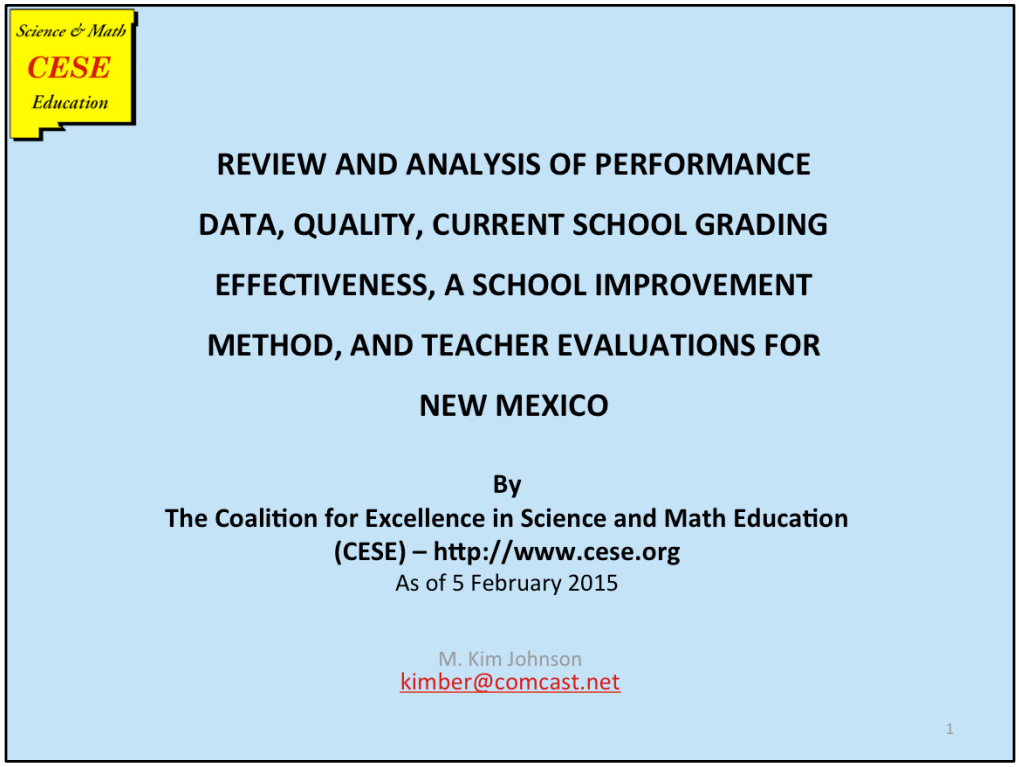
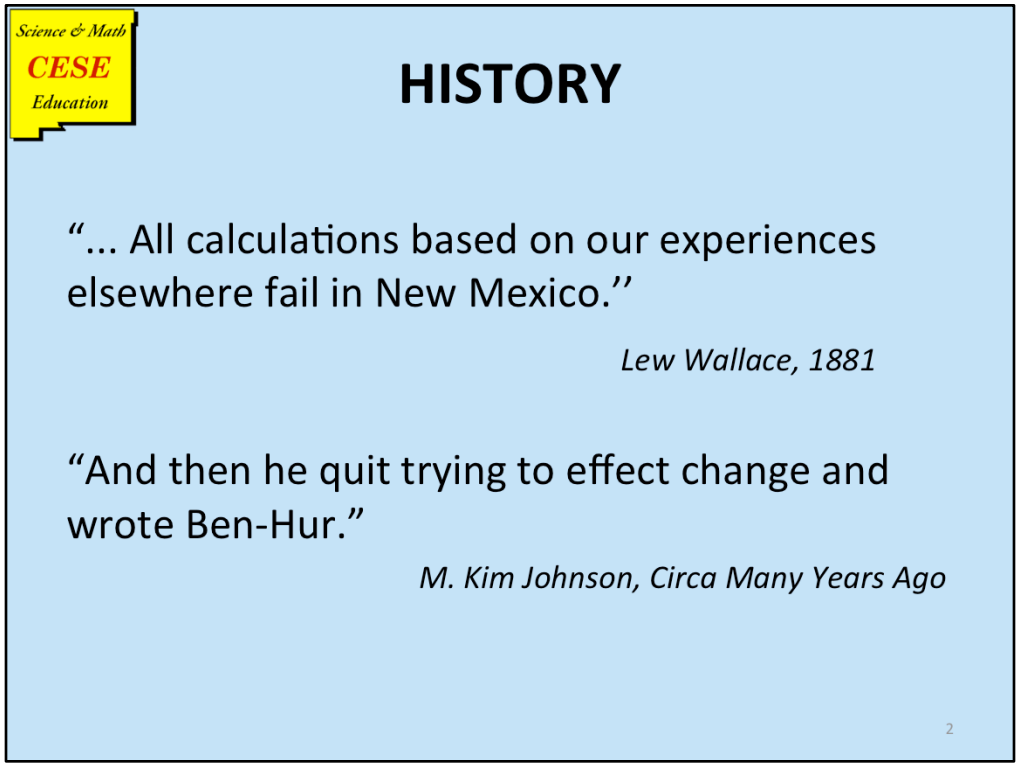


One of the more controversial actions taken by the New Mexico Public Education Department (PED) recently was the implementation of a teacher evaluation system that was based on Student Growth, a measure that was used in the creation of the ABCDF Grading system that has been in use here in New Mexico now for four years. That grading system itself has been the subject of controversy because it is so complex in its structure that it cannot be easily understood even by individuals who have the statistical and mathematical backgrounds to so do. The results of both the grading system and the more recent teacher evaluation system have also shown many cases where the results do not reflect other data.

It should be said from the outset that the PED has been severely constrained in how they have had to implement these two interrelated systems of evaluation, driven initially by legislative guidance in the A-B-C-D-F School Ratings Act in 2011 that was not based on peer-reviewed research (for example in the selection of evaluation criteria that represent meaningful measures of relative performance), and then subsequently by federal requirements associate with waivers from the No Child Left Behind Act, which further restricted the ability of PED to create a fair and useable system.

This presentation provides the background analysis of the Grading system that is in place, and the subsequent roll-out of the related teacher evaluation system. It demonstrates mathematically the flaws in the current system, and more importantly how the resulting data provides no useful information to help schools and teachers improve their performance. The results of our scientifically-based research has been broadly corroborated at a national level by organizations such as the American Statistical Association on the concerns about using the Value Added Model that is the basis for the New Mexico Grading System. Similarly, other prestigious and peer-reviewed studies have shown the influence teachers have on student performance is only in the 10-20% range, whereas all the other factors influencing student performance, including demographic effects simply swamp the ability of teachers to improve any given student performance. Using student performance improvement, therefore, as a measure to evaluate teacher performance is a non-sequitur.

What we hope to do with the analysis you see here is start an informed, bi-partisan dialogue to work toward a system that helps our schools and teachers make improvements, one that respects the professionalism of teachers and recognizes the difficult job they have, as well as the individual circumstances that each unique student faces in the challenges that lie ahead for them as they pursue an educational path and a meaningful life.



New Mexico IS unique in many aspects. Particularly, our demographics are unique. A few other states are similar, but what we can do when we look at New Mexico data rarely translates to what we can glean from other's data. Just because something works in another state or country does not mean it will work in NM. It MAY. But it requires a very careful analysis before implementing. Lew Wallace was correct.

CESE BACKGROUND

- CESE, is a non-profit, non-partisan 501(c)(3)charitable corporation
- Members include National Laboratory personnel and retirees, industrial scientists, educators, parents, college professors, etc.
- We have analyzed New Mexico public education data and policy issues for over 15 years
- Our primary focus is to help improve New Mexico schools using New Mexico unique data

3

Science & Math
CESE
Education

CONTENTS

- NMSBA Test based performance results in the last 7 years – School and student performance results and short analysis
- ABCDF Act, as implemented – Some background, good news, and critique
- How to Supplement the ABCDF Act – We must show schools HOW to improve and close the Achievement Gap
- Teacher Evaluations – The 50% based on Student Growth – This is a serious problem that needs addressing

4

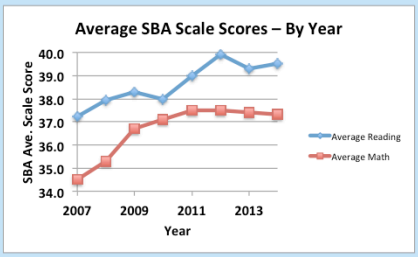
Science & Math
CESE
Education

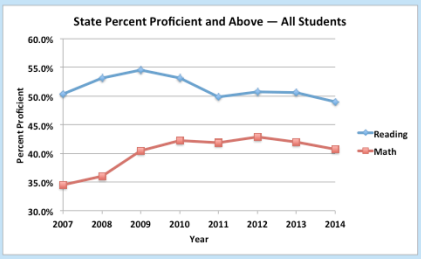
Past Performance from 2007 Using New Mexico Standards Based Assessment Tests

5

Science & Math
CESE
Education

AVERAGE NEW MEXICO STANDARDS BASED ASSEMENT (NMSBA or SBA) SCALE SCORES AND PROFICIENCIES FOR THE STATE BY YEAR SINCE 2007



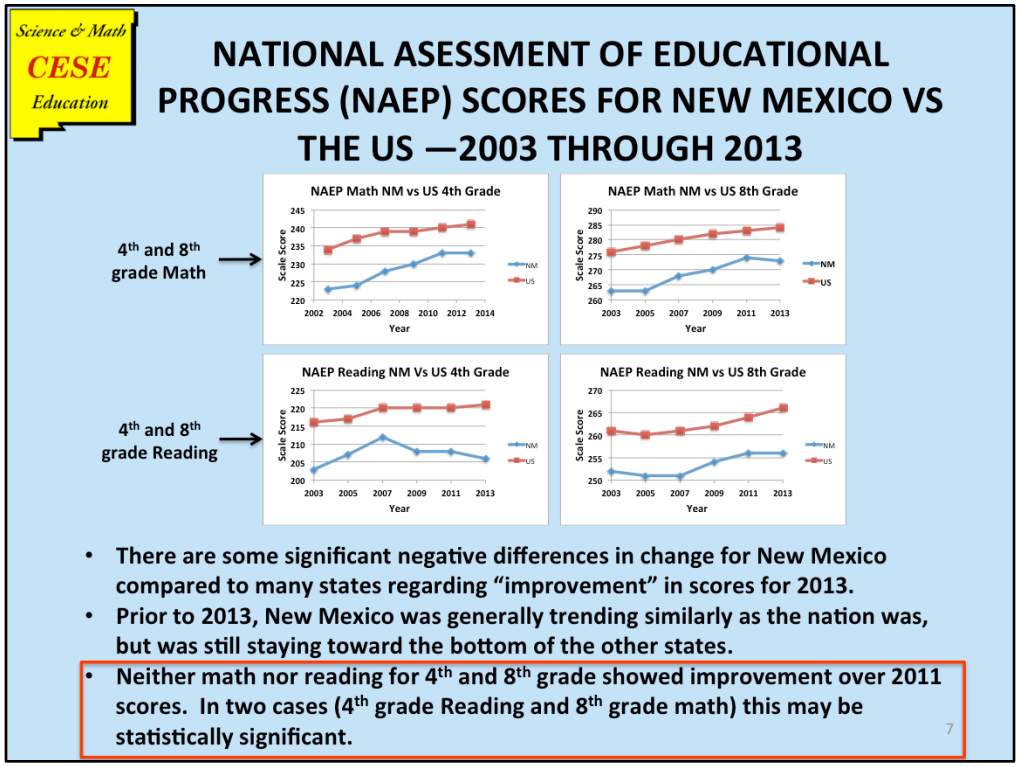


- **Average scale scores have been trending generally upwards for math until 2011.**
- **Reading scale scores have trended upward with an anomaly in 2010 and have slightly decreased in 2013 and 2014 from its high in 2012.**
- **Proficiency percentages follow a similar but smoother pattern as scale scores.**
- **The adoption of Common Core Standards probably does not explain recent trends (NMSBA is based on different standards).**

6

Average scale scores and proficiency levels have trended upwards between 2007 and 2011. After 2011, reading still trended upward, but mathematics trended downward. When comparing with proficiency scores, the similarity is, at is should be, about the same. There are differences because proficiencies are the percentage of students scoring above a certain scale score (40 on a scale of 0 to 80), which makes this a cutoff percentage, rather than a continuum of scores or percentages. Also, NM has often shown a "bi-modal" distribution of scores in standardized tests, where a part of the school population peaks below the average, while another part peaks above the average. By having these peaks move relative to each other, the proficiency can also appear to be not quite synced with the scale scores. But the actual trending for both are very similar with the exception of the scale score for 2012 in reading. This could be caused by the above, or it might be a simple one-year aberration. It is not easy to "back out" the data.

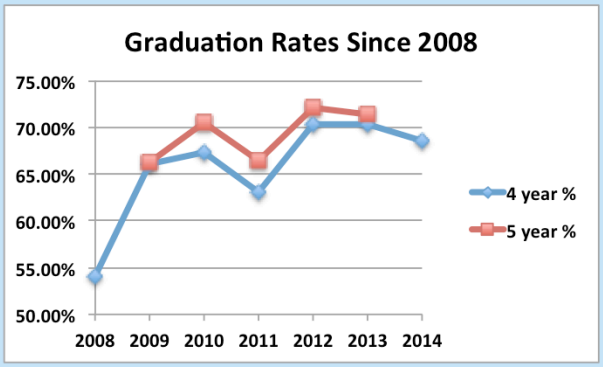
It is clear that the introduction of the use of Common Core Standards (2013/2014) for math and reading appeared to have little, if any, effect on the NMSBA standardized scores. Since 2011, proficiencies for math and reading have trended downward slightly, but perhaps not significantly. However, prior to 2010, both scale scores and proficiency scores were trending upward at what appears to be a significant rate with the exception of 2010 reading, which went down suddenly. This could be a simple one year aberration, also. Overall, scores have not improved since the spring tests of 2011. It remains to be seen if the PARCC test can actually be equated to the NMSBA tests, such that there is a smooth, reliable transition.



The National Assessment of Educational Progress (NAEP) or the nation's report card, as it is often called, is one of the most comprehensive standardized tests given in the nation. It is given every odd numbered year (beginning in 2003) and shows relatively smooth trending, except for 4th grade reading. NM has either decreased in 2011 and 2013 or stayed at approximately the same level.

Science & Math
CESE
Education

STATE GRADUATION RATES – ALL HIGH SCHOOLS



- **The rates have been trending generally upward except for 2010 to 2011, and 2012 to 2014 is level to trending down (4 year rate).**

8

Graduation rates as measured by the current formula that the PED arrived at during the Martinez administration shows that there was some change since 2009. There was a rather significant decrease in 2011 and a corresponding significant increase in 2012. But the 2012 increase is only significant because of the decrease in 2011. The last data point is the 4 year average for 2014. It has decreased from the previous year. 2008 may have been an aberration. It is difficult to tell at this late date. (The NMSBA was contracted to a new testing group and "rescaled" in 2011. It is, however, hard to link graduation rates in 2011 with NMSBA scores.)

There has been some rather significant bragging about how the magazine "Education Week" has NM ranked as the highest gaining state in the nation with respect to graduation through 2013. They use this formula:

Percent Graduating = 10th graders (this year) / 9th graders (this year) X 11th graders (this year) / 10th graders (this year) X 12th graders (this year) / 11th graders (this year) X # Diplomas Spring (this year) / 12th graders (this year)

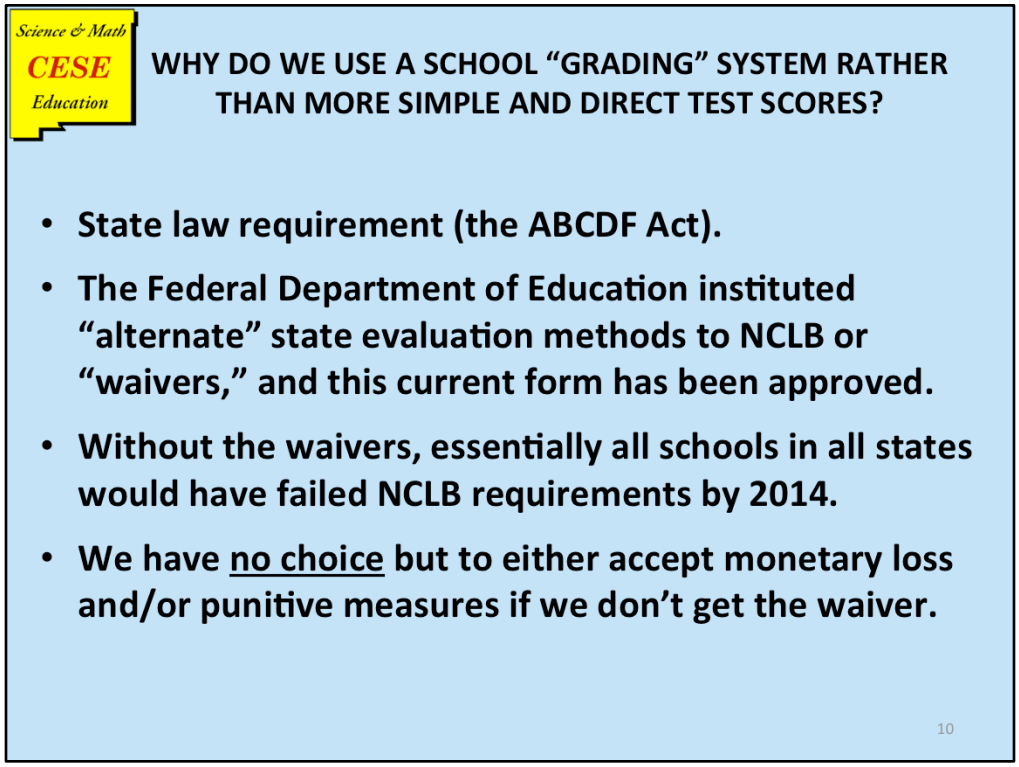
This is a single point estimate, and not a cohort rate, and may contain degrees from students older than any of the grades shown in the formula, since it uses the number of degrees handed out, and does not necessarily track the degrees with the age of a student (e.g., an older student receiving a degree from a GED test.). Also, it does not account for mobility. This is incompatible with the way New Mexico calculates graduation rates and also almost certainly gives a number that is NOT the same as dividing the number of diplomas given at the end of the fall of this year by the number of 9th graders entering school 4 years ago, accounting for mobility When we do look at these data, the greater part of the growth in graduation rate occurred from 2007 to 2009. Unfortunately, we do not know what formula NM was using to account for graduation in 2007, though that data may exist, so we do not show it in this graph. Also, these data may decrease in accuracy the further back one goes, simply because errors that do occur are simply not caught nor scrutinized as one goes back in time. For example, note that the 2009 5-year rate is the same as the 2009 four year rate. Some of this can be attributed to how the data were accounted for, and some may simply be a lack of accounting, or an error in accounting.

Science & Math
CESE
Education

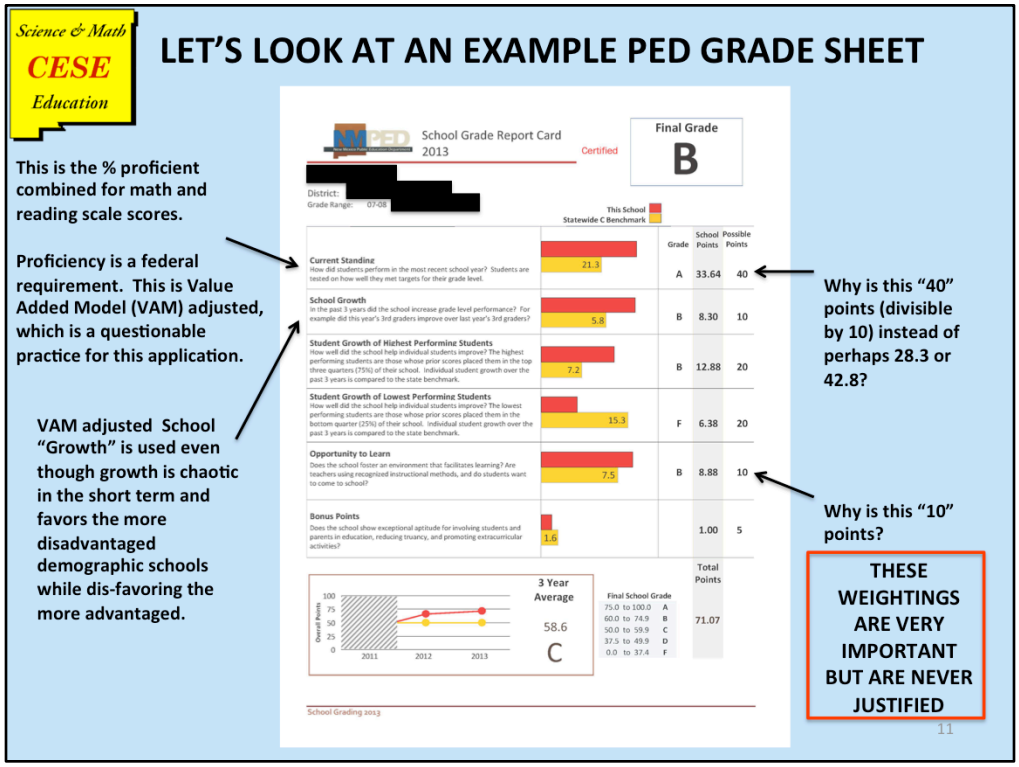
ABCDF Grading System

Analysis

9



We have to be careful in saying just what the "monetary loss" and the "punitive measures" might be without a waiver. It is not clear how the overall penalty might impact anything but Title I money. In fact, when reading the NCLB Act, it appears that only Title I money or use of the money would be impacted, though the other related Titled money could be impacted, too. We have not looked at this in any great detail and probably will not, since there are provisions to make executive driven discretionary exceptions, and these tend to change over time. What would a possible impact be? In 2013 Title I money for NM was a bit over $112 million. Would that be withheld, if any? Would it be redirected in terms of how it is to be used? We are not sure that question has a good answer at this time, or even if it did that the answer may not change over the next few years.



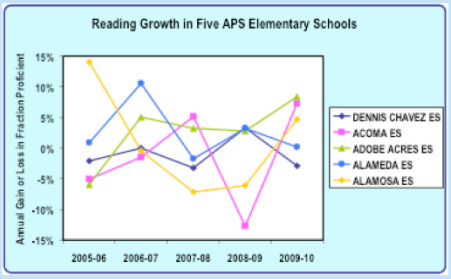
The Current Standing, School Growth, Student Growth of Highest Performing Students (upper 75%, on average), and the School Growth of Lowest Performing Students is all adjusted by a Value Added Model (VAM) based on parameters that are not fully justifiable (more later). Additionally, "growth" has been shown to be chaotic, even over the three years allowed in these areas. For the large scale growth characteristics of all students, there is a slight trend to favor those students with less favorable demographics compared to those with more favorable. This division of growth into two categories appears to have been an effort to work on the "Achievement Gap." However, this is not necessarily reflective of the achievement gap in New Mexico nearly as much as is the difference between schools with less favorable demographics compared to those with more favorable demographics, as will be shown later. It is not clear that these categories used for grading are the most appropriate, and in the case of growth, it appears that there is a reasonable chance of mis-grading a school's performance because of its chaotic nature.

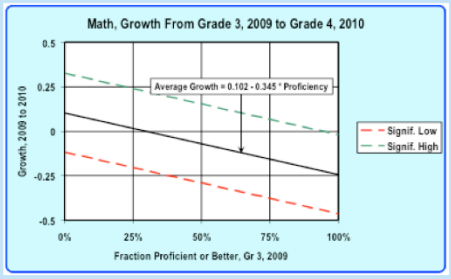
The weighting of these factors is NEVER justified. Being divisible by a factor of 10 or 5 implies that it is a best guess as opposed to a carefully calculated value that applies rigorously to all similar schools in the state.

Bottom line, the question has to be asked how does this information help a school get better? In many cases schools have been driven to improve one particular area on the scoring sheet to raise their grade, with no understanding of what the unintended consequences are if other important areas receive less attention.

Science & Math
CESE
Education

LET'S LOOK AT GROWTH





Short-term growth is somewhat random and NOT a good measure of how a school is performing overall.

Direct growth measures favor the more disadvantaged demographic schools while disfavoring advantaged demographic sectors.

12

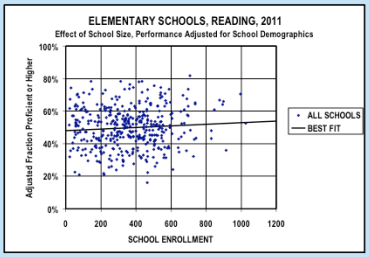
Short term growth for any given school may be rather chaotic with large swings in test performance by students. This is a sample of proficiency percent variation for some random Albuquerque schools. The second graph show that there is a slight overall trend for schools at the advantaged end of the demographic spectrum to show less growth than those at the disadvantaged end. These graphs are from 2010, but the nature of growth does not change over the years, whether looking at proficiencies or scale scores.

Science & Math
CESE
Education

LET'S LOOK AT VAM AS IMPLEMENTED

The NM PED VAM adjusts for*:

- **Proportion of student body that is FAY****
- **School size (total enrollment)**
- **Students' prior scaled scores aggregated by school**

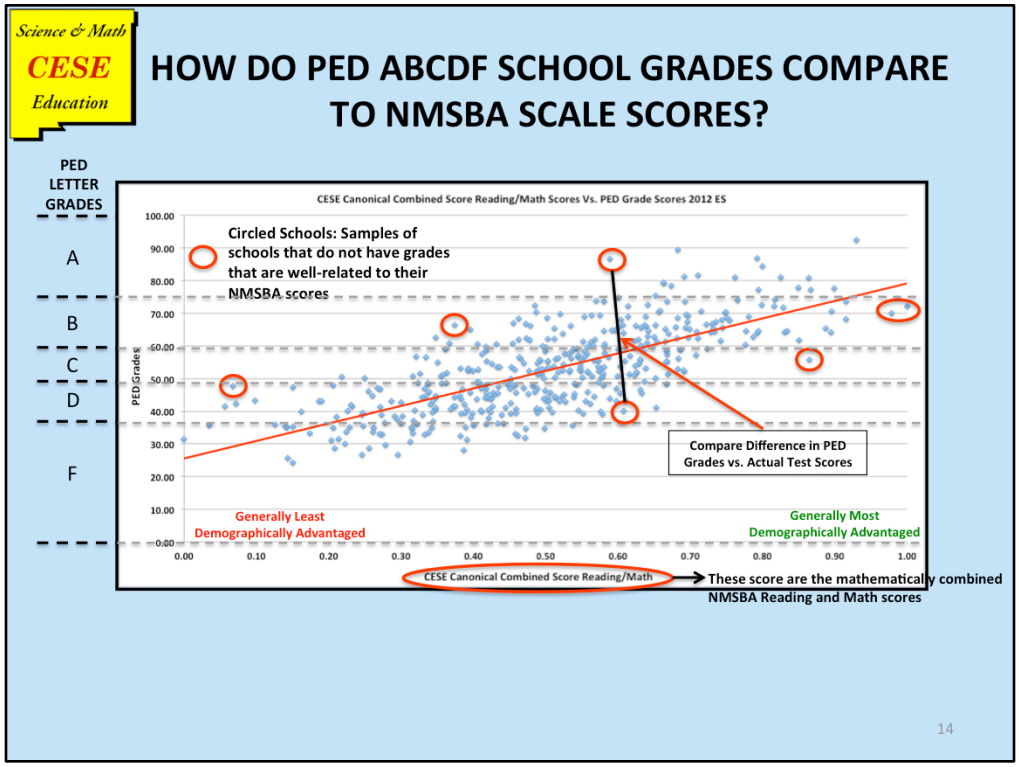


- **School size does not significantly correlate with NMSBA***.**
- **Prior performance correlates with demographics AND everything else. (According to W.L. Sanders, prior performance contains all demographic information, but does it do so in a useable manner?)****
- **FAY provides relatively low correlation to performance.**

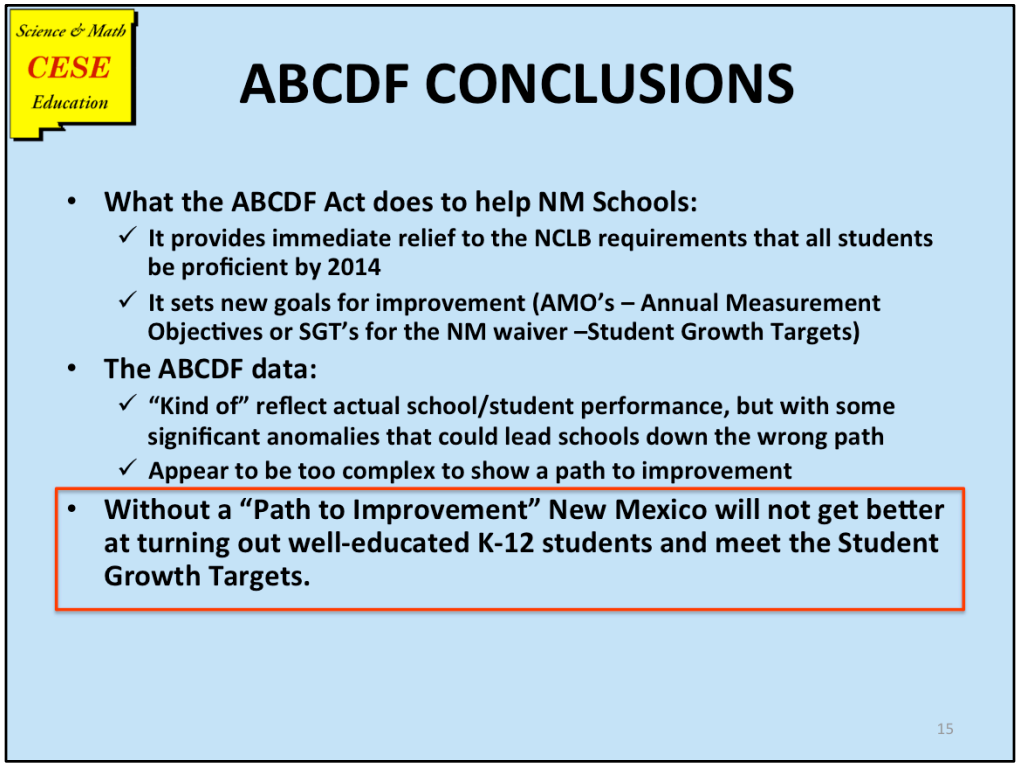
* From the PED "New Mexico "School Grading Technical Guide Calculation and Business Rules" (Date not shown, but posted in 2012)
** Full Academic Year
*** This may be an attempt to "adjust" results for schools with statistically small quantities of students.
**** When prior scores do not correlate with demographic predictions, this correction will probably lead to anomalies.

13

Value Added Models (VAMs) cover a very wide category of mathematical models, usually used to manipulate data so that it is "adjusted" to allow something to be seen that would not normally be seen, or to adjust the real data to fit an another format. It is our understanding that originally, the PED VAMs were used to simply adjust for certain demographic effects, but that these were also adjusted to account to some extent for manipulating results into a standard format. This is probably more than most people want to know about the "arithmetic" portion of VAMs. The point is that using any form of a VAM with the parameters of FAY (Full Academic Year – a measure of mobility of students) proportion, School size (enrollment) and student prior scores have little, if anything to do with how a school performs TODAY. In fact, to correct for school size appears to be nearly meaningless, FAY is barely correlated with student performance, and prior score is so well correlated with student performance that it ceases to have any real information. That is, one can predict almost anything from student performance, even totally wrong demographics. An example will follow further on.



People almost always want to know how well the PED ABCDF scores compare with the actual test scores of the students for schools. This graph shows the PED scores and letter grades on the vertical axis versus the mathematically combined reading and math scores for 2012 elementary schools in New Mexico on the horizontal axis. Each blue marker shows one school – how it was scored by the PED and what the students actually scored, on average, per school. It is clear that the PED scores do not account for demographics, since the combined scores on the horizontal axis correlate very well with demographics, and a "best fit" line through the schools show some decent correlation with the PED scores. But there are some significant problems. That is, the PED scores are used to assign grades to the schools. The red circled schools show where a school is mis-graded. That is, look at the actual score (below on the horizontal axis) and compare it to the PED grade as shown on the vertical axis. Note that the grades often do not match the performance. In fact, there are numerous examples on this graph with only a few red ones circled as examples. Many schools are scoring at what would normally be a low level, but are given high grades, while many schools are scored a a high level, but are given low grades. This is very confusing to the schools and is not at all certain that the specific categories being scored by the PED are really meaningful along with the weightings used.



To properly analyze the effect of the federal "waivers" based on school grading, AMOs (annual measurable objectives), and teacher performance as partially determined by student test performance (to be addressed later in this briefing). A cost benefits analysis should really be performed showing whether or not the state could do away with the grading scheme and not be impacted by the attendant penalties of the No Child Left Behind Act. It appears that this may not have been done using due diligence in considering all the unintended negative consequences of creating a questionable school grading scheme and not have a way to address school improvement, rather than simply saying "We won't be penalized if we do this by the federal government, and the schools will be graded" with no real attempt to tell the schools HOW to improve.

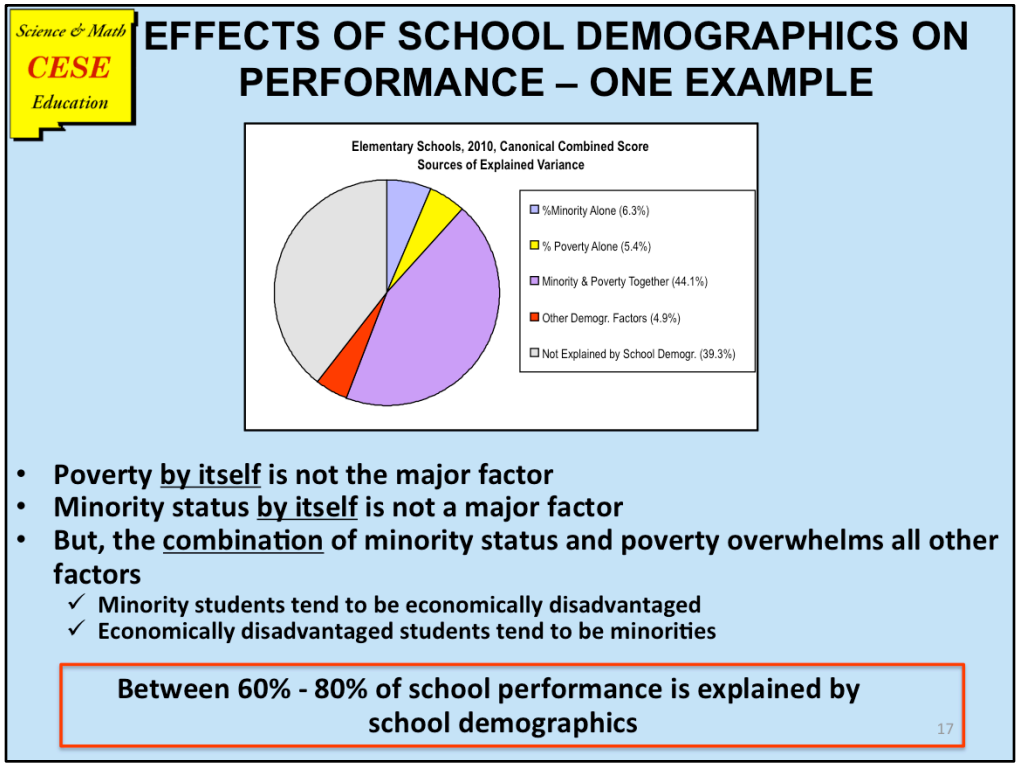
Science & Math
CESE
Education

What Do We Suggest?
–THE CESE METHOD–

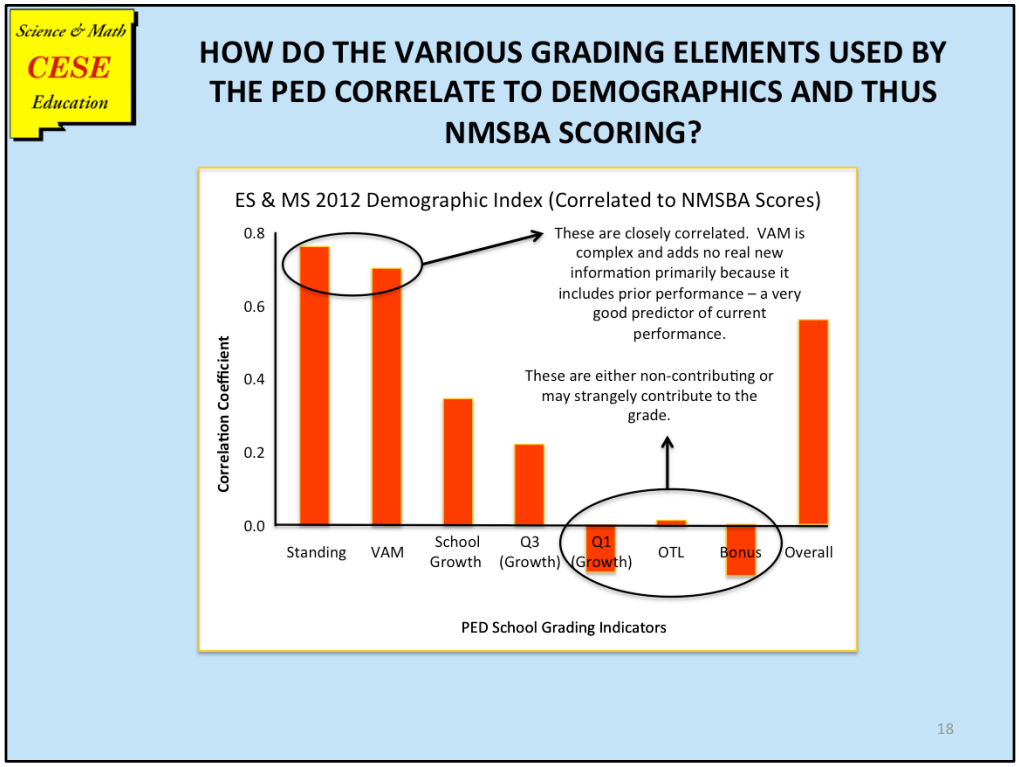
How we might get better

Without Silver Bullets

16

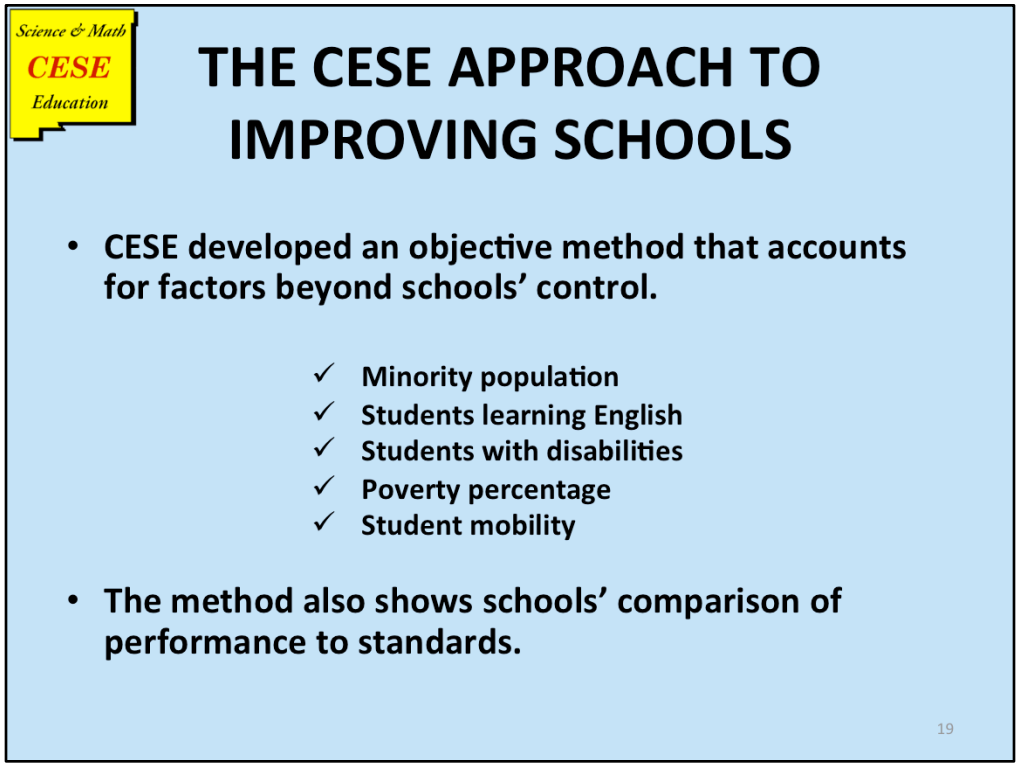


The key things to note on this slide are: 1) minority fraction explains just a little bit more than poverty fraction regarding student test outcome, but the combination explains the majority of the outcome compared to all other demographics measured in New Mexico, and 2) when looking at all combinations of year versus elementary, primary, and high schools, demographics explain from 60 to 80% of the student standardized test outputs. The other 20 to 40% is explained by either unknown demographics not being accounted for, or systemic effects of the schools, districts, and state. The amount of student performance explained by the teachers is (from numerous other studies), generally less than 15%, regardless of the demographics associated with a school or student. This is not the case for each and every school in the state, but it is the case for at least 95%. The other 5% are either scoring significantly higher or significantly lower than the demographics explains. Unknown demographics effects, school, district, or state system effects, or teacher/administrator effects may explain the performance of the schools that lie within that 5% of exceptionally well or poor performing schools.

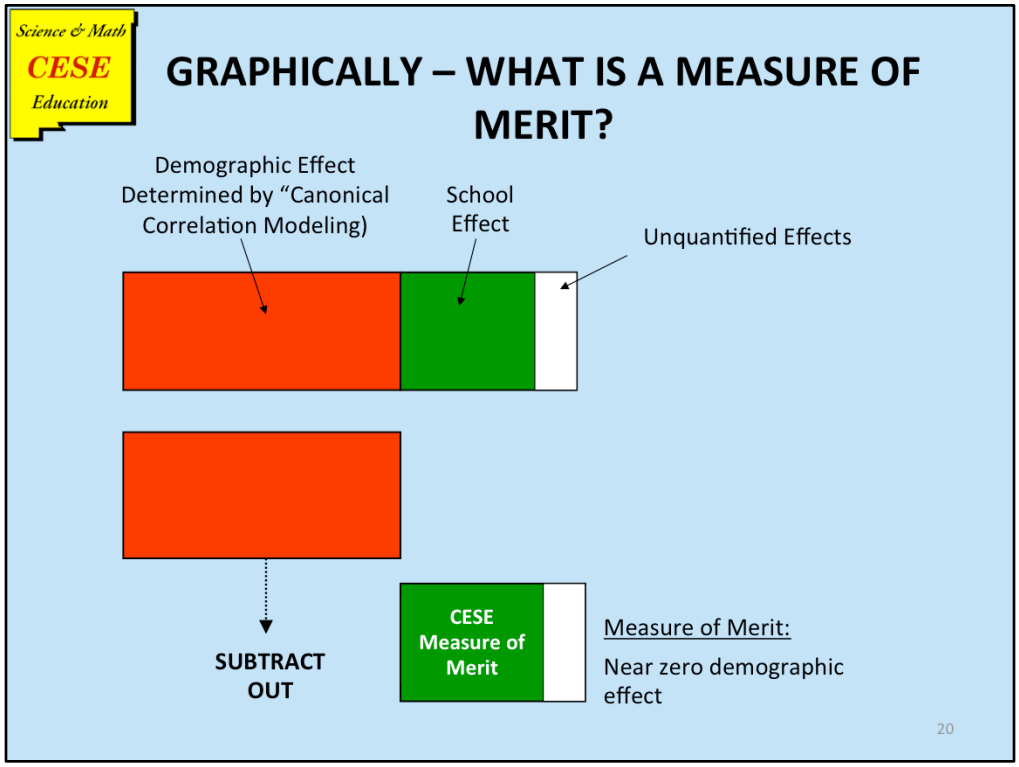


Interestingly, the factors chosen by the legislature to measure in the ABCDF Act and its implementation by the PED are not necessarily good metrics to use to determine how a school's students are actually performing on standardized tests. This slide shows the correlation between each factor and all schools' demographics factor. Demographics are used to represent the expected performance on the NMSBA and has greater than an 80% correlation to the performance. So this slide shows how well the ABCDF Act factors correlate both to demographics and, by inference, to schools' standardized test performance.

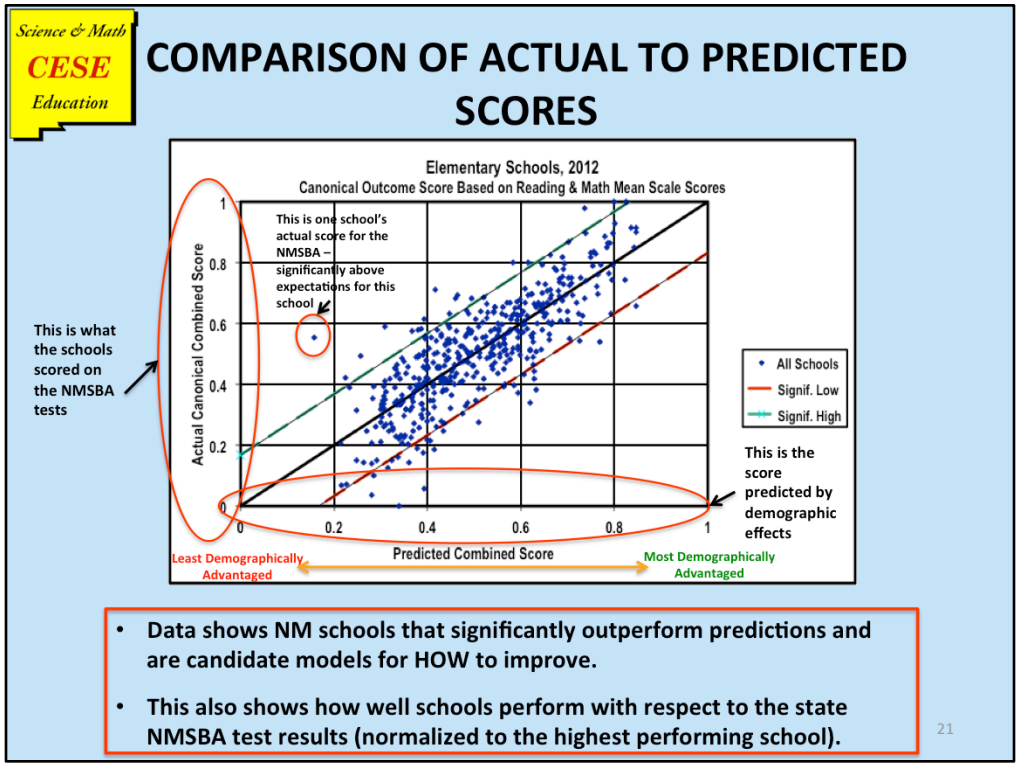
Note that the VAM measures have essentially the same correlation as does the standing (VAM adjusted proficiency percentage). This is redundant and adds no new information to how well a schools is performing with respect to proficiency measurements. School growth still does correlate with demographics, but we have already shown that this is fairly unreliable on a year-by-year basis for any given school. The upper 75% growth rate (Q3 growth as originally labeled) and the lower 25% growth rates (Q1 as originally labeled) are supposed to be a measure of the achievement gap, but really reflect the amount of growth for lower performers compared to higher performers. The real achievement gap is generally for whole schools rather than a school's lower 25% performers. It is not clear that this really provides any very useful information in the aggregate, though it may be something a school would wish to know to make internal adjustments to how it applies its resources. Also note the the Q1 growth is negatively correlated with the demographic index. The better the growth of the lower 25% of the students, the lower the correlation to the demographic index. This is also true of the Bonus points. It is not at all clear how this shows the schools how well they are performing when aggregated to make a single grade. OTL is almost not correlated (within the noise for this particular year – 2012) and adds nothing to showing a school how well it is performing. And finally, the Overall correlation is clearly driven by the standardized test results in terms of proficiencies, which begs the question of what the other factors really add to telling a school how well it is doing. There is clearly no information that tells a schools how to do better, except "Score higher on the standardized test!" And, of course, that is a goal that schools already understand.



The biggest problem, other than unnecessary complexity, with the ABCDF Act is that there is no mechanism to tell the lower performing, or even the higher performing schools how to improve. We believe that this should be the primary function of grading schools – to determine which schools need to improve the most and HOW they should go about doing so. We use the elements shown in a *Canonical Correlation*, a well known and used method of combining explanatory factors that are related to measured factors (test scores) to determine how well a school performs. Note that the factors shown are demographic factors. Also note that no single factor explains performance, nor does the simple addition of the individual factor. However, when they are canonically combined, the cross correlations (cross covariance, actually for the mathematicians) are accounted for such that an optimum linear combination can be made to show the optimum relation to performance, including the multiple performances measured – math and reading along with others in the future, if needed.



The CESE method of determining WHERE to look for best practices to improve schools involves the isolation and subtraction of the demographically correlated effects. The red in this slide represents demographically explained effects, and the green/white are those other effects attributable to the schools or unknowns that impact how well a school performs. The next two slides show this in some detail, but the key point is that to isolate schools performing considerably about expectations, one must subtract out the effects explained by demographics, first. Then those schools that are significantly outperforming schools of similar demographic makeup can be studied to see what it is (probably multiple factors) that causes its exceptional performance that overcomes the demographics effects. The results of the studies would then be used to determine the best practices to use on similar type schools with similar demographics that are not performing at exceptional levels.



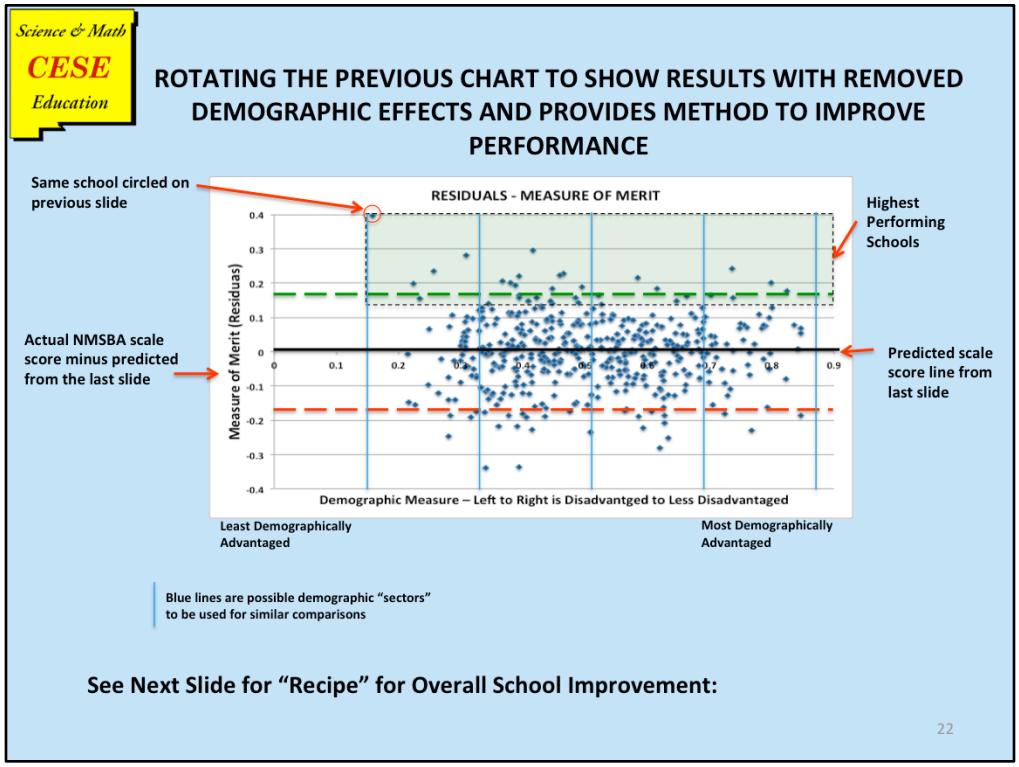
Acknowledgement: this and the next slide appear to be very "busy' and complex slides of this section. But they are probably the most important in showing you how the CESE method for improvement works. The horizontal axis shows the school scores that are predicted using canonical correlation with the previously shown demographic factors as the inputs. Note that the left end represents the least advantaged (demographically) and the right end is the most advantaged. The scores are normalized so that 1 represents the highest possible over the range measured. The vertical axis shows the actual canonically combined scores of reading and math from the NMSBA test (4th grade elementary schools in this case), normalized to the highest score in the state. Generally speaking, this is very similar to adding the reading and math scale scores together and dividing by 2, but that is not the case when the two individual scores are far apart, which indicates there is probably a data error or sever problem at the given school. Each blue marker represents one school.

This graphic shows the REAL achievement gap. As schools become more and more demographically favored, the difference between individual students with the most disadvantaged demographics and other students in that school become less and less noticeable, eventually blending such that there is minimal, if any, way to predict how an individual student will perform based on his or her individual demographic status. This demonstrates that there is absolutely no physiological reason that any student (without mental disabilities that would cause an inability to learn at grade level) cannot learn at high levels in this state.

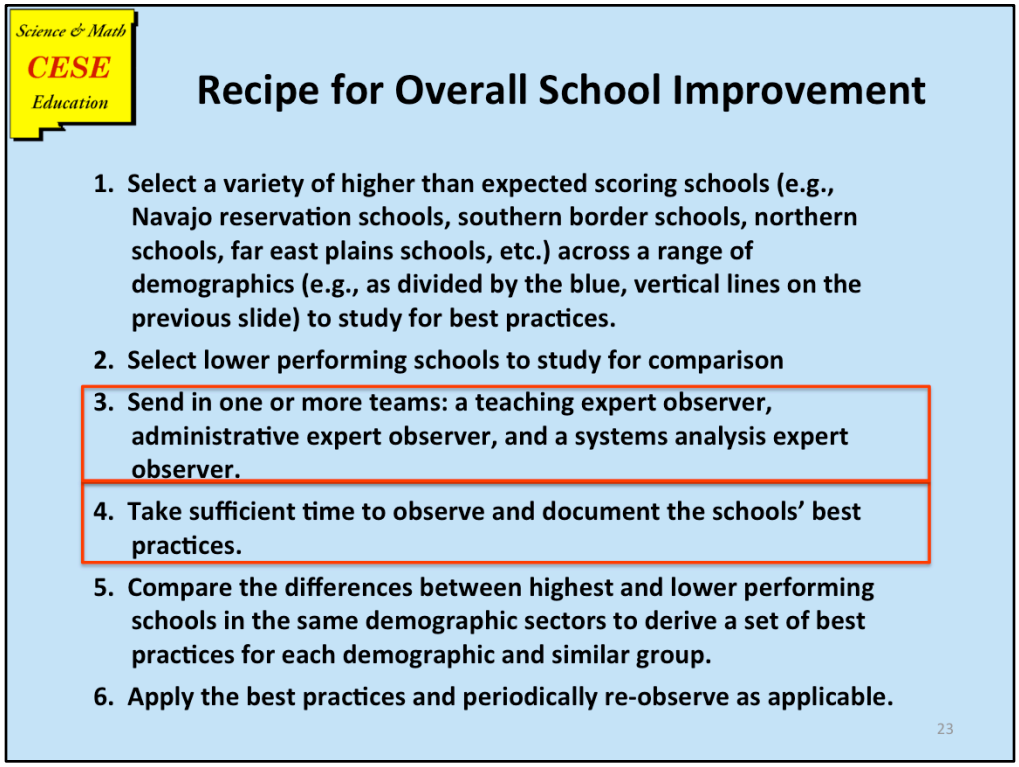
Now, note that there are some schools, such as the one circled in red, that significantly outperform other schools within the same demographic range. In fact, there are many such schools in this state that perform near or above a statistically higher level (greater than one standard error) that is predicted. Additionally, there are about the same number of schools that are performing significantly lower than is predicted, with the rest at some level in between. This suggests a way to improve performances of schools by carefully observing those that are scoring at or above a significantly high level based on the demographically predicted level. These observations should be performed by trained observers in classroom teaching, school administration, and systems analysis. The observations should take a sufficiently long enough time to provide a definitive guide to what these schools are doing with respect to best practices that allow them to overcome demographic effects to score as high as they do. At the same time. The significantly lower performing schools and a sampling of mid level performing schools should also be observed to see what they are doing that apparently causes their lower performance. These observations should be grouped according to slices of demographic similarity. For example, there may be five divisions of demographic similarity shown here. Or there may be more, depending on initial surveys. Additionally, it is probably the case that the not every school in a given demographic slice of above significant performance is doing the same things to excel. For example, an outperforming school on the Navajo reservation may be doing some things differently than would a school on the southern or eastern border of the state to achieve the higher performance. This is something that will have to be studied and accounted for in selecting which schools to observe.

The end result is to have a set of best practices for the various demographic slices and perhaps culturally different schools that can be applied to schools at lower levels. And that is not just to significantly underperforming schools, but to all schools. We understand this is not a trivial undertaking, but believe it will be far better than the "Silver Bullet" approach that has been used for so many years to accomplish very little. A great deal of studying has been performed, but so far, the New Mexico schools have not progressed significantly and show a tremendously large, and in our opinion, unacceptable achievement gap linked directly to demographics. We believe that this method will pave the way to closing that gap and significantly improving schools in the state, regardless of any individual's silver bullet approach.

(Note: this chart also illustrates how the use of prior performance to predict current performance with a VAM is potentially wrong. Note that the underperforming schools performance for the year shown would cause a VAM to predict the school SHOULD underperform. In fact, the school should NOT underperform. It is simply not appropriate to use prior performance for any given school or student as a VAM parameter, except for those schools close to the middle, black regression line.)



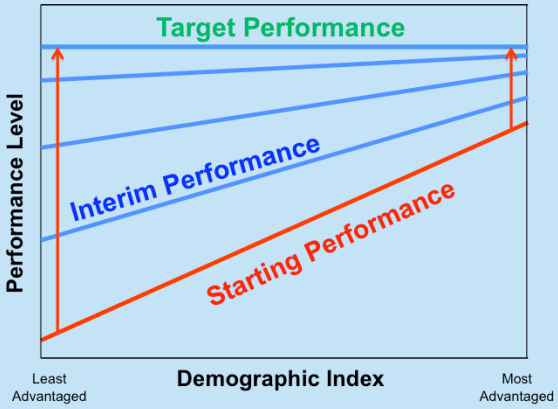
This slide is derived by subtracting the predicted canonical scores from the actual canonical scores. In essence, it rotates the previous graph such that the demographic effect is not seen, however, the effect manifests on the horizontal axis. The light green shaded box at the top represents candidate schools to study for best practices based on their "Measure of Merit" which is that difference in predicted from actual scores by school. Note that the red circled school in the upper left of the shaded box is the same school circled on the previous slide.



This slide summarizes the previous discussion, step-by-step. It is very important to again note that not just any set of observers will do. They must be expert in the areas they are observing, trained observers, and they must contain one systems person to provide an overview input. Also, this cannot be done in a day. It will take time – perhaps several months per school (though it may take less time) the first time through after surveying the schools for appropriate selection.

Science & Math
CESE
Education

WHAT ARE THE ULTIMATE GOALS?



- **To lift the disadvantaged demographic end so that performance is minimally dependent on demographics and any other factors**
  - ✓ **We predict this provides a path to help close the Achievement Gap**
- **To raise total performance so that all students perform to their potential**

24

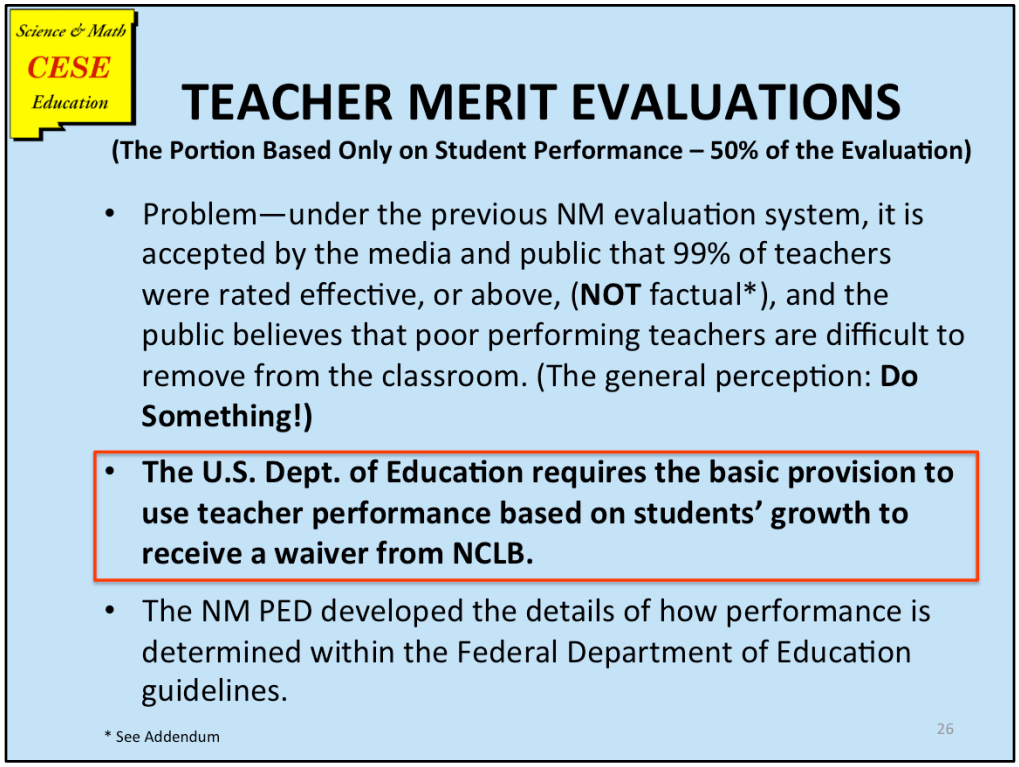
This is where we think the state should wind up compared to where it started. It is not a short term process, but it should be a lasting improvement and do away with, if not significantly eat into the really unwarranted achievement gap.

Science & Math
CESE
Education

TEACHER MERIT EVALUATIONS

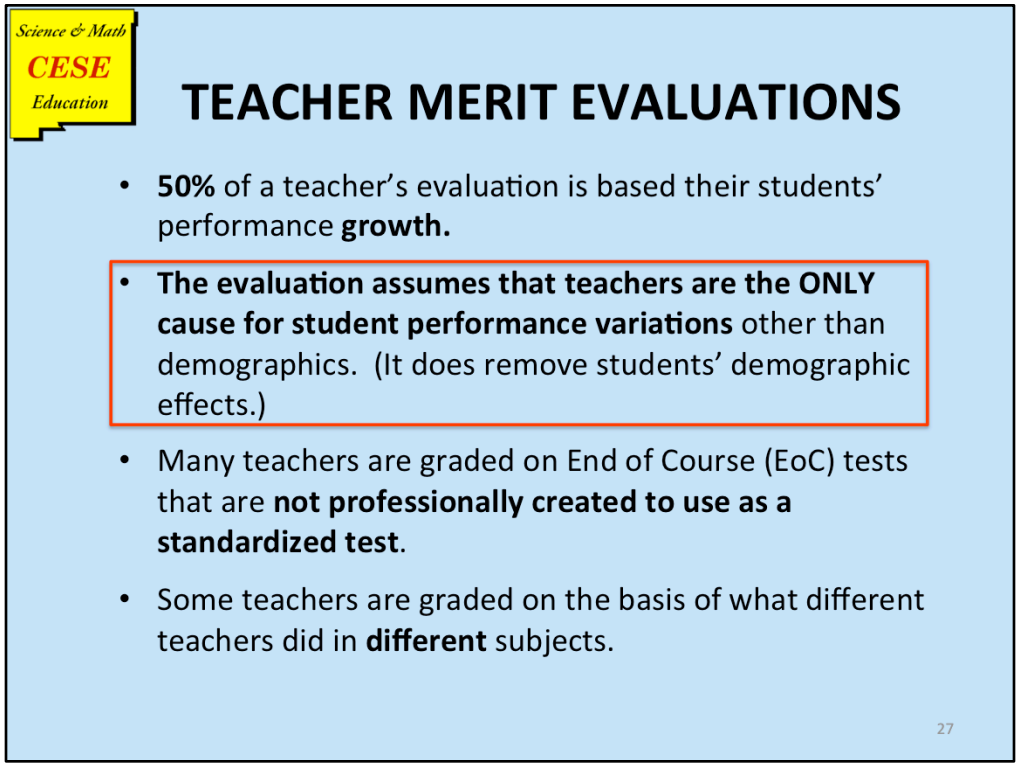
The Portion Based Only on Student Performance – 50% of the Evaluation

25



The use of a 99% rating of "effective" or higher for teachers since the three tier system went into effect has been broadcast far and wide with no one actually validating that number. There was a survey a few years ago that I spoke about with Walt Murfin, former CESE statistician who passed away last spring. I have actually seen some of those results recently on the web, but there were only a few school districts represented, and the net number evaluated as effective or higher was closer to about 80%. But that is from memory and the small sampling I saw recently. In fact, the 99% number seems to have come from the PED, itself, but in response to how many teachers pass the PED Tier I to Tier II and Tier II to Tier III licensure level. The actual number quoted in a political advertisement by Governor Martinez was 99.8%, and she used it in the sense of individual teacher evaluations, not licensure between tiers. This was a misuse of the data. The 99.8% licensure passing percentage is backed up by the PED through an IPRA request made by the NEA that explains it the way I have. Of course the licensure between tiers is a very high pass rate. In order to be approved, a teacher must meet certain criteria ahead of time, be approved within their district (i.e., the dossier required for the PED to review must be polished), and then reviewed by two individuals who contract to the PED for their review. One would expect that any teacher who passes all the wickets before ever submitting anything to the PED is almost certainly going to pass to the next tier. The system is set up to filter out those who would not be eligible. But to conflate that number with the number of teachers at the district level scoring as effective or higher is simply put, lack of performance of due diligence. However, the press, the Governor, and perhaps even more importantly, the public appears to take that number as being true and factual. So, based on faulty data, there is a demand to "do something." This provides added support to grade teachers based on students' scores on standardized, or supposedly standardized tests.

Additionally, the federal government, as backed by President Obama and the Secretary of Education, Arne Duncan, are insisting that teachers be evaluated based on standardized test results in order to get a waiver from the No Child Left Behind Act which requires the impossible goal of all schools and all students scoring at or above the proficient level on state standardized tests. If a state does not do all the things necessary for the waiver, then the federal government can dictate numerous ways in which federal Title I money can be spent (over $100M in New Mexico per year). The details are laborious, but this amounts to a penalty that most states were not willing to take on, though several have said "No!" to the waiver. New Mexico, and specifically the Secretary designate of Education, Hannah Skandera, decided that the waiver was needed. However, that waiver's requirement that some part of the individual teacher evaluations be done based on their students growth was a requirement that she and the Governor apparently agreed with. In fact, the federal requirement is less strict in terms of how much of a teacher's evaluation should be based on their student's growth than was Skandera. When Federal Secretary of Education, Arne Duncan, said that this requirement would be relaxed for a year, because of the many implementation problems, Skandera said: "**[**New Mexico will] not be breaking the commitment we made in our waiver. Delayed accountability won't help the students of New Mexico." So, the PED has implemented the waiver requirements and this has been accepted by the federal government. But Skandera believes so strongly in them that she will not relax the requirements to grade a teacher based on standardized test results.



Consider that the New Mexico teacher evaluation system requires that 50% of the evaluation be based on teacher performance, the maximum amount recommended by the federal Education Department. (The minimum is 33%.) New Mexico chose to use *growth* as the measure of performance. That is, each student's growth performance (compared to a scale score of 40) for the last two years is averaged and compared to projected (scale score of 40) for all students for the current year. The resultant comparison actually does negate the effects of any given student's demographic situation if all have come from the same or similar schools. But we have already seen that growth per year varies significantly, on average. But the biggest, and most obvious fallacy in this method is **the assumption that even with demographic effects removed, ONE INDIVIDUAL TEACHER PROVIDES THE ONLY IMPACT ON STUDENT PERFORMANCE.** This has no basis in evidence, though one can cherry pick studies and find some that conclude this. However, when reading between the lines, the real answer is that teachers can have a profound effect on some students and little effect on others. The effects may be positive or negative. But the largest effect is probably the cumulative demographic plus systemic effect. Truly outstanding teachers can promote growth better than an average teacher as can a truly bad teacher promote lack of growth in learning compared to an average teacher. But most teachers have a very complex relationship with each student that depends on a number of possible systemic factors. (More specifics in slide 31.)

Additionally, in New Mexico, teachers are often being graded on the results of End of Course (EoC) tests that are not professionally designed or vetted. Yes – multiple teachers from New Mexico are used to construct and review the tests and they are first given to selected students on a trial basis. But they are not professionally constructed, and they do not appear to reflect the results of professionally constructed standardized tests, in general.

Furthermore, many teachers students' "growth" are compared to subjects that the teacher is not teaching. For example, a science EoC (End of Course test) may cover biology, physics, chemistry, etc. But these are very different subjects, and often are really not totally related to the the subject the teacher is responsible for (compare physics and biology, for example.) Sometimes, there are not two years to compare with, and the comparison data are calculated, but not demonstrated to be accurate. For teachers at the 3rd grade or under, separate tests are given, and no attention is paid to potential systemic causes of lower than "effective" grades. In short, a teacher may be graded on a mathematical guess or even by comparing to a subject not directly related with the one the teacher is teaching.

Finally, none of this methodology has been truly and fully tested in New Mexico, which as first stated in this briefing, is different than other states. No attempt has been made to control for that nor has any attention been given to the fact that teachers only have some control over the learning done by the student – not 50% by any recent measure.

Science & Math CESE Education

COMPARING TEACHERS TO TEACHERS
FEATURES AND CONCLUSIONS

The hidden assumption: for this VAM approach, only teachers control how well a student is performing. The inescapable conclusion: An average teacher's ranking is determined primarily by the previous two teachers' performance.

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

L = Low Perf., M = Median Perf., H = High Perf

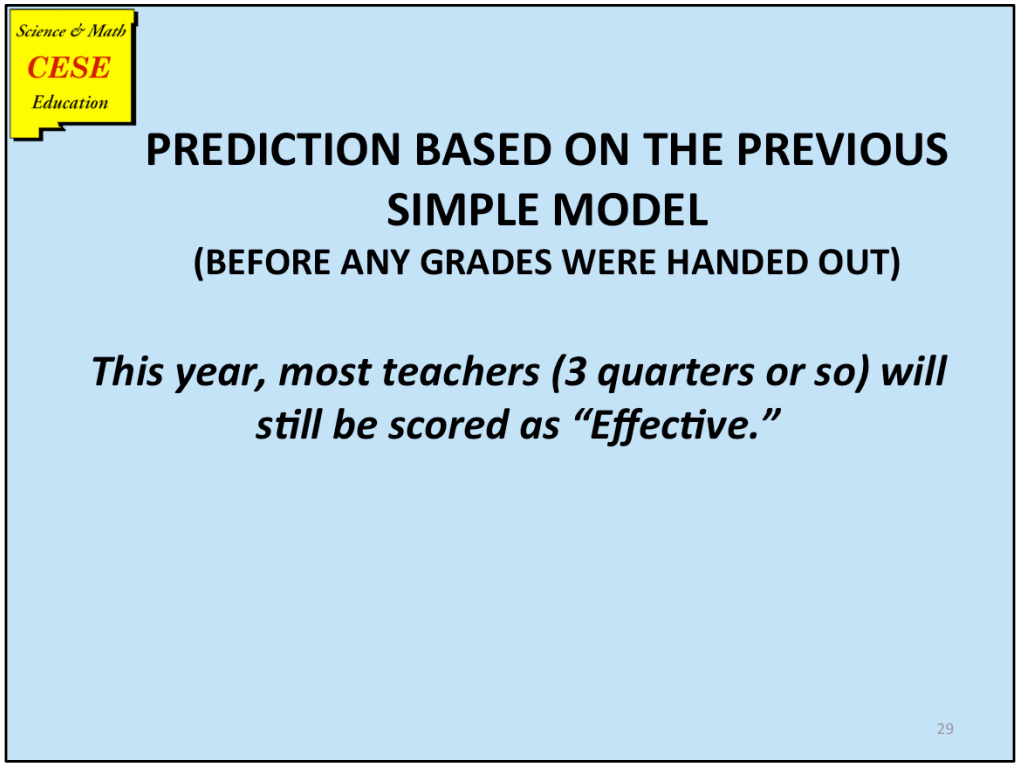
28

If you assume that a student's learning rate is solely determined by the teacher, a bad assumption, but the one that is used in the New Mexico performance portion of the teacher evaluation, then it becomes apparent that any teacher's score will be solely determined by what the teachers for the last two years have done. This table says to let us presume that assumption is true. What does that mean for ME, an average teacher (of which one would expect there to be many. We can do a simple comparison and look at all possible student performance variations from previous teachers, presuming that these consist of high performing, medium performing, or low performing teachers. So this table shows those possible combinations and how I, an average teacher, will compare based on the previous two teachers.

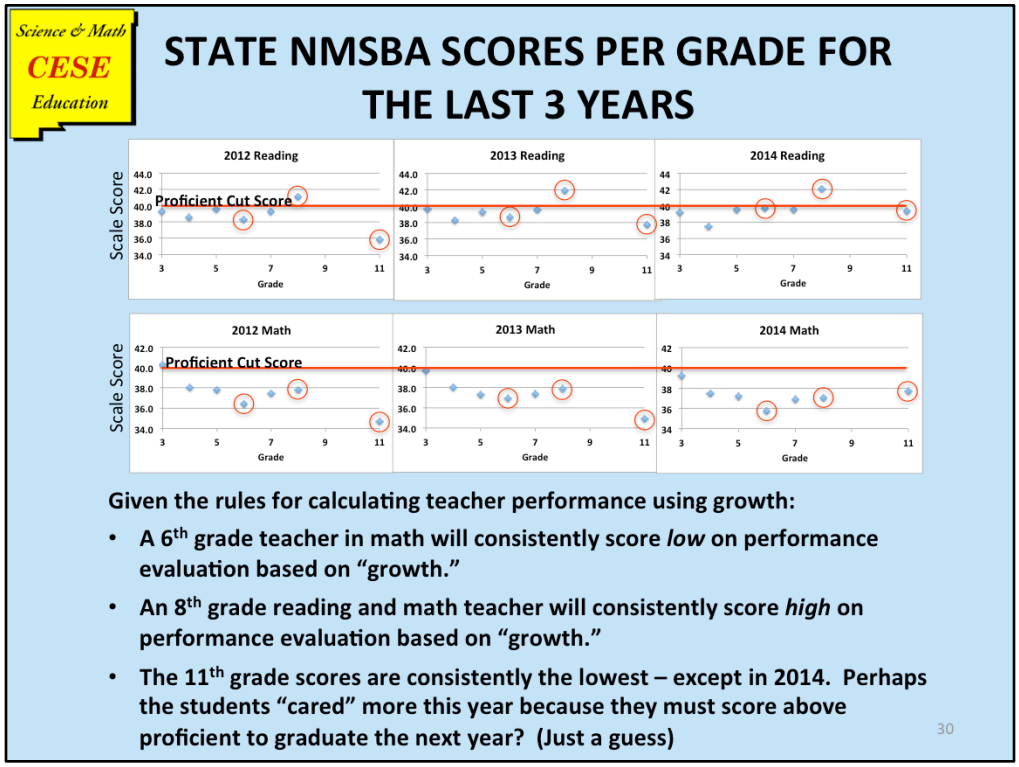
In the first line, you can see that we have the one possible combination of the two previous teachers performing at a low level (as would be indicated by the students' average growth for that year), and me performing at my average medium level. Based on the other two teachers, it appears that I am a high performing teacher, **even though I am only average.** The next line shows the possible combination of the previous two teachers students scoring at average and low growth, respectively. This makes it look as if I am somewhat higher than a medium performer, though that is not the case. And so on until we come to the last line where it looks like I am a low performing teacher, because the previous two teachers' students had shown more annual growth than they did under me. So my evaluation of performance based on my students' previous two years growth is governed as much or more by those teachers than by me! Even if I were outstanding, there is a 1 in 9 chance I would be scored as just an average teacher.

This (though simplified in this example) is happening, and is simply based on flawed reasoning. Plus, the performance data shows this (second slide, following).

Can this work in any circumstances? It probably can pick out extraordinarily high and extraordinarily low performing teachers. But all else is suspect.

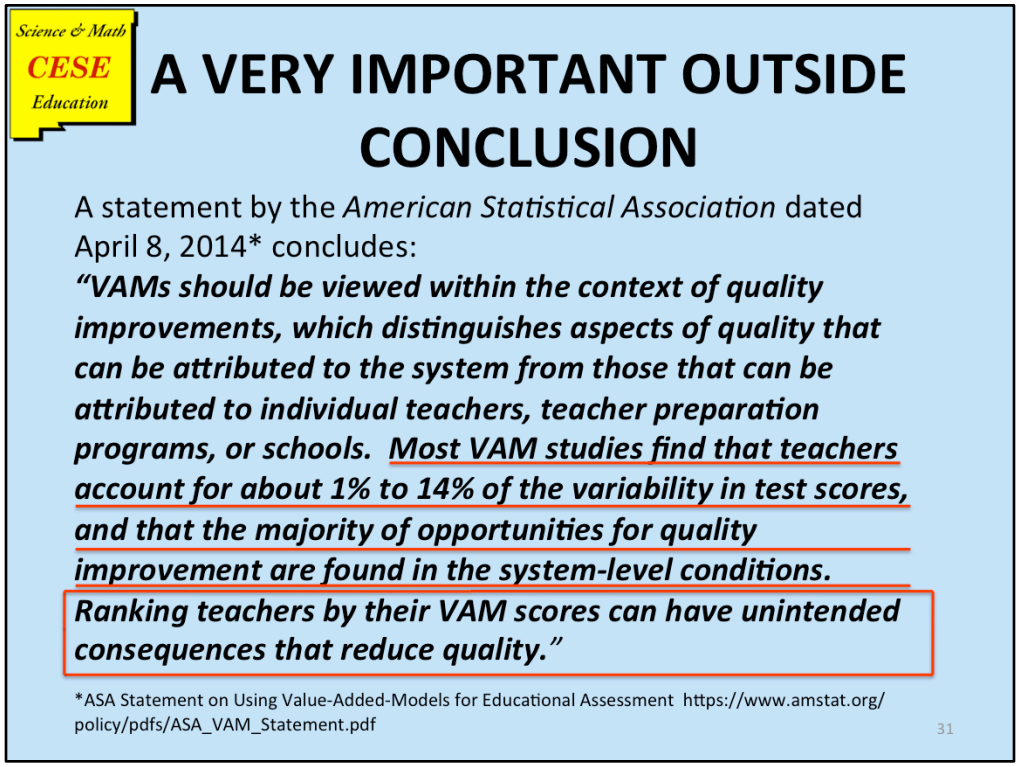


This slide had to be included, since that simple model shown on the last slide led to this prediction. And the teacher performance evaluation results for the 2013/2014 school year are very close to this. Of course that is not proof, but it doesn't hurt!

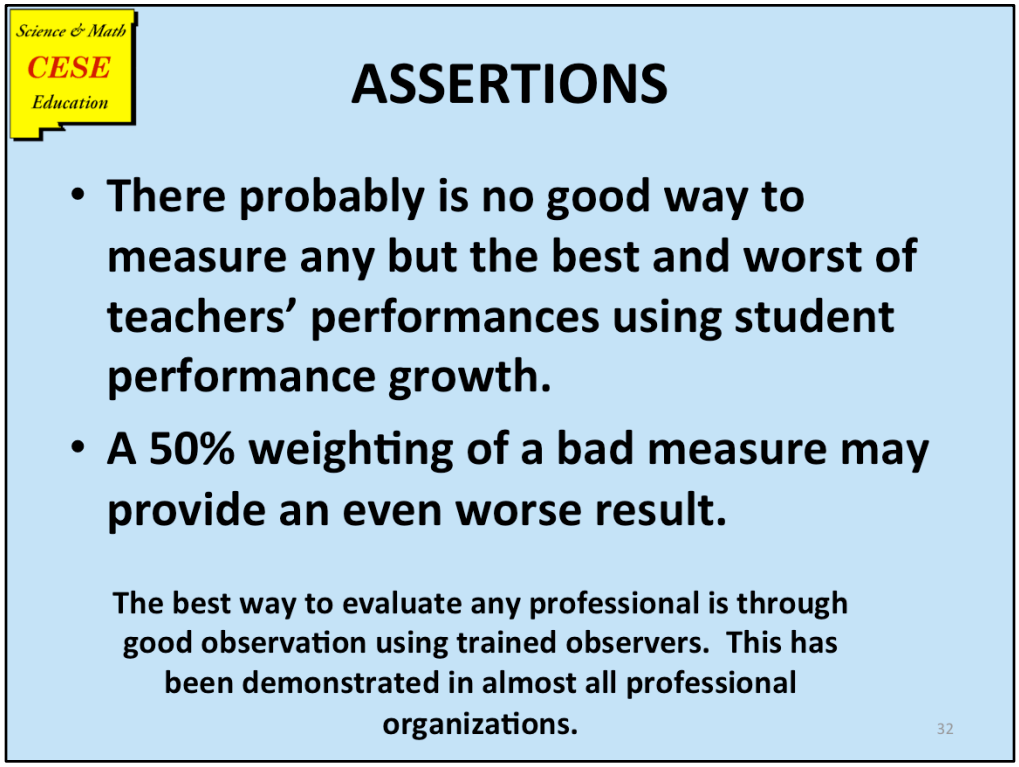


These data actually show the fallacy of grading a teacher based on the previous two years' growth. Note the first two red circled points on each graph. Each of these points is the average scale score for the NMSBA taken for grades 4 and 8 (each point is an average scale score for other grades, as indicated.) Grades 4 and 6 are particularly interesting, since they show that the average 6th and 8th grade teacher in both reading and math (except for 2014 reading) has students who would cause them to score low for the sixth grade and high for the 8th grade. This is the *average for the state. This means that if you are a teacher and wish to have the odds on your side that you will be scored at a high level, teach 8th grade math or reading. If you do not wish to be scored low on a student growth basis, do NOT teach the 6th grade!* Of course, this is ridiculous as far as choosing what grade level to teach at, but it illustrates the fallacy of grading teachers on a student growth basis. Other things are happening that are not captured by how well or how badly the teachers are performing. And they are big enough drivers that they appear to overcome a teacher's capabilities at instilling the testing capabilities and knowledge by the NMSBA. (Prior years show the same trends, but were left off for readability reasons.)

One might also note that grade 11 is circled. One can see that from 2012 to 2014, the 11th grade scores increased. This could mean that students at the 11th grade did not start taking the NMSBA test seriously until they determined that it was key to their graduation. That's just a guess, but it raises the same question about earlier standardized testing that does not count towards students' grades. Do they care enough to perform at their best?

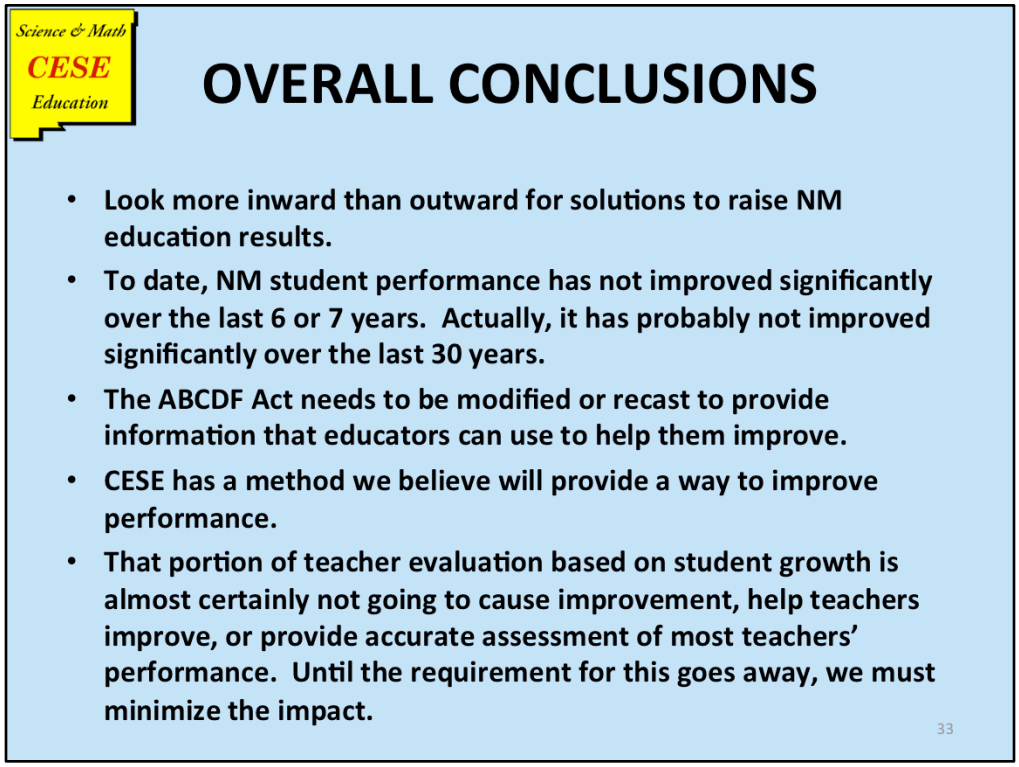


This is a fairly conclusive statement. The ASA performed this study specifically to answer this question, theoretically without any pre-bias. This is a world recognized expert organization. The results in the red underlined text are very straight forward – only 1% to14% of the variability of test scores are accounted for by teachers. THIS DOES NOT MEAN THAT TEACHERS DONOT CONTRIBUTE SIGNIFICANTLY WITHIN THE SYSTEM TO STUDENT LEARNING – JUST A SMALL AMOUNT TO STUDENT STANDARDIZED TEST SCORES. Therefore, what sense does it make to use standardized test scores, presuming they even are applicable to what the teacher is teaching, to determine how well the teacher is doing?



There are a few studies (e.g., the Melinda and Bill Gates Foundation being the most often quoted that we have found) that say teachers should be evaluated on student growth. However, once this evaluation method was subject to wider scrutiny, professional studies are nearly unanimous in concluding that it is flawed and should not be used – witness the American Statistical Association study on the previous slide. There are no other professionals we have been able to find who are similarly evaluated. In fact, the test of time has shown what all manuals we have found regarding professional personnel evaluation say: Evaluate professional personnel via good observational techniques using trained evaluators with teaching experience. Determine criteria, get inputs from others, evaluate formally at least once a year, but evaluate continually – with feedback – at all times, etc.

This slide speaks for itself.



This slide speaks for itself.

Science & Math
CESE
Education

ADDENDUM

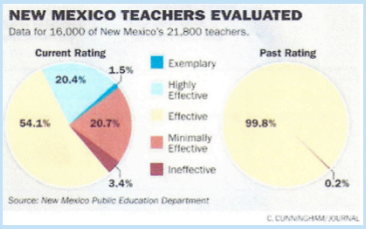
Additional Information on the "99% Error"

34A

Science & Math
CESE
Education

THE CONFUSION

The following graphic and excerpts are from an Albuquerque Journal story from May 16, 2014. The Journal was simply reporting based on the information presented to them.



WRONG

Education Secretary-designate Hanna Skandera said the new system was needed because the old evaluations were flawed. They didn't give insight into teacher performance, Skandera said, noting that over 99 percent of teachers were found effective under the old system.

Skandera has said a strength of the evaluations is that 50 percent of scores are based on how a teacher contributes to student achievement, measured by progress on standardized test scores.

Teacher unions and some local school officials have blasted the evaluations, saying they rely too heavily on student standardized test data, among other criticisms.

"Of course we don't believe the 76 percent figure is accurate," said Stephanie Ly, president of AFT New Mexico. "It's the same old story. The system is flawed."

Ly said the percentage of effective teachers in the state is higher.

SPECULATION

WRONG

Something is very wrong with the data and quotes.

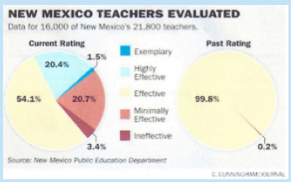
35A

The Three Tier Teacher Evaluation System was put in place by law during in 2003. It was used to provide 3 levels of teacher licensure, each requiring more and more formal demonstration of advancement in capabilities. Though this has been sometimes controversial, depending on your point of view, it effectively caused an overall increase in teacher salaries to occur and, theoretically, it promoted increased levels of performance (we have no data to know one way or the other if this is true). To achieve the next level (tier) of licensure, very specific criteria had to be met by the teacher seeking the advancement. The applications for advancement were first filtered through the teachers' principals and districts. After passing the specified requirements at the district level, the applications or "dossiers" were passed to the PED for final review which took place with outside, contracted reviewers.

One might relate the PED's function as making sure that all the requirements were checked off and properly completed. It was logically expected that the approval rate at the PED level or even at the district level after being reviewed thoroughly would be nearly 100%. And that is what the data demonstrated as shown in the right pie chart in the figure. But this had NOTHING TO DO with a teacher's annual evaluation rating. Many teachers leave the system within the first five years of beginning teaching. There is an "up-front" filter, so to speak, that rids the educational system of people not suited for a career in teaching. Is this filter perfect? No, it is not. There are also a number of instances (anecdotal) of teachers with inferior skills or an inappropriate temperament who still teach. But some of these teachers are also dismissed from the system – again, we have no specific records, only anecdote. Unfortunately, these data do not appear to be kept in other than a

Science & Math
CESE
Education

AND HERE IS WHAT IS WRONG



- The data are apples and oranges. The left pie chart represents the initial results (later modified) for teacher growth comparisons based on state tests. The right pie chart is NOT a summary of teachers' annual evaluation results using the previous year's method of observation only.
- The right pie chart represents the percentage of teachers approved for a change in licensure level – a formulaic requirement that never makes it to the PED until a teacher has met predetermined qualifications. One would expect a near 100% acceptance rate.
- Skandera's statements are simply wrong though the press and the Governor has repeated them in many different fora. Ly's statements have no known empirical basis.

36A

Science & Math
CESE
Education

HOW DO WE KNOW THE PIE CHARTS ARE APPLES AND ORANGES?

The following is the answer to an IPRA (Inspection of Public Records Request) request that was filed with the PED on May 27, 2014 after the Journal story in which Secretary Designate Skandera stated that 99% of teachers evaluated received an effective or above:

"Along those lines, there have been questions around the citation of 'today, 99.8% of our teachers meet competency.' This data point comes from a study in the early part of 2010 that examined the total number of Professional Development Dossiers submitted between 2005 and 2010. The study indicated that of the nearly 6,800 submissions, less than 15 did not meet competencies as verified by local superintendents.
This data point comes from a PED examination conducted in the early part of 2010 that examined the total number of Professional Development Dossiers submitted between 2005 and 2010."

A subsequent, formal query (IPRA request) to the PED asking what the real evaluation percentages were received no answer.

37A

Science & Math
CESE
Education

WHY IS THIS IMPORTANT?

- The use of the 99% number to justify a new teacher evaluation method is bogus. Teachers understand that and are upset about it.
- Still, the evaluation system that was in use before 2014 was not up to the best practice standards.
- Change may be required, and attention to evaluation by EXPERT observers is almost certainly in order.
- No one (on record) appears to know what the cumulative teacher evaluation results were prior to the 2013/2014 school year. If anyone should have the data, there are many, many people who would like to see it.

38A