie chart (the large lavender area on the right side).

The other factors that are tracked (the red slice in the lower, left quadrant of the circle, are such things as percentages of English language learners (ELL), Full Academic Year (FAY) student (that is, mobility), etc. **On average, only 20 to 40% of a school's performance is determined by the school (light gray area to the left). This is a significant determination.** (And since in New Mexico, poverty and ethnicity generally are concurrent, one may use either as a proxy for both. But both must be included in calculations, so that the combined impact is present.)

Figure 2 explains what we have discovered in another graphical form and is a prelude to what we suggest as a way of determining **how** schools can be improved.

Note that we have identified three areas that the school impacts students: the first is the school/individual

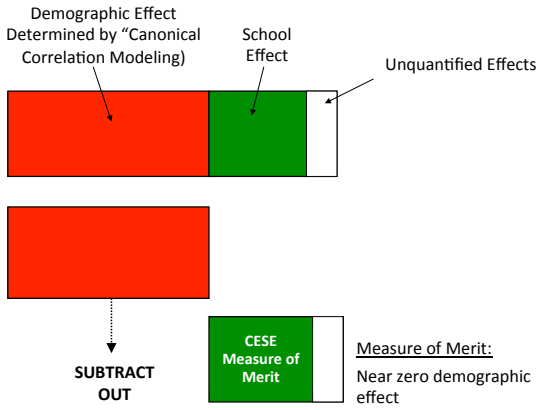

Figure 2. This shows a simplified concept of what the CESE method is doing with demographics.

demographics; the second is the effect of the school alone; and the third are the unknown effects. When we mathematically subtract out the demographic effects, we are left with just the school and unknown effects. This is where we can potentially have impact. In fact, we believe that the school itself can overcome many potentially negative impacts that are an effect of the demographics as the data will show in a moment.

Figure 3 (page opposite) results from graphing elementary schools' average test score for math and science (mathematically combined using *canonical correlation*) versus the scores predicted from using just demographic effects. Each dot is one elementary school, with the vertical axis being the actual combined test score and the horizontal axis being the predicted score. The least demographically advantaged schools are on the left and the most demographically advantaged on the right. The demographic factors used were percentages of minority, poverty (as measured by free or reduced lunch numbers – FRLP), FAY (mobility), students with disabilities, and ELL. Using more or different factors becomes redundant and adds no new information.

The black line on the graph is the predicted scores (a simple regression line for the mathematicians). The upper green line and the lower red line represent "standard error" boundaries that are commonly used to represent where there is a significant variation from the predicted values. That is, any school that scores above the green line is significantly outperforming the prediction, and any school scoring below the red line is significantly under-performing. And we ask: "how good are these predictions?" The answer is: Very good. The correlation is 0.8 (for those mathematicians and scientists reading this). Or, in the common vernacular, you can almost certainly take this specific predictive method to the bank – for New Mexico.

So what have we shown? We have identified schools across the continuum of demographics that are significantly outperforming (and under-performing). This is shown in figure 4 in which the predicted performance (black line of figure 3) is subtracted from the actual performance (vertical axis score on figure 3). This has the same effect as simply rotating the black line and all the schools until the black line is horizontal. The red circled school in figure 3 is the same school as the one circled in figure 4. Also, note the rectangle drawn with its bottom edge near the top, significant error line. This rectangle contains all the schools that are performing considerably above expectations.

By dividing the horizontal axis into four or five areas as shown by the example vertical black lines, those schools to select for observation are easily seen in the

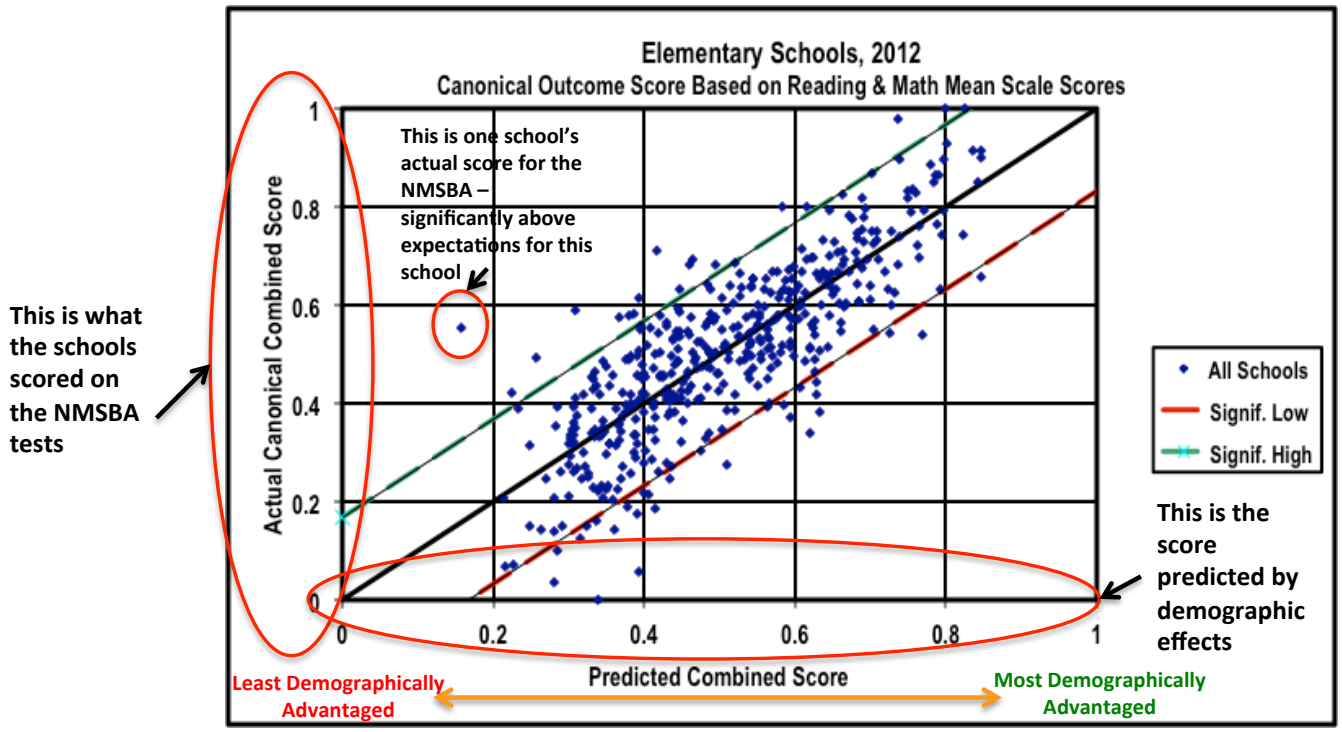

Figure 3. This shows the effects of accounting for demographics in actual observing school scores vs. Scores predicted by demographic considerations.

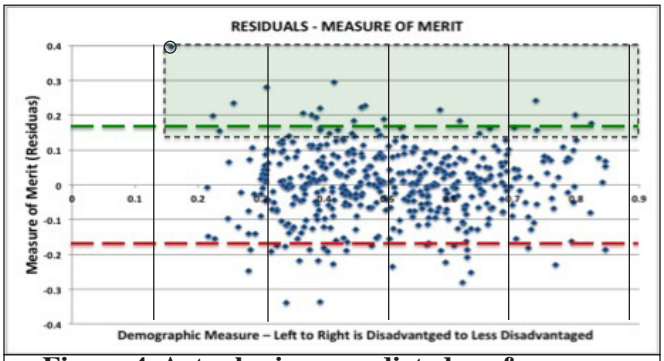

Figure 4. Actual minus predicted performance based on demographics.

rectangle mentioned above. Realizing that there are a mixture of demographic types (e.g., Navajo, east side ranchers, south central recent immigrants, etc.), we select samples of each type of demographic and study those selected schools for best practices. That studying is performed by trained personnel who are expert in teaching, administration, and systems analysis. They then take sufficient time to determine the best practices that appear to be responsible for the over-performance. They compare these to schools scoring much lower in the same demographic areas and determine which best practices apply, passing them downward. And these observers may take a month or more to do this properly. We do not know, because it has not been done, yet. Just pairing the principal of a high performing school with the principal of a low performing school is insufficient to get the right answers. The observers must be well trained and independent of the schools to do this properly. They cannot have a vested interest or unknown prejudice that impacts their study results.

If this formula is followed and the best practice results applied to lower performing schools, we believe that the achievement gaps will start to close and that New Mexico schools will become significantly better. But it will take time and effort to make this happen. And this method does not depend on shooting "silver bullets" at those left-most humps caused by demographics in order to move them to a higher performance level. The state has had over 30 years of that approach to improve education, and it has not happened. This method assumes nothing until it is observed.

It is time to change approaches that rely on best guesses. **Let us look and see what works in New Mexico**.

[Before leaving this specific topic, it is worth noting an important observation from looking at these data. First, each student does carry along his or her individual demographic effects. But in the aggregate, these effects tend to vanish when students are placed in highly advantaged demographic environments. In other words, **there is no physiological reason that the majority of students cannot perform at high, cognitive levels. Also, many believe that the achievement gap is formed by the lowest performing 25% of any school's students. This is simply not the major contributor. Look at figure 3, and you will see where the achievement gap really is—the schools on the left, NOT the lowest 25% of all schools' students.]**

Continued from page 5

IS THE CURRENT TEACHER EVALUATION METHOD THAT REQUIRES 50% OF A TEACHER'S EVALUATION TO BE BASED ON STUDENT GROWTH REALLY VIABLE?

It is clear that the teacher in the school is the primary interface with the student regarding the learning process. We have also seen how there is a rather significant impact (in New Mexico) explained by, or correlated with, demographics. Therefore, to fairly evaluate a teacher's performance, it is necessary to account for the impact of demographics. The PED has done this for that part of a teacher's evaluation based on student *growth*. But, **is it even valid** to evaluate teachers based on the growth of that teachers's students, especially when growth is measured by students' performance on standardized tests?

The answer to that question is NO. Evaluation based on growth will sometimes get the right answer, but will also sometimes get the wrong answer. This can be shown with New Mexico data and with some fairly simple logic. Let us start by looking at the logic involved. Figure 5 shows a table of results derived by applying the logic used in the PED's growth evaluation.

Teacher 1 (Two Years Ago) Student's Perf.	Teacher 2 (One Year Ago) Student's Perf.	ME (This Year) Student's Perf.	My PED Performance (I appear High relative to two Lows, etc.)
L	L	M	H
M	L	M	MH
H	L	M	M
L	M	M	MH
M	M	M	M
H	M	M	ML
L	H	M	M
M	H	M	ML
H	H	M	L

Figure 5. One way of seeing that teacher evaluation based on growth is not always reflective of a teacher's performance

The PED uses the "growth" as determined by test data (often not totally relevant to what the teacher's subject matter is) from the last two years. The growth scores of each student are averaged for the last two years and compared to the current year. So in figure 5, the first two columns represent the average growth scores of all students for each teacher the students had over the last two years ("Teacher 1" and "Teacher 2"). This year's score is represented by the column labeled "ME." To determine my ranking, I compare the aggregate average of all students test scores subtracted from their expected scores (40 on the scale score for the NMSBA test) over the last two years to their similarly calculated scores this year with "ME." Figure 5 shows all possible combinations of comparisons. Note that we are aware that there could be more than just two different prior teachers a student has had over the last two years and that there are really five teacher rankings by the PED and not just 3. But we have looked at an expanded table of four teachers possible and five rankings, and the results are similar. So this table simplifies the structure to demonstrate the point, discussed, below.

To illustrate the key problem with ranking teachers this way, the figure is quite adequate. First, it assumes that **all** test performance by a student is dependent on the teachers when demographics are accounted for. That, as it turns out, is almost certainly a bad assumption, which we will address in a moment.

Note that the table's last column shows how "ME" (I) would rank compared to the other teachers based on student test performance (growth). The first row shows that if the actual scores (High, Medium, or Low) of the student under the previous two teachers, in the aggregate, were low, and I am that medium performing teacher, then it appears that my student's performance is much better than it was under the previous two teachers. In other words, the earlier teachers are actually a significant determinant of my apparent performance. So in the first row, I would appear to be a high performing teacher, though my real performance (or student test results) is simply average. This same logic holds true for the rest of the table. My apparent performance depends on the average performance of students on tests under the previous two teachers. One can look down the last column and see how my performance changes, depending on my students' test performance over the two years.

This admittedly simplified scenario does demonstrate the fundamental problems of trying to evaluate teachers in this manner: **a teacher's level is determined as much by the previous two years of student test grades as it is by the current year's test grades**. I as a teacher have absolutely no control over what the students did over the last two years. And also, this method assumes that I, as a teacher, have total control over how a students will score on a standardized test. In fact, the American Statistical Association says otherwise after researching this assumption (https://www.amstat.org/policy/pdfs/ASA_ VAM_Statement.pdf). To quote from the association's assessment of VAM usage and teacher impact on student test scores, "Most VAM studies find that teachers account for about 1% to 14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions."

So the New Mexico PED assigns a weighting of 50% based on student growth that is only partially under the control of the teacher, and the data indicate that the teacher has little control over their students' standardized test scores from which the growth is measured? Frankly, this makes no sense. Teachers, as well as other professionals in any field are best evaluated by professional observations, **not by factors the teacher has little control over.**

There is not nearly enough room to speak on all the issue, such as the End of Course tests that have not always been working, or the use of student scores for topics having nothing or little to do with the subject "Me" is teaching. Or the fact that, e.g., though biology and physics are both sciences, they are worlds apart in students interests and abilities to score well in standardized test. But figure 5 should get the major logical error of VAM evaluation across.

Finally, let us look at some New Mexico specific NMSBA data that directly illustrates this point. Figure 6 shows the New Mexico average scale scores by grade for the last three years. (We could go back further, but the graphics becomes very cumbersome to read, and the pattern remains basically unchanged.) Here are the keys things to note about these data:

- A 6th grade teacher in math will generally score low on performance evaluation based on a "growth VAM."
- An 8th grade reading and math teacher will consistently score high on performance evaluation based on "growth."
- The 11th grade scores are consistently the lowest – except in 2014. Perhaps the students "cared" more year because they must score above proficient to graduate the next year in order to graduate? (Just a guess. CESE does not shoot *silver bullet* interpretations or fixes.)

We must ask, why are we using a teacher evaluation method that can so easily be shown to be flawed? Part of the answer is that our NCLB waiver required it. Then we must ask, why not minimize this impact to the minimal acceptable weighting until the waiver is unnecessary? So far, no one has been able to answer to this question.

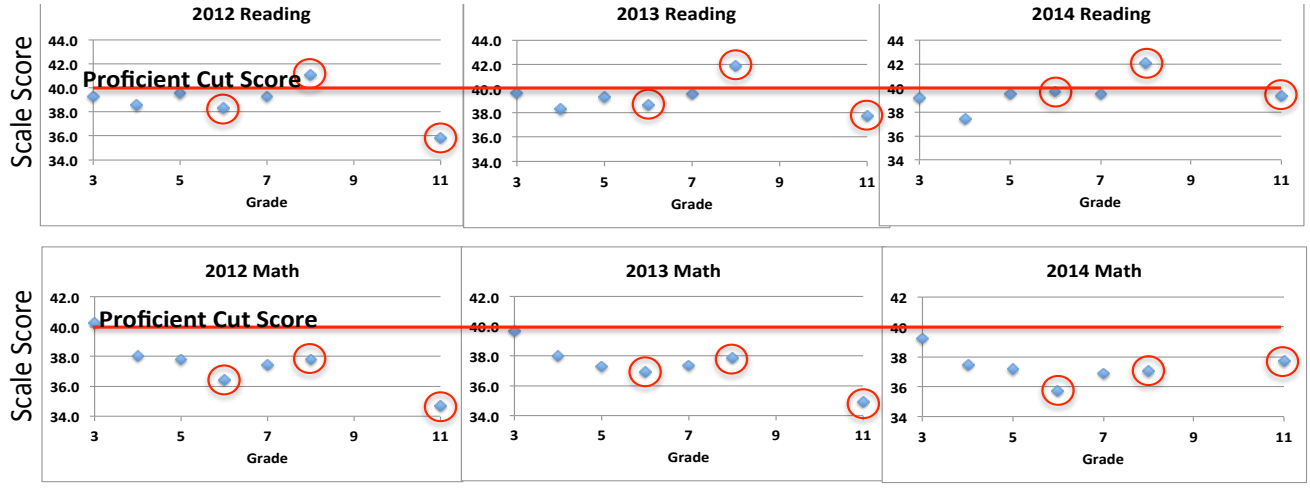

Figure 6. These NM average NMSBA test scores show how the average teacher will be evaluated based on growth. We are fairly certain how the average 6th grade math teacher and 6th and 8th grade math and reading teacher will be scored based on the past trends. We can see this BEFORE a student takes the NMSBA.

http://www.cese.org

Coalition for Excellence in Science and Math Education
803 Maverick Trail SE
Albuquerque, NM 87123-4308

Return Service Requested

Membership dues/Donation Form

Coalition for Excellence in Science and Math Education (CESE)
501(c)(3) non-profit, tax deductible

Dues and Donations cheerfully accepted year round
(Expiration date is found on address label)

Member $25.
Family $35.
Student $10.
Lifetime: $500 Individual, $750 Family.

You may contribute through United Way, PayPal or snail mail.
Snail mail checks to CESE, 803 Maverick Trail SE, Albuquerque NM 87123.

New Membership [ ] Renewal [ ] Donation [ ]

(Please indicate any changes for renewing members. Don't forget your name!)

Name _____ Date _____

Profession and/or affiliation(s) _____
(e.g. Science teacher, member of APSD)

Mailing Address _____

_____

Phone _____ Cell _____ Fax _____

E-mail _____

Most of our communication is by E-mail

Please let *Marilyn Savitt-Kring <marilynsavitt-kring@comcast.net>* know if your e-mail address changes.

http://www.cese.org