Steve Cordogan
Director of Research & Evaluation
Township High School District 214

To: Website Visitors
Subject: What You Need to Know about Using the EXPLORE, PLAN and ACT Tests for Instructional Improvement
Date: 3-14-11

1. I'm Dr. Steve Cordogan, Director of Research and Evaluation in Township High School District 214, a northwest suburban school district with six high schools (Buffalo Grove, Elk Grove, John Hersey, Prospect, Rolling Meadows, and Wheeling) and around 12,300 students.

The following presentation is a comprehensive overview of District 214s' experience with the EXPLORE, PLAN, and ACT tests. This ACT testing series is called EPAS, which stands for the Educational Planning and Assessment System.

2. Here are the topics we will cover in this presentation, which will be divided into six sections.

Many of us in public education have spent years trying to make sense out of high stakes testing data. Those of us in Illinois, particularly in high schools, have seen test selection, administration, and scoring changes that were generated by political considerations more than by educational ones. And what we knew about the tests generally came from the same people who were selling them and telling us how great they were.

The ACT has been used for decades in District 214, and we began using the other EPAS tests in the mid 1990's. Our EPAS focus became much stronger in 2004 when our research showed that the PSAE and AYP measures were very limited in their usefulness for instructional improvement. For example, AYP omits English, and it focuses on pass/fail measures rather than on scale scores which give much more useful information. For the past six years, EPAS scores and growth have been our primary district-wide instructional improvement measures at the departmental, school, and district level.

We have learned a lot about EPAS through the years of use and the conducting of dozens of research studies based upon the data. A fair amount of what we have found differs from what ACT has claimed, but regardless of the differences, the tests probably are the best tools currently available at the high school level for measuring large groups of students and using the data for instructional improvement.

3. So let's take a closer look at ACT EPAS.

4. Here's how ACT describes the tests

5. This slide describes how EPAS tests have become increasingly important in Illinois over the past decade. While the Common Core Standards will bring about a new set of assessments in a few years, ACT is very likely to be a major player in those assessments. And colleges and scholarships will be focused on the ACT for the foreseeable future. Next to AP, the ACT is the fastest growing of the major high-stakes tests, with almost 50% more students taking the test than ten years ago. Nearly 1.6 million students will take the ACT in 2011.

6. Here is a summary of the EPAS tests. We modify the EPAS structure so that we have an annual test from 8th through 11th grade. We give 8th graders the EXPLORE, 9th graders the PLAN, 10th graders a retired ACT, and 11th graders the ACT as part of Illinois' Prairie State test in late April. Much of our instructional improvement focus is on the EXPLORE and PSAE ACT, since they are higher stakes for our students and have less variability, but all of the tests have been useful in tracking our student's progress.

7. All EPAS tests cover four subject areas: English, Math, Reading, and Science. For the EXPLORE, shown here, each test is 30 minutes long. ACT identifies content areas and provides scores from each test, as well as English test subscores for rhetoric and usage/mechanics. ACT also provides a composite score, which is the average of the test scores (always rounded upward).

8. In addition to the test, the EXPLORE provides a survey to assist students with educational and career planning.

9. The PLAN is similar, but adds math subtest scores, and…

10. provides a more extensive educational and career planning survey. Please note that if you check the box to accept the Educational Opportunities Service, it will provide you with possibly useful information, but it also allows ACT to sell your information to colleges and other vendors, resulting in a lot of junk mail and e-mail. Additionally, I had one student concerned with identity theft complain because their birthdate appeared on the envelope of a college mailing. A checkmark in the seemingly innocent-sounding EOS item may give your students a lot more than they had bargained for.

11. The ACT described in this slide has more subtests as well as an optional writing test, and is much longer – at least 46% longer without the writing test, and 71% longer with it.

12. Here is a list of our primary reasons for using EPAS. The first bullet was important - students and teachers were not always inclined to take some of the state's earlier high-stakes tests seriously. But since our students' futures can depend on ACT scores, they're a little hard to ignore. As for the third bullet, the college readiness standards are easily understood for diagnosing student strengths and deficiencies, as well as for purposes of designing curriculum. We found them easier for teachers to work with than the state standards.

As for the last bullet, please realize that no test scoring is absolutely stable across different versions of the test. While the ACT is more stable than most, changes will happen. Additionally, the changes may affect some students and schools more than others. For example, one version of a subject area test may be harder for students scoring at the higher ranges of the test but not for students scoring at the lower ranges. Last year, statewide scores dropped 1/5 of a point in English and science, and I am quite sure that it was not due to the sudden onset of a brain freeze affecting an entire class of students in only two subject areas. The test versions administered at PSAE testing, which account for most of the state's final ACT scores in the ACT Class of 2010 report, were harder than in the previous versions.

13. This slide is included to illustrate that ACT tests are a valid measure of student achievement for groups of students. This graph for the 71 Chicago area suburban districts with high schools illustrates the relationship between a single demographic characteristic on the horizontal axis, and student performance as measured by the ACT composite score. Research has conclusively found that parent education level is the best predictor, beyond income, mobility level, or any other factor, although they are all very interrelated (e.g., the

higher the income level, the lower the mobility rate). While I don't have specific parent education level, I do have the next best thing, the education level data for the districts in Illinois, as shown here. That demographic characteristic predicted 80.9% of the variance in ACT composite scores, an enormously large amount. Additionally, when we add student racioethnicity to the prediction equation, we can predict between 94% and 97% of the between district variance in overall ACT performance in the suburbs.

Although it requires an inferential leap or two, given the strength of the relationship between parent education and student achievement, and given the strength of the relationship between bachelor's degrees in the community and ACT performance, the use of the ACT composite as a measure of a school's or district's student performance basically is validated. It is not some sort of disconnected or irrelevant test.

Nevertheless, significant differences remain between districts beyond what is predestined by demographics. It's the difference that remains after demographics have been considered that identifies which schools and districts really do a better job.

14. Next, we will take a closer look at the ACT College Readiness Standards. If you would like to examine or download them, the URL is provided on this slide.

15. Here is the first of two slides displaying ACT's description of the CRS. The College Readiness Standards have been at the heart of our school improvement efforts. They have been used as a skill spine for curricular work in many of our schools. Please note that this is not to say that they comprise a curriculum in themselves, but they do define essential broad elements of a curriculum. Where they were used the most, the John Hersey English department, they have been integral to a remarkable 4.3 point increase in ACT English scores within seven years (from 21.9 to 26.2). And this was done through a creative, engaging, and constructivist curriculum, not some drill and kill test prep approach.

16. ACT's description continues, explaining how the standards connect score levels with specific skills that students with such scores are able to do, with ideas on how to help them improve.

17. The next three slides provide examples of the college readiness standards. On the left, we see the score level bands for the EPAS <u>math</u> tests – the first is for very low scoring students with scores between a 1 and a 12. In each score range, ACT provides a set of general standards and a set of skills areas listed across the top of the chart, such as "Basic Operations and Applications" and "Probability, Statistics, and Data Analysis." Within each of these score bands and skills areas, ideas for helping student progress are provided.

18. Slide 18 shows more of the skill areas, such as "Expressions, Equations, and Inequalities."

19. We see that the chart continues with ideas for progress for each of these skills areas for students scoring in the 16-19 range. Additional pages of the chart deal with the rest of the skills areas and score ranges up to the 36 point maximum for the test.

20. I will not discuss the contents of the next four slides in detail, but I want everyone to be aware of what the student score report looks like.

21. The score report includes percentages to explain where a student ranks within their school the state, and the nation.

22. an analysis of career possibilities, and

23. and item analyses for each test, telling students which ones they got right or wrong, with recommendations for improving skills in the various content areas.

24. And here's the ACT version.

25. Since we're on the subject of ACT Scores, I know that everyone is interested in how students can do their best on the test. Retaking the test leads to higher scores more often than not, though some of this improvement may be due to the fact that the retake comes later and the student has learned more since the first testing. But practice probably helps. Additionally, I would recommend a second try since, as we discussed earlier, one test version may be slightly different and your child may do better on one version than another. They need to try hard on both, since you cannot count on doing better on the second one.

26. We have seen that ACT provides a lot of data. Unfortunately, not all of it is good data. A huge problem with ACT data is its College Readiness Benchmarks.

27. The benchmarks, based upon a 2005 ACT study, were an attempt to establish minimum ACT score levels at which a student would be considered college-ready. According to ACT, "The benchmarks are scores on the ACT subject-area tests that represent the level of achievement required for students to have a 50% chance of obtaining a B or higher or about a 75% chance of obtaining a C or higher in corresponding credit-bearing first-year college courses. These college courses include English composition, college algebra, introductory social science courses, and biology."

This slide then shows the benchmarks for the EXPLORE, PLAN, and ACT. For example, an 18 on the ACT English test is considered on-track to be college ready. A 13 on the EXPLORE English is expected to lead to an 18 on the ACT.

28. Here's the ACT score report again. The students' scores are contrasted with the benchmarks on the score report, highlighting which benchmarks the student did or didn't make. In this case, the chart in the red circle shows that the student met the benchmark in English and reading, but did not meet the benchmark in math and science.

29. All of this warrants an ACT-type multiple choice test question - you have ten seconds to answer… Now I've gotten money from the tooth fairy, candy and eggs from the Easter bunny, and presents from Santa Claus, but I've found nothing in my experience to support the college readiness benchmarks.

30. Unfortunately, the benchmarks are not a laughing matter. While a logical idea, they ignore the fact that the tests alone can be weak predictors of college success, particularly the ACT reading and science tests. This was complicated by ACT horribly flawed research methodology and interpretation of the findings. In my university work with graduate students, we NEVER would have allowed students to conduct such bad research.

31. Do you see a problem with the scores? When I first saw them, the English jumped out at me as a score too low to gain entrance to college-level courses at most colleges in Illinois, and the science was impossibly high relative to the mathematics score.

ACT combines the benchmarks and declares that anyone who does not meet all four is not college-ready. Since so few students reach the artificially high science benchmark, ACT can claim that only 23% of Illinois students are ready for college. This simply is untrue and a gross abuse of their role in Illinois testing. Unfortunately, <u>it appears that ACT is trying to create the perception of a crisis, which creates a better market for selling its products and services</u>.

32. ACT maintains that a student is not college-ready if they do not meet all four of the benchmarks. Yet they came out with a subsequent study that showed that 65% of the students who met NONE of the benchmarks persisted to their second year of college. If meeting the benchmarks was critical, how could so many do so badly and still succeed?

The truth is that both studies are not credible. In the case of the second study, most students cannot do that badly on their tests and remain enrolled that long.

33. Here are the main reasons for the conflicting findings. I don't want to spend too much more time on this, but you need to know what to say to students, parents, and the public when they see these grossly inaccurate science and English figures.

Much of the problem is with sampling issues. ACT claimed that the study was "Based on a nationally representative sample of 98 institutions and more than 90,000 students." Well, 98 schools out of the 4,200 colleges and universities across the country would require careful random sampling to be representative. But ACT used a non-random convenience sample, limited to those using a service offered by ACT. And samples ranged from 76,122 students for English down to only 14,136 for science, which would be only 1% of the students taking the ACT science test that year. Most incredibly, the science benchmark was based only on science majors, who would be significantly higher performing than the students sampled in the other subject areas. So the sampling could be best characterized as junk science at best.

In addition to the sampling issues is the weakness of some of the ACT tests in predicting college classroom performance. A study of 2,432 Northern Illinois University students showed that the science ACT score predicted only 6.9% of the variance in first year general biology grades; i.e., 93.1% of the variance was NOT explained by the test score. In fact, the English test explained almost as much as did the science test. The science test is such a seriously weak predictor of student performance that using it to create a benchmark borders on the unthinkable.

For those of you who feel a need to have some sort of college-ready guideline for scores, I think that a composite score of 21 is as close as we can get. As we will see later, the composite score for the EPAS tests is much better overall predictor of future classroom student performance. Obviously, this is simply a generic standard. The standards at the college at which a student plans to attend will set the benchmark.

Incidentally, when I describe reading as not predicting much, you may want to click back to the last slide, # 31. What's missing in this study? Reading apparently was omitted because its inclusion would have reduced, or made no difference in, the prediction level. We will see more on the weakness of the reading test later.

34. Here's the College Board's take. They simply sat that you can't use a test alone to determine college readiness. Educators need to get out the word that the ACT College readiness Benchmarks are nothing more than an ethically indefensible attempt to misuse bad research for marketing purposes!

35. The next few slides deal with why we don't use AYP, the ISAT, and the PSAE for school improvement.

36. The first is that testing and scoring changes by the state or its suppliers over the past six years make it very difficult to compare scores across years. It is not possible to identify whether a school's score changes were due to instructional changes or to changes in the state tests.

37. Also, the scaling is substantially different between the ISAT and the PSAE. The pass rate for the ISAT is 56% to 59% higher than for the PSAE overall, so meeting standards on one is not remotely the same as meeting standards on the other. Students who were labeled as

meeting standards in 8th grade suddenly are labeled substandard in high school. And with some K-8 school score levels in the 96% to 98% range, improvement is almost impossible to measure. The test lead schools to assume that they already are near perfect, which is not what the subsequent EXPLORE testing tells us. This false sense of high performance undermines the drive to improve K-8 instruction.

This does not mean that the ISAT is a bad test. Students who do well on the ISAT also do well on the ACT and other standardized tests. But the grading scale on the ISAT is much less rigorous than that on the PSAE.

38. These are three reasons that much of our data, particularly current state NCLB data, don't meet our needs very well.

39. This revised version of the scatterplot graph seen earlier incorporates both bachelor's degree levels and at-risk racioethnicity levels to create an expected ACT score. This illustrates the fact that demographics predict an incredibly high 96.5% percentage of suburban school district ACT performance. So only 3.5% of the variance in suburban ACT composite scores is not predicted by these two demographic factors. Hence, as we all know, **overall school performance largely is determined by initial student characteristics**. If we judge a school only by achievement scores, we're largely judging them by who enrolls there, and not on how well the school helps students to learn. **This punishes schools who receive low performing students regardless of their efforts, and rewards schools who receive high performing students, even if they do little to increase their learning**.

40. Much of the year-to-year change in NCLB meets/exceeds levels clearly can be connected to changes in incoming student performance; changes **that are highly statistically significant**. So not only are there differences between schools in initial performance, there also are real differences between years at the same school. Some classes simply do better than others. I'm sure you've seen it. Again, such differences are **NOT** connected to what the school did for the student, only what the students brought to the school.

Some day very soon we need to examine what we as schools do to help create these "good class, bad class" differences. There shouldn't be so much difference between classes. But for now, there is.

41. Finally, let's look first at the issue of random variance, or chance. Like the flip of a coin where you might get heads or you might get tails, meets exceeds levels may vary between years for no identifiable reason. These are differences that are not statistically significant. And by taking a test score and making it pass/fail, those meaningless year-to-year changes seem much bigger. For example, you take widely differing PSAE reading scores of both 155 and 200 and call them both passing. That's like saying Bill Gates and I are both in the top 50% of American wage earners – we lose a lot of information when we make such simplistic groupings.

42. When we look at the statistics behind using pass/fail data, we find that for a target subgroup of 45, a change in meets/exceeds level as large as 20% or even more can be due simply to random chance. A 20% increase, doesn't that seem like something huge to celebrate - and yet statistically it can mean nothing. And the following year, when there's a big drop, we wonder what went wrong, even though nothing may have changed in terms of student ability. Incidentally, AYP meets level can actually go up while average scores go down (and vice versa), due to small changes in scores near the cutoff level between meeting and does not meet.

43. Meets/exceeds data are not worthless. Changes sustained across multiple years are likely to be meaningful. But the one-year fluctuations on which on which we currently are being judged by the state and feds often can be meaningless.

44. We can't, and shouldn't ignore final achievement levels - and the public and government won't let us. However, we need to focus on what the schools are value-adding to the student. We need to look at growth.

45. So what happens when we contrast AYP data with ACT-type data or other performance data? We get very different pictures of a district or school's performance. Our district's AYP performance looks pretty mediocre across the past years. Our math has made only minor growth, and our reading growth has largely reversed direction and ended up not far from where we started. Incidentally, we are using the graduating class, not the year of the test, to label the data.

46. When we look at our final ACT performance data, we see that all subject areas are demonstrating gradual, significant, and overall substantial growth. When we look at the Classes of 2004 through 2011 data that match the previous graph (the striped patterns denote the pre-AYP classes), we see that reading increased steadily throughout those years, in contrast to the AYP findings. So AYP was not giving an accurate picture of our progress in those subject areas, nor does it describe important growth in other areas critical to our 90% college-bound student body.

47. When we look at the composite scores only, over the course of the past 25 graduating classes, and contrast the increases in scores with the state and national trends, as well as with incoming EXPLORE scores, we see how dramatic the improvement has been, a projected 1.6 point growth across a very large district that, initially, already was performing at a respectable level. The focus on the College Readiness Standards and EPAS performance was instrumental to the instructional changes that led to this growth.

Please also notice that, despite all of the recent publicity about concerns over student performance across the state, ACT scores have increased by .6 in Illinois since the beginning of the state's universal testing, three times the improvement level for the nation. Additionally, they had been improving substantially, by .7 points, over the twelve-year period preceding universal testing, a rate 75% higher than the national average.

48. So which is the real measure of the district, EPAS or the PSAE? Well, to cross-validate the ACT findings of significant improvement, we see that the number of passing AP tests per student has increased 149% in the past nine years, and

49. The number of students receiving D's and F's has declined by over 24% in the past three years. So the cross-validating data show that EPAS gives us an accurate picture of our district's improvement, which the PSAE AYP did not. This does not mean that the ISAT or PSAE are bad tests, particularly since the PSAE is partly the ACT. We simply see that with the skewing of ISAT scores to an "almost everybody passes" scaling, the many annual changes in the ISAT and PSAE, and the reducing of the scale scores to simplistic pass/fail measures, the tests cannot measure progress across years.

50. Now we'll look at the relationships between EPAS tests and various academic performance measures.

51. EPAS tests do correlate strongly with other tests. This slide of some 2005 research shows the relationship between student EXPLORE scores from one of our large sender districts to the Iowa Test of Basic Skills and the previous version of the ISAT. The figures represent the squares of the correlations between the two tests listed for each bar. So, for example, the

EXPLORE reading score explains 61.5% of the variance in ITBS scores, which is a very high level. Basically, as we would expect, students who do well on one test do well on the other.

52. Similarly, this slide shows that among our sender schools, those with high meets/exceeds levels also have higher EXPLORE scores. Again, I want to emphasize that the ISAT levels may be inconsistent between years, but at any given year, ISAT student performance is consistent with that of EXPLORE scores.

53. One of the most important uses of EXPLORE scores is for placement of sender school students in District 214 classes. While 15% to 30 % explanation levels between test scores and future grades may not seem like a lot, this is much stronger than the levels achieved by the ACT in predicting college performance.

But the interesting discovery of three years of research on almost 9000 students is that the composite score EXPLORE composite test scores remain the best predictor of 9th grade student performance for every department's courses. They are consistently better than the subject area EPAS scores. The composite score explains 16.1% more of the variance in 9th grade English class grades than does the English test, 24.8% more than the math test predicted math grades, 36.9% more than the reading test predicted social science grades, and 22.8% more than the science reasoning test predicted science grades.

The strength of the composite score relative to the subject area tests is in part due to the simple psychometric consideration that it can be considered the score of a longer test; the score is derived from testing that is approximately four times the length of the subject area test. A longer test (within reason) is more reliable and valid than a shorter test. Also, the content area score is incorporated in the composite score. Finally, initial classroom performance may be dependent on more than simple content area testing ability (e.g., an overall ability to "do school," to show up on time with homework completed, etc., which may be better measured by an overall score).

It may seem inappropriate to use a composite score when a subject specific one is available, but the data clearly demonstrate that the composite provides a more complete picture of student potential for 9th grade in all subject areas. And although the displayed data are only for District 214, two other suburban districts replicated the study and also found that the EXPLORE composite score was the strongest predictor for 9th grade classroom performance.

In later grade levels, the PLAN and ACT composite scores also are the best overall predictors of District 214 classroom performance, but the math and English tests emerge as strong predictors and are useful for assessing the learning taking place in those classes. The science test, while it represents valid skills and knowledge that should be part of the curriculum, does not predict science grades as well as the composite or even the math score. Similarly, reading predicts the least amount of performance overall of any of the tests for any class, just as ACT found for predicting college-level persistence.

Please note that the lack of relationship does not necessarily indicate a problem with the tests, since a badly aligned curriculum also can be an issue requiring attention. Nevertheless, a test should not be assumed to represent what is important to teach, and when real-world evidence shows a disconnect between the classroom and the test, we need to question the test as well as the curriculum.

Given the limited relationship between EPAS tests and some subject areas, only English and math test growth can be used to examine teacher performance in English and math on an individual basis. Even those must take a multi-year perspective and can only be a small part of any evaluation. EPAS cannot be used in a meaningful way for other subject areas, since there is no way to isolate the effects of one teacher's impact with tests that could be significantly and substantially impacted by learning in other courses. These issues must be considered when trying to tie state assessments to teacher evaluation, which is likely to occur in the next few years.

Incidentally, we focus only on non-ELL students, since these are the students for whom the test is a valid measure and growth. We also focus only on students that have both an EXPLORE and PSAE ACT, since the district should not be held accountable for students who may have attended only a short time.

54. One of the most valuable features of EPAS testing is that can provide measures of annual student growth. Since students change in incoming performance levels from incoming class to incoming class, school and curriculum evaluation should focus on growth between EPAS tests. We place the most focus on overall growth between the EXPLORE score and the PSAE ACT score. This time period encompasses most of a student's high school career. And since the EXPLORE is used for 9th grade placement, and the PSAE ACT is the final ACT for the majority of our students, both tests are very high stakes and are taken very seriously by our students, providing more consistent data than our PLAN and retired ACT testing in 9th and 10th grade. Additionally, all students take the PSAE ACT at the same time, and a PSAE focus provides ACT feedback a year before final ACT scores are generated, so the data are more consistent and more timely than final ACT scores.

The graphs shown provide a simple depiction of such growth. The dark blue bars illustrate the entering EXPLORE score averages for the past nine years of testing for District 214 schools overall. The light blue bars represent the average growth that those students made from their EXPLORE scores to their PSAE ACT's. Then the yellow strip on top of each bar shows the ACT scores. So for example, for the Class of 2011 (the last PSAE test), students entered with a 17.1 on the EXPLORE English test and grew 6.5 points to the ACT, ending up with a 23.6 composite score average.

55. When we focus on the growth only graph, which simply regraphed the light blue part of the bars from the previous graph, the growth levels becomes clear. We end up with a relatively pure measure of student growth. We see an overall growth in student performance of 19.4% (from less than 5.4 points of growth to 6.5 points). Again, this change is very substantial, hardly the picture we would get from the AYP data.

56. However, in District 214 and in a couple of other districts that have shared with us, this growth does not occur evenly across the years or EPAS tests. The EXPLORE to PLAN growth is 2.3 points, but with the almost 18 month period between the two, this equals an annual growth of less than 1.6 points. And although the tests are vertically aligned in theory, the scale differences (EXPLORE has a score ceiling of 25, the PLAN 32), we might have expected even more growth. Even more striking, the PLAN to retired ACT growth was only 1.1, despite another increase in the post-test's ceiling. But the retired ACT to ACT ceiling was a much larger 2.9.

The low 10th grade growth may be due to the proverbial "sophomore slump." But more likely, it is due to the greater motivation on the higher stakes-PSAE, and the fact that the 11th grade covers much more material relevant to scoring well on the ACT.

57. Furthermore, the growth does not occur evenly at different EXPLORE score levels. ACT tells us that the average student composite score grows around 4.5 points from EXPLORE to ACT nationally. But the average can be very misleading in terms of expected growth. This graph show the amount of growth from EXPLORE to PSAE ACT at each level of the EXPLORE composite score, Although the average growth at District 214 was 6.3 points for the last five graduating classes, this does not mean that all students have grown at that level. As we see from the graph, students with an entering EXPLORE score of 12 only grow an average of less than 4 points, but students with score of 21 points grow an average of over 8 points, more than twice as much. Basically, EXPLORE composite scores explain a huge 75% of the variance in PSAE ACT scores three-and-a-half years later. If you multiply a child's EXPLORE composite score by 1.38, you get a very good estimate of their PSAE ACT score in District 214. We don't necessarily want to use this when we talk to students though, since it is likely to reinforce low expectations. While a small amount of the difference in growth at each EXPLORE level may be due to the scaling of the two tests, most of it is due to the fact that at-risk students tend to grow less each year than higher performing students.

58. Why is this? This only is a very crude depiction of student growth, since in the real world, hypothetical grade equivalency growth is not linear. But it is meant to depict that when students come in with an achievement deficiency, the gap continues to grow with time unless a specific and successful intervention is made. In this example, if at-risk students grow at a hypothetical average rate of ¾ of a grade equivalent per year and mainstream students grow at one grade equivalent each year, the gap between each group will increase by a full grade every four years. The 8th grade gap of 2.25 grade levels grow by over 44% to 3.25 by the end of high school. Under NCLB, we are supposed to close that gap to 0. Realistically, our first goal should be to make sure that it stays at the 2.25 level with which they started high school.

Data from our most recent classes shows that we have only begun to shrink such gaps. We have made great growth with lower performing students, through timely interventions and a focus on moving students up from skills-level to mainstream classes. But these have been matched with higher-level student enrichment, such as moving them up to honors and AP classes, and offering more of the latter. So because all students have grown, the reduction of the gap has been very small.

The significance of growth being dependent on incoming performance is huge. <u>Using growth levels does not erase the effects of prior performance level, which of course is very dependent on student demographics.</u> It shows that we must evaluate schools relative to their demographically comparable schools, even when using growth levels. Using a growth model alone will not equalize between schools with widely differing demographics. Furthermore, using growth data for teacher evaluations also has to consider demographics. Without considering demographics and initial performance, using growth models is not much better than using final scores.

59. An additional consideration of extreme importance is that **no test scores are ever as strong a predictor of student classroom performance as from prior classroom**

performance grades! I can't emphasize the point strongly enough. Grades from earlier years have proven to tell far more about classroom performance in later years than test scores or teacher recommendations.

ACT scores do relate well with grade, with ACT composite scores predicting a fairly large 51.5% of the variance in grades in our District 214 schools. However, there are a fair number of students who are performing well in school who do not have college-level ACT scores. With a post-secondary attendance rate of over 90%, we can assume that our mean weighted GPA of 3.93 (weighed for honors and AP courses) shows that our grading standards are not too inappropriately inflated. But there are a significant number of our students who do well at our schools and yet have ACT scores that would significantly limit their access to college. While 36.5% of our students have ACT composite scores of 20 or lower, 24.1% of these have above average GPA's. So we see that a large number of students do perform well in our high school classes and probably would perform similarly well in college classes, yet have rather low ACT scores. Please also note that the GPA for lower EPAS scoring students may have been reduced through teacher perceptions of what such low scoring students warrant. We can only speculate how many teachers give a lower grade to a student because their test scores suggest that they couldn't be worthy of a higher one (e.g., should I really give an A to a student with an EXPLORE score of only 13?). Similarly, they may be reluctant to place otherwise high performing students into honors-level courses due to their low test scores, depriving them of the rigor that they need as well as a weighted grade that would increase their GPA.

With more universal testing, I hope to see a model that is focused on a student's relative performance within their school. Rather than class rank, a more accurate Z-score of their GPA's standard deviation above or below the mean could be used along with the school's average school's test score. For Illinois, the school's average ACT would work for placement decisions for within-state students. This assumes that the GPA reflected course weighting that would differentiate between more-and less-rigorous courses. The student's own test score would still be a useful data point, but only after the student's overall grades and the school's overall test score.

The logic behind this approach is that a test score sometimes can be very inaccurate in measuring individual students. But in aggregate, scores can be a very accurate measure of a school's overall student performance. A student's past track record in the classroom remains the best predictor of their future classroom track record, and knowing how the student's school performed overall would provide the most important additional perspective on their within-school performance. I currently am beginning such a study on our sender schools that might shed some light on this.

60. EPAS scores have their limitations. We can't expect a 2 to 3 hour of test to accurately predict a lifetime of academic potential. But used with measures of past classroom performance, they provide a valuable piece of additional data to assess student performance. Used across groups of students, they can provide very accurate and meaningful measures of department, school, and district performance, information that cannot be found in other currently available sources. These measures, combined with the use of the standards upon which they are based, can provide valuable guidance for instructional improvement efforts.