

*Science & Math*  
**CESE**  
*Education*

---

**A Data-Based Approach for School Evaluation  
and Improvement in New Mexico**  
\*\*\*\*\*

**A Rational Methodology for Applying the  
A-B-C-D-F Schools Rating System**

**Coalition for Excellence in Science and Math  
Education**

Contact: Walter Murfin  
murf2345@aol.com

Presented to LESC  
Alamogordo, NM  
August 25, 2011

1

There are some complex issues to be covered. Time constraints make it impossible to cover every point at length. It is unlikely that everyone will completely absorb all this in the short time available. We are available for longer presentations or face-to-face discussions.

Our proposal is not specifically aimed at NCLB, which we believe is flawed beyond the possibility of repair. On the other hand, NCLB showed the general public that some NM schools are seriously lagging. Our proposal starts from the premise that current methods of evaluating schools are flawed and do not give a true picture of school performance. We will demonstrate a better way that is technically defensible and statistically correct.

## OUR GOAL: TO IMPROVE EDUCATION FOR ALL NEW MEXICO STUDENTS



- **The First Task: Bring struggling students and schools up to speed.**
- **Next: Raise the level for ALL students and schools**
- **This presentation focuses mainly on the first task, but we must keep our eyes on the longer range target as well.**

We can demonstrate a clear and technically correct method for accomplishing the first task – improving the performance of struggling students and schools. We believe the program has a good chance of success, and it is well worth trying. A pilot program could be started promptly, with some nominal training and technical assistance being required for PED staff. We must also be working on a method for raising the level of ALL schools.

## GRADING SCHOOL PERFORMANCE



THE “A-B-C-D-F SCHOOLS RATING ACT” (A-B-C-D-F), AT THE URGING OF SECRETARY-DESIGNATE SKANDERA, CALLS FOR THE GRADING OF SCHOOLS WITH AN A, B, C, D, OR F BASED ON PERFORMANCE INDICATORS.

**EXCELLENT!** THIS COULD GIVE THE PUBLIC, THE PUBLIC EDUCATION DEPARTMENT, LEGISLATORS, DISTRICT ADMINISTRATORS, AND SCHOOL STAFF A CLEAR PICTURE OF SCHOOL PERFORMANCE.

### **BUT – WE NEED TO DO IT RIGHT!**

- THE PERFORMANCE INDICATORS MUST BE COMBINED BY A RATIONAL AND TECHNICALLY DEFENSIBLE METHOD.
- THE COMBINED PERFORMANCE INDICATORS MUST BE USEABLE TO GET A CLEAR PICTURE OF TRUE SCHOOL MERIT, UNCONTAMINATED BY FACTORS THAT ARE OUTSIDE OF A SCHOOL’S CONTROL.
- THE GRADE SHOULD BE USED TO HELP STRUGGLING SCHOOLS.

3

Some groups might oppose any attempt to evaluate schools. Assessment methods have to be technically defensible, with every point nailed down. We have to be able to assure schools that they will be judged on true merit. We have to keep in mind that our goal is to improve the education of all NM students, especially those who are now lagging.

This method focuses on scores on the NMSBA. We recognize that test scores give only a limited picture of the state of education, but they are the measure we have NOW. We should be developing additional measures that will give a more complete picture. Those other measures will supplement (but not replace) test scores. We will offer some suggestions at the end of this presentation as examples.

## HOW WE CAN HELP



CESE has many years of experience in analysis of New Mexico and national school data. Our goal has been to find ways to help struggling schools to improve. Based on research, we will show the following:

- Some of the traditional measures will not give the intended results if improperly used. We suggest a rational method for combining scores and growth in several subjects as well as other measures.
- Data on school demographics (fractions of minorities, poverty, limited English proficiency, disabilities, etc.) and current scores and other outcome measures can be used to develop “best fit” predicted scores for each school.
- The actual score and the predicted score can be compared for each school.
- The difference between the actual score and the predicted score is called the “residual.” It is an accurate and fair measure of true school performance. This can be used for more meaningful letter grades and to help struggling schools to improve.

This is the heart of our proposal. Each point will be covered in more detail in later slides. Keep the term “residual” in mind. It is the technically correct term for the quantity, but might not convey the full meaning to the public. Perhaps you can help to find a more readily understandable term.

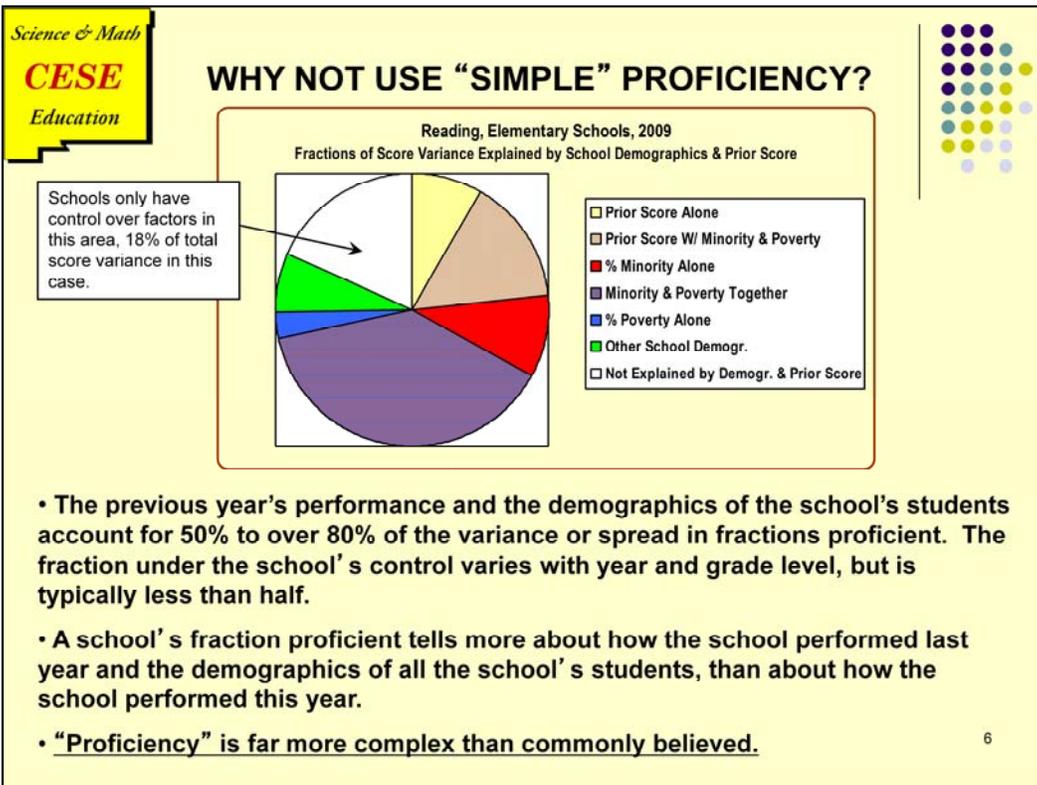
In our analyses we have usually included Hispanic, Native American, and African-American students as “minorities”, regardless of their actual fraction in the population.

## WE HAVE FOUND DIFFERENCES



- OUR RESEARCH HAS REPEATEDLY SHOWN THAT RESEARCH CONDUCTED NATIONALLY OR IN OTHER STATES MUST BE CAREFULLY VALIDATED FOR USE IN NEW MEXICO. WHAT WORKS ELSEWHERE, INCLUDING WELL ACCEPTED CONCLUSIONS, MAY NEED TO BE MODIFIED TO WORK HERE.
- OUR DEMOGRAPHICS, HISTORY, AND CULTURE ARE DIFFERENT.
- CONVERSELY, WE HAVE FOUND THAT ANALYSES THAT HAVE BEEN REPEATEDLY VALIDATED HERE DO NOT APPLY IN SOME OTHER STATES.
- OUR ANALYSIS – TO BE DESCRIBED IN SOME DETAIL IN THIS PRESENTATION – HAS POINTED TO MODIFICATIONS THAT COULD MAKE LETTER GRADING MORE VALID FOR NEW MEXICO SCHOOLS.

Example: it is true that economic status is a major predictor of scores nationally and in many other states, and it is commonly supposed that that this is universally true. We have found that for NM, minority status is invariably a somewhat more important predictor than economic status. However, by far the greatest effect is from poverty and ethnicity together. On the other hand, we found that the methods that always worked on NM school data failed when applied to Minnesota school data. It is absolutely necessary to look at the actual data in the actual schools for which we want to draw and apply conclusions. Commonly accepted beliefs – the things everyone “knows” -- are often incorrect.



In this example, only about one-sixth of the variance or spread in test results can be attributed to factors possibly under school control. Public schools are unable to choose the demographics of the student body. They can't go back to rewrite last year's performance. Even the white slice might not be completely under the school's control. Factors such as home environment, type of neighborhood, district support, etc. probably also have some influence.

It is important to note that these factors do not indicate the actual cause but instead indicate consistent correlation for all situations in New Mexico that we have studied. All of our analyses have shown that a school's fraction of combined poverty and minority students is a far more powerful predictor than the fractions of either poverty or minorities alone. Simplistic explanations like attributing reduced performance to poverty alone are often false and misleading.

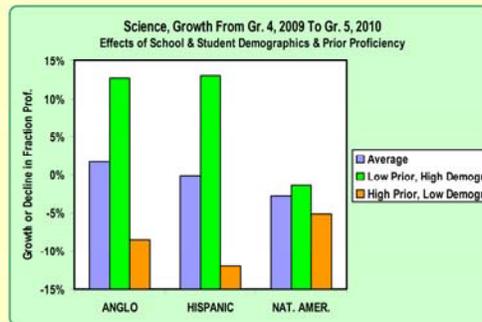
## GROWTH



- **A-B-C-D-F (Section 2.A) defines student growth. A student who meets any one of three criteria is counted as having made adequate growth.**
- **“School growth” can be defined as the fraction of a school’s students who meet at least one growth criterion.**
- **We did not have access to student-level data. A proxy for school growth was used: the increase in a school’s fraction proficient for a cohort of students, for example from 3<sup>rd</sup> grade in 2009 to 4<sup>th</sup> grade in 2010. These are mostly the same students, except for a few who transfer in or out or drop out.**
- **The NM PED will have access to student-level data. Their results would differ from ours in details, but we are confident that our general conclusions will be fully confirmed.**

We had to use a substitute for growth as defined in the legislation. We understand that the PED would use the defined measure. Our substitute measure demonstrates the method, although we are fully aware that the method if finally adopted will give results that might be different in the details. That does not matter. Our substitute measure demonstrates the effectiveness and practicality of CESE’s approach.

## WHY SIMPLE GROWTH IS A PROBLEM

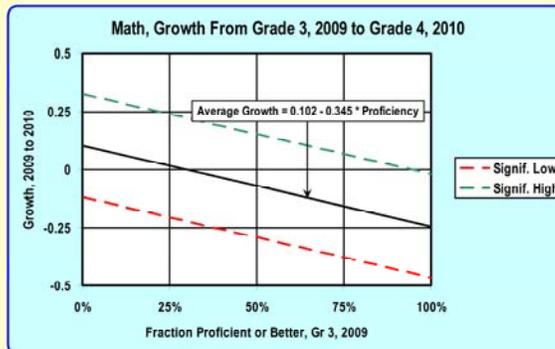


- Growth varies widely on factors that have little or nothing to do with student or school improvement.
- Growth is a complex function of the students' demographics, of the schools they attend, and of prior performance.
- Raw growth is typically unstable and irregular. High growth in one year is often followed by low growth or decline in the next year.
- Raw growth is more complex than commonly believed, and often tells little about real improvement in performance of either students or schools.

8

The blue bar shows average growth for each ethnic group when school demographics and prior scores are at their average values. The green bar shows growth when other factors have values associated with high growth. In both cases, Native American students have lower growth than Anglo or Hispanic students, on average. The orange bar shows average growth when other factors have values associated with low growth. In this case, Native American students have higher growth, on average. Growth, however it is defined, is complex and often counter-intuitive.

## Proficiency and Growth Together



90% of schools fall between the dashed lines.

- High initial proficiency is generally associated with low growth, and vice versa. High proficiency means less room for improvement. Low proficiency means more room for improvement.
- There is a great deal of scatter in the data, but the relation is significant.
- High proficiency is good, high growth is good, but the two do not necessarily go together.

9

Growth and proficiency tend to work against each other. If a school is at or near 100% proficiency, it can only stay the same or decrease. If a school has extremely low proficiency, it can only stay the same or improve. We can find schools with high proficiency and high growth by cherry picking the data, but the typical case is as shown here. Policies must be based on the averages, not on extreme cases.

## A-B-C-D-F (Section 4.B) MEASURES OF SCHOOL PERFORMANCE



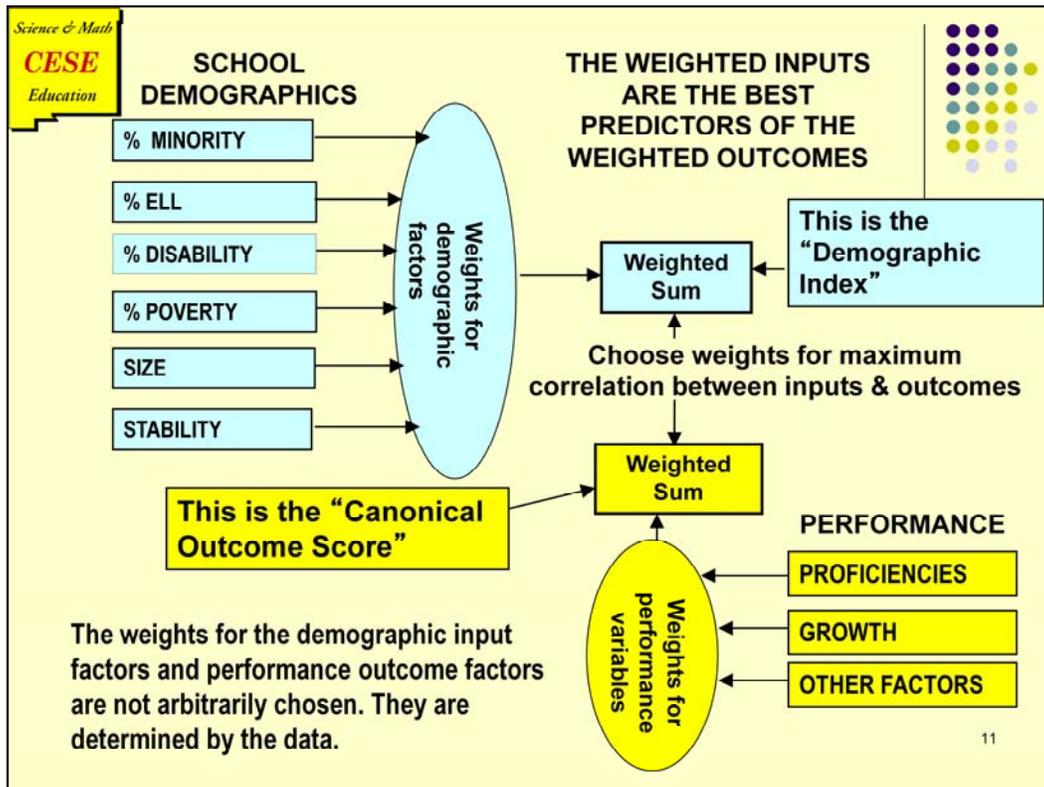
- Fractions proficient or better on the NMSBA
- Growth in reading and math
- Growth at the 25<sup>th</sup> percentile in reading and math
- Additional measures for high schools (e.g., graduation rate; AP and dual credit enrollment; and SAT and Act scores).

How can we put them all together to make a single defensible measure for evaluating school performance?

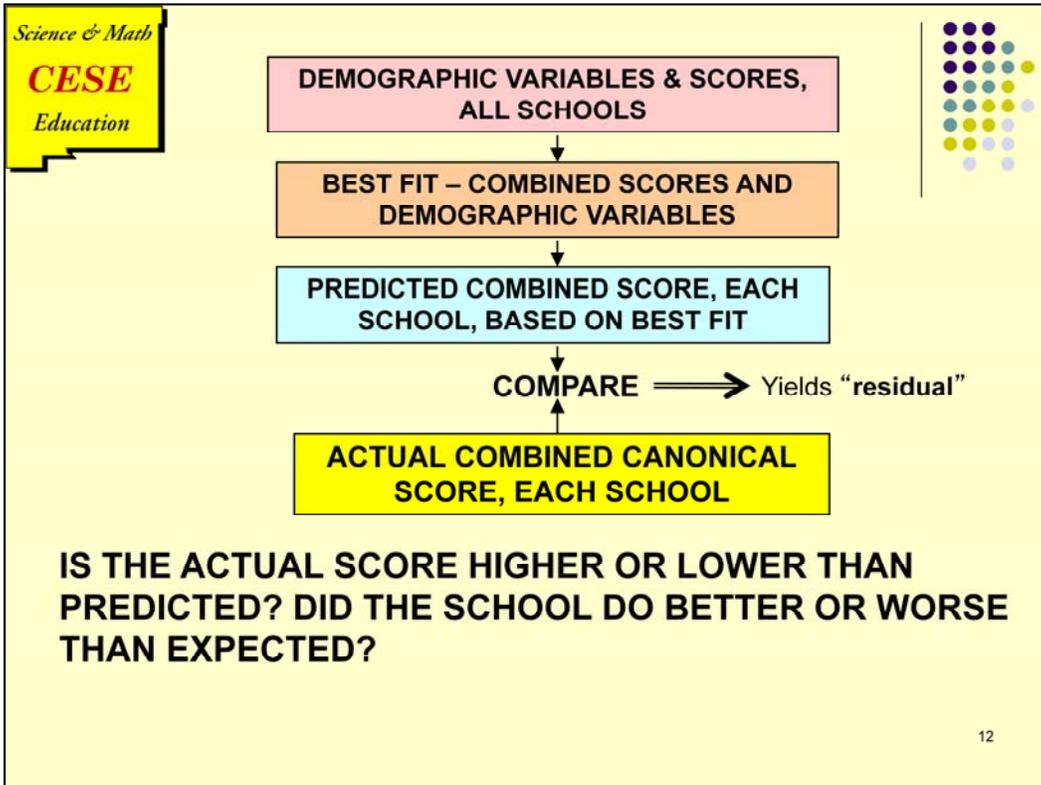
- An arbitrary or estimated combination will almost certainly not be the best. Simple averaging is easy, but is not technically defensible.
- As will be explained later, we want the combination to be most useful for determining a fair and defensible measure of true school performance.
- Combine the measures to have maximum correlation with school demographics. The combination is the “canonical outcome score.” No other combination will be as accurate in displaying the relation between school demographics and combined scores (see next four slides).

10

Requiring the combined score to have maximum correlation with school demographics might seem paradoxical. The reason will appear in later slides; it allows us to remove all the demographic effect from our final measure of performance. There is a well recognized method for this process.

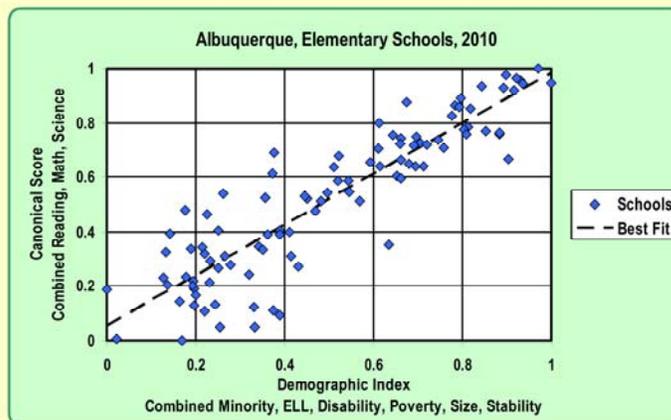


The weights for the demographic and performance factors do not in any way represent relative importance. They are only chosen to give maximum correlation between combined demographics and combined performance factors. Simplistic attempts, such as giving equal weight to each factor, will not result in true, demographic-independent measures, and cannot be technically defended. An additional advantage is that individual beliefs (often mistaken) will not influence the results.



This shows the sequence of calculations. The objective is to find a true measure of performance that is completely independent of school demographics. The “residual” – actual combined score minus predicted score for each school is what we are looking for.

## AN EXAMPLE OF "BEST FIT"

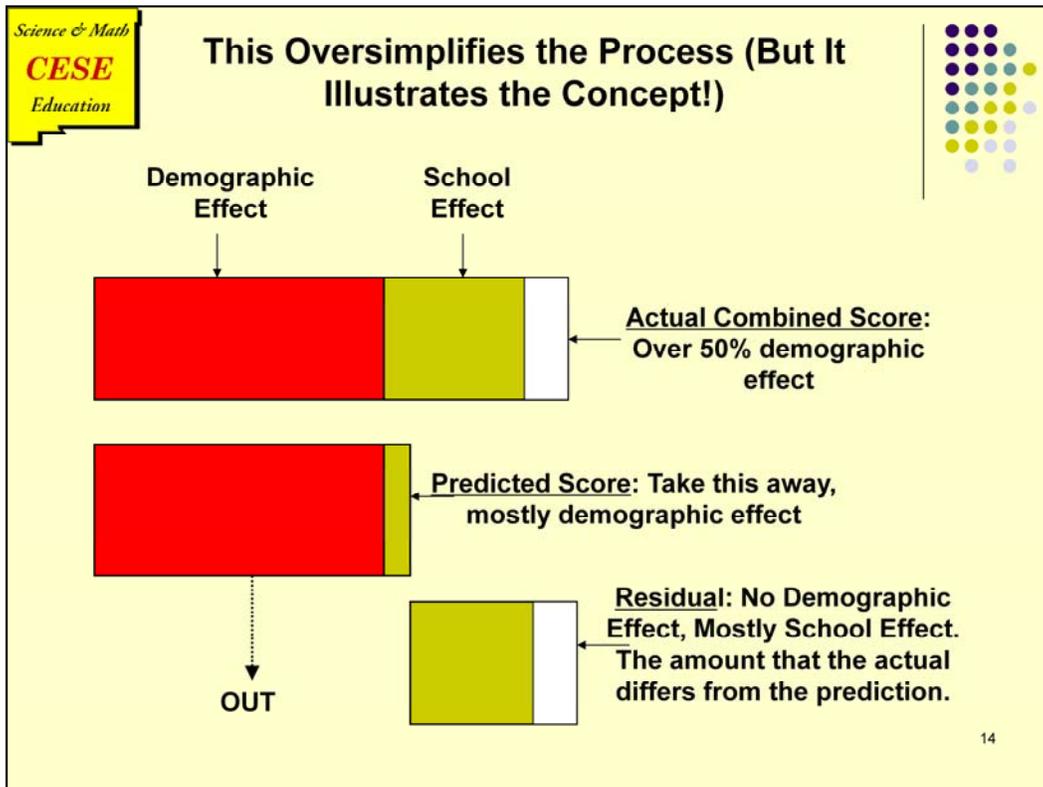


- Each symbol represents the performance of one school.
- Combined school demographics on the horizontal scale.
- Combined school fractions proficient in reading, math, science on the vertical scale.
- We find the line that is the best fit (in a mathematical sense) for all the points.

13

The idea is that, given a specific set of school demographic factors, we want to be able to find the most probable score, based on the demographic factors alone. The black line shows the best fit – the prediction. The scatter in the school data is real and is not a defect. Some schools do better than we predict from their demographics. Some schools do worse. That's exactly what we are looking for. We expect deviation from the best fit, and will show later how it can be used.

In the actual calculations for school evaluation we use all the individual demographic factors. The combined Demographic Index was only used in this example for simplicity of presentation



This shows why we need the combined measures to have maximum demographic effect. It means that all the demographic effect will have been wrung out when we calculate the residual. Most of the remainder will be under the school's control.

The predicted combined score has the highest possible estimate of demographic effect. When it is subtracted from the actual score, we are able to see whether the school has performed better or worse than could be expected from a prediction based on the school's specific demographics.

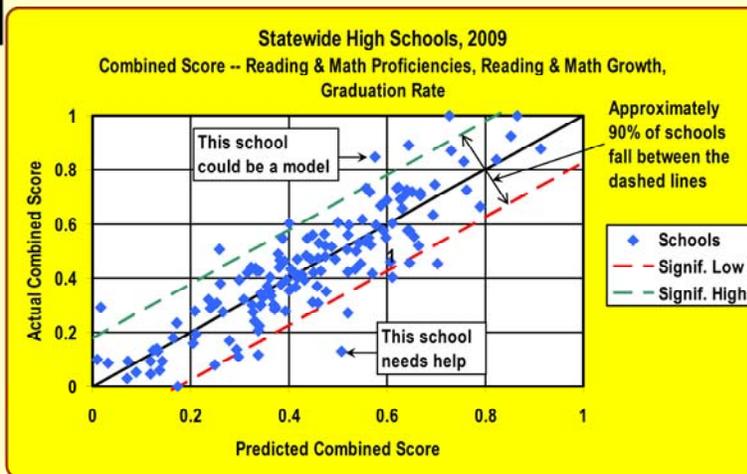
## A TRUE MEASURE OF EXCELLENCE, INDEPENDENT OF DEMOGRAPHICS



- THE PREDICTED SCORES HAVE MAXIMUM CORRELATION WITH DEMOGRAPHICS
- THE ACTUAL SCORES SHOW BOTH THE DEMOGRAPHIC EFFECTS AND THE EFFECTS OF TRUE SCHOOL MERIT.
- THE ACTUAL SCORE MINUS THE PREDICTED SCORE (THE “RESIDUAL”) FOR EACH SCHOOL IS INDEPENDENT OF DEMOGRAPHICS, AND IS AN ACCURATE MEASURE OF TRUE SCHOOL MERIT. DEMOGRAPHIC EFFECTS ARE COMPLETELY REMOVED.
- THE ACTUAL SCORES FOR SOME SCHOOLS WILL BE HIGHER THAN PREDICTED. THOSE SCHOOLS HAVE BEEN ABLE TO OVERCOME THE EFFECTS OF DEMOGRAPHICS AND SHOULD BE LOOKED AT AS MODELS.
- THE ACTUAL SCORES FOR SOME SCHOOLS WILL BE LOWER THAN PREDICTED. THOSE SCHOOLS HAVE NOT PERFORMED UP TO THEIR POTENTIAL AND NEED HELP.

The term “residual” will be used often in following slides. It’s important for understanding what is done and why we do it.

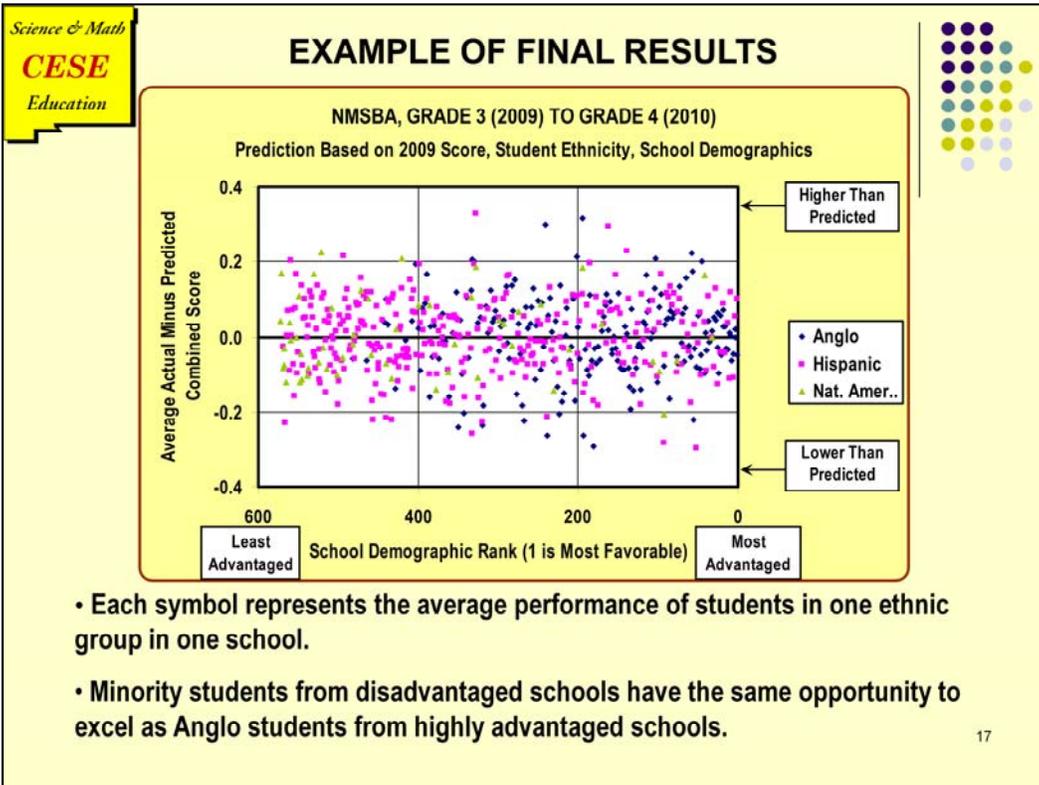
## EXAMPLE



Each blue symbol represents the performance of one school. The distance above or below the solid black line is an accurate measure of school merit, completely independent of demographics. Schools with unfavorable demographics have as much opportunity of achieving a high measure as schools with the most favorable demographics.

16

This graphical representation of actual and predicted scores of high schools in New Mexico demonstrates how the removal of demographics can identify schools that have model performance and schools that need help. Schools with similar demographic indices can be compared and model schools may be able to assist schools in need of help. Some schools are truly successful, but some others fall short. Schools that fall above the diagonal line do so because they have been able to overcome any disadvantages of unfavorable demographics. Schools that fall below the line do so because they have failed to overcome demographic effects or have not taken advantage of favorable demographics.



The residual – the actual minus predicted score is very nearly equally distributed over schools with every level of demographic advantage and students of any ethnic group. The working field is as nearly level as it can be made. Minority students in some schools, even with unfavorable demographics, do extremely well. Minority students in schools with disadvantaged demographics have the same opportunity to achieve a high residual as affluent Anglo students in a school with highly advantaged demographics.

## HOW CAN WE USE ALL THIS?



### (1) BETTER ASSIGNMENT OF LETTER GRADES

The “residual” (actual minus predicted combined score) is a fair and accurate measure of school performance, boiled down to a single, technically defensible number.

- Base the letter grade on each school’s residual
- Any that fall “significantly high” (the top 5%) probably deserve an “A”
- Any that fall “significantly low” (the lowest 5%) probably deserve an “F”

Letter grades assigned by this method will be fair, technically defensible, and completely independent of demographics. They are more likely to be accepted by schools and districts than scores based on raw proficiencies.

## HOW CAN WE USE ALL THIS?



### (2) HELP STRUGGLING SCHOOLS

For each failing school:

- Find a demographically similar successful school
  - Study both schools.
  - What is being done in the successful school that could be applied in the struggling school?
  - Help the struggling school to apply those methods
  - FOLLOW UP FOR A FEW YEARS! Is it really working? If not, why not?
- The method can be extended to individual ethnic and economic groups. There are schools in which poor and minority students outperform affluent Anglo students in most other schools in the state. If some schools can do it, why can't every school?
- At the same time that lower performing schools are improving, we develop programs to help all students in every school to improve.

This is the heart of our proposal. Teachers in schools that fall far below expectation would no doubt use strategies to improve performance if they knew what to do. This gives them a chance to learn and to utilize the techniques being used in schools with similar demographics that perform far better than would be predicted. Helping the struggling schools will surely be more effective than applying punitive sanctions.

We have demonstrated the utility of the method. The PED is free to use our work, with citation, in applying for a waiver from the requirements of NCLB. The method has been developed to the point that the PED could make a very strong case for a waiver.



**EXAMPLE:**  
**HELPING STRUGGLING SCHOOLS**

SCHOOL	ACTUAL SCORE	PREDICTED SCORE	ACTUAL MINUS PREDICTED	DEMOGR. INDEX
SCHOOL "A"	0.326	0.563	-0.237	0.597
SCHOOL "B"	0.791	0.570	0.221	0.606
SCHOOL "C"	0.297	0.407	-0.110	0.351
SCHOOL "D"	0.636	0.408	0.228	0.353
SCHOOL "E"	0.534	0.412	0.122	0.359
SCHOOL "F"	0.313	0.433	-0.120	0.392
SCHOOL "G"	0.539	0.433	0.106	0.392

- School "B" could be a suitable model for School "A".
- Either School "D" or School "E" could be suitable models for School "C".
- School "G" could be a suitable model for School "F".
- There are many other examples.

Just a few of the many possible matchups.

## SUMMARY



- We do support letter grades for schools in the interest of clarity, but some conditions need to be considered.
- Raw proficiency and raw annual growth are not suitable indicators of school performance. Both are far more complex than commonly believed, may not give the intended information, and can actually give a false picture.
- There are technically defensible methods for combining performance measures. We suggest a method in which the combination has the maximum correlation with school demographics.
- The combined measures and school demographics can be used to calculate a best fit prediction for the outcome measure.
- The actual outcome measure minus the predicted measure (“residual”) is completely independent of school demographics and is an accurate and fair measure of true school merit.
- Letter grades based on the residuals will also be accurate and fair.
- Residuals can be fruitfully used in a program to help struggling schools, and in the long run to help all students.

We have shown that raw scores strongly reflect factors over which the schools have no control. There is a well-tested method for eliminating school demographic effects. We have demonstrated that a measure of merit independent of school demographics – the “residual” – can be found for each school. We have shown how this measure can be used as a basis for grading schools fairly and effectively. More importantly, the measure can be used to help struggling schools to improve.