Script for **High-Stakes Academic Assessment & Educational Evaluation in Illinois**

Hi. I'm Dr. Steve Cordogan, Director of Research and Evaluation at Township High School District 214. We are a district of six high schools, and while my research focus of the past 18 years has been focused on high schools, much of this presentation is relevant to K-8 schools.

High-stakes testing truly is high-stakes. It usually requires many hours of student and teacher time. It provides students with often highly stressful hours of testing and labels them with a score rating that can have life-long implications. The aggregate scores can label schools as desirable or undesirable, impact teacher and administrator careers, and ultimately be used to judge the educational systems of an entire state or even country. As we stand on the threshold of a new testing era, we have to get it right.

Both schools and the public are drowning in academic assessment data. Academic assessment data include test data, grade data, and survey data covering topics such as student social emotional health or parent satisfaction. The volume of data can be overwhelming to those trying to make sense out of it. But not all of the data are accurate measures of performance. Furthermore, many of the attempts to make sense out of the available data provide us with inaccurate information.

2. Test data are vital for measuring student, school, and district academic performance . Test data can be used as a reality check to a school's perception of their own performance by providing comparison data for other schools. Schools can use such data to guide improvements in academic performance, such as identifying which curricular changes enhance student learning.

3. But while tests are highly useful, they have their limits. We cannot expect standardized tests as short as twenty minutes (and, with at least one major ill-conceived computer adaptive test, less than ten minutes) to be as accurate of a measure of student ability as a teacher's appraisal of student performance over nine months of observation and testing. Future student classroom performance is almost always best predicted by past classroom performance.

Test scores are not as potentially authentic a measure as post-secondary information, particularly college data for areas with higher levels of college attendance. But currently, such postsecondary data remain limited.

Additionally, some students may not test as well as they perform in class. Performance in class will almost always predict future classroom performance and future job performance better than a test, since it is a more comprehensive measure of content knowledge and related behaviors (such as organizational skills).

An extremely important factor in test scores is demographic characteristics of the students. Across Illinois, the percentage of a school's free and reduced lunch students can predict 70% of the variance in school ACT performance. In the suburbs, a combination of the percentage of adults in the community with bachelor's degrees and

the percentage of students who are Black, Hispanic, or Native American can explain 93% to 95% of the difference in district ACT performance. So school and district performance largely is determined by demographics.

A proposed way around this issue is to use growth models, such as where the average of student growth over several grades is used instead of a final performance score to evaluate school performance. The idea is, for example, that if a school grows from a student average ACT EXPLORE score of 12 to a score of 18 over the course of the first three years of high school, it is equal to the performance of a school whose students grow from 15 to 21. However, demographics also predetermine most growth levels. Students with more academic at-risk characteristics have both lower initial scores and lower growth. For example, the students with an average score of 12 on the EXPLORE only grow 4 points on the EXPLORE in District 214, while students with an average score of 21 grow over 8 points, more than double. The use of growth as a percentage on initial score helps to ameliorate this issue, though not completely; e.g., our students with a 12 grew 33% to reach a 16, the students with a 21 grew 38% to reach a 29. The magnitude of this percentage difference is still large, but nowhere near the point difference.

Statistical variation, i.e., random fluctuations, can cause substantial differences in school performance between years, particularly for smaller schools and subgroups like special education students.

Tracking school improvement across years also is confounded by consideration of where the school was performing initially and demographic changes in the student body. An example of the first is that it is much easier for a school to improve if it was seriously underperforming initially; the initially high-performing school has much less room for growth. An example of the second is that an increase in the at-risk student population usually will lower academic performance, a change that cannot be attributed to school quality.

The different versions tests themselves are not always consistent. For example, we occasionally have statewide fluctuations in ACT test averages of .2 or .3 points which disappear the following year. These are due to inconsistencies in test difficulty, not some miraculous increase or decrease in the content knowledge by students across all schools in the state.

And finally, some tests are better than others, and some subject tests within the same full test are better than others. For example, the ACT math and English tests are very useful. The ACT reading and science reasoning tests are not. We will look at the spectrum of currently used and future tests shortly.

5. There are a lot of players in testing. There are for-profit and (supposedly) not-for-profit test-making companies. I say supposedly because when salaries reach into seven figures, the not-for-profit company incentive to sell products is not much different from the for-profit companies. There is nothing wrong with trying to make a profit - I

have been a for-profit businessman too, as have many of you. But testing companies are too often and mistakenly assumed to be somehow utterly altruistic and totally trustworthy. Such an assumption is very naive.

When organizations make announcements about test results, they want to receive publicity. Publicity helps to sell their products. Unfortunately, negative news gets more attention than neutral or positive news.

Magazines and newspapers that promote school ratings also are profit-driven. They are not negativity-driven. They are driven to promote the idea that they have the best answer to the question "what are the best schools." The problem is that trying to distill school quality down to a simple measure or even combination of measures does not necessarily accurately reflect school quality

Ultimately, we have to accept the fact that much of the data that we see in the media are bad data, and even good data can be presented in a misleading way.

6. Here is a recent example of such 'misleading' data. This was one of the more extreme examples of a media outlet misrepresenting data, but there have been so many in recent years. And ACT also made no mention of Illinois improvement on their own test. Some local newspaper outlets (such as the Daily Herald) have been willing to acknowledge the flaws in data, particularly sensationalistic negative data.

7. Incidentally, here's the actual Illinois performance, using the accommodations-excluded sample that was used for every year of ACT's reporting since it began. Instead of celebrating the fact that we passed the national average, despite the fact that we test all students, including non-college bound. and the rest of the nation mostly tests college-bound only. So please take what you read in the papers with a grain of salt.

8. I'll single out one of the most extreme misrepresentations of the data, though other organizations have been guilty of similar misrepresentations. ACT has maintained for the past few years that only 25% of U.S. students are ready for college. The conclusion was due to a combination of bad sampling (e.g., only science majors were used to generate the science benchmark, the one that was set higher than the others), ignoring the limited predictive power of its weakest tests, and ignoring the reality that students do not major in subjects in which they are not strong. ACT's own research and its own publications, such as the High School to College Success Report, refute its findings. I would consider the state's School Report Card figure of 45.7 as being much more accurate, but then I am biased, since I proposed and lobbied for the use of a college-readiness percentage based on a composite score, not individual test performance. In fact, college readiness cannot be distilled accurately to a set of test scores, as even the testing company College Board itself has acknowledged. For a full presentation on the topic of the ACT benchmarks, please visit the website on the title page of this presentation.

9. If we want to understand educational data, we first need to look at the essentials of measurement. In order to assess student learning, we need to establish what they need to learn. Illinois created its standards back in 1997. Since my supervisor at the time co-chaired the committee that assembled the standards, I got to be involved with the process, which involved considerable politics as well as genuine academic considerations. They were a start, but they were very broad and difficult to distill down to classroom level curricula.

Our district began to use the ACT standards for Transition, which were renamed the College Readiness Standards. These standards became the standards of choice for our district and many others because they clearly defined the knowledge that was needed in different subject areas. This enabled us to use the standards as a curriculum spine, which was very useful in creating a coherent curriculum. The standards also were linked to different performance levels on ACT corporation's EXPLORE, PLAN, and ACT tests, although the degree to which those tests measure the standards differ between subject areas; some tests are more authentic measures of the standards than others.

The latest standards for most of the nation are the Common Core standards, which currently are being incorporated into the school curriculum at all grade levels. At the high school level, they are fairly similar to the ACT standards, not surprisingly, since ACT was integral to their development. While there is some controversy around the standards, they seem like a comprehensive set of standards that will help to increase rigor in the classroom without going too far. One of the main issues will be how to incorporate the standards into a school's curriculum without stifling all flexibility and turning the classroom experience into an overly simplistic test prep experience. The other will be what test is used to measure them.

10. There are many high-stakes tests upon which public schools are focused. When we look at college placement tests, we see the array presented on this slide. In Illinois and the middle states, ACT is the primary college placement test, with the SAT being used on the east and west coasts. The COMPASS and ACCUPLACER tests are for community college admissions which, according to recent major research, have little validity and tend to over-place students in remedial courses when they could have handled college-level work.

The AP program allows students to take courses with college-level rigor in high schools. They traditionally have been focused on students at the honors level. However, we and other districts have expanded our AP course access to students with average and even below average placement test scores when they demonstrate sufficient motivation to attempt the courses. There has been minimal negative impact on our AP test pass rates, and we have far more than tripled our number of successful AP test taking in the past twelve years, demonstrating that far more students should have access to such high-rigor courses than was thought in the past.

11. The next slide shows tests currently used by the State of Illinois to evaluate students at the high school level. The Prairie State test consists of the useful and meaningful ACT test, a WorkKeys workplace readiness test that never has been scientifically evaluated and is of questionable validity or meaning to our students, and a state-developed science test that, given the wide range of science topics, cannot meaningfully measure science knowledge; it largely is ignored. There are tests for highly cognitively disabled students which do not have much credibility in the world of special education. Finally, there is an ACCESS test for English Language Learners which is used to test language proficiency, but not course content. It has a reasonable amount of credibility in the English Language Learner community.

12. Many schools in Illinois currently use the EPAS testing system, which consists of the ACT college entrance exam and two scaled down versions of the test, the EXPLORE and PLAN, tests soon to be replaced by similar ACT ASPIRE tests. The EXPLORE often has been used for making 9th grade placement decisions. The PSAT test is a scaled-down SAT test which is used to identify student eligibility for the National Merit Scholarship Program. The AP test was described earlier.

13. The ISAT is the state test that will be given for the last time this spring. While it is a reasonably good test, the level that was set for meeting standards on the tests was much lower than for the PSAE until 2013, leading to the confusion of having students labeled as meeting standards in grade school but as substandard in high school. The EXPLORE was described above. The other tests have been used for many years for K-8 evaluation and placement decisions, and have good reputations for accuracy. The MAP test currently is the most commonly used K-8 test in the Chicagoland area. It (as are most current tests) is marketed as being aligned to the Common Core standards. Its computer adaptive nature is a plus, and it seems to be reasonably valid, but it can time consuming (particularly if given multiple times during the school year).

14. But the testing horizon is changing substantially. Illinois has affiliated itself with the PARCC consortium of around 15 states, which have banded together to offer an assessment system developed from scratch from the Common Core Standards. ACT withdrew from its role as subcontractor under PARCC to develop a competing suite of tests. While Illinois has repeatedly declared that it will use PARCC, there are reasons to question whether such a transition will occur.

The next slides look at the pros and cons of using PARCC and ACT.

15. Staying with ACT will mean that the students will continue to take the ACT test for which we have been preparing students and a test that they take seriously, which enhances its validity as a measure of student performance. The current ACT suite of tests have proven very useful for school improvement efforts.

16. However, the ACT tests only have two genuinely valid subject tests, math and English, though the composite score also is very useful. Please note that these tests are too short to be ideal instruments. Additionally, the reducing of the EXPLORE and

PLAN to one test will limit the validity of the test, since it is only slightly longer and must cover two grade levels. ACT's tests are not much different from their original 1959 test. We can do better.

Additionally, in the past few years, ACT has been much more interested in survival and market expansion than in producing good products. They have misused data to bash schools in order to promote themselves and their products. All recent research has found that their community college test is of very limited validity, and their other product offerings have not been found to be of quality. Their presentations and follow-up discussions have shown that they are at least as unready as PARCC to roll out their suite of tests. We cannot assume that they will develop a sufficiently useful test for the future.

17. I wanted to provide a brief illustration of the limits of the ACT tests. In a 2005 study meant to support the benchmarks, the first bar in their graph showed that 65%, almost 2/3, of the students who met NONE of the benchmarks, the lowest testing students of all, persisted to their second year of college with a C+ average. If meeting the benchmarks was critical, how could so many do so badly on the test and still succeed?

Additionally, meeting only the science benchmark added a mere 2 percentage points to the persistence rate and .11 points to the GPA. This is a much lower level of improvement than if they had only met the English or mathematics.

Finally, you may notice that something is missing from this study. Reading apparently was omitted because its inclusion would have reduced, or made no difference in, the prediction level. From ACT itself, we can see that the ACT reading test is relatively useless. This confirms our own research.

Besides illustrating the problem with the design of the setting of the benchmarks, the graph shows that ACT's have limited predictive power for the college classroom, even though that is what they were designed to do. We can do better.

18. On the other hand, PARCC at least has the promise of recognizing the need for better, more valid instruments, with greater length and alignment to the Common Core. They began with a fairly clean slate, although most of the major players in PARCC development were existing testing companies. They have not yet displayed the bad behaviors that ACT has exhibited.

19. However PARCC quality is a complete unknown. They already have been behind schedule for their rollout, and the defection of several states makes it uncertain that they will even continue to exist. Their tests are considerably longer. While this is a good thing in terms of greater reliability and validity, the extreme to which they have taken it will be a huge burden for schools. A significant piece of the above time issue is their open-ended allowance for students needing additional time, which can extend their testing time to an additional 50% for any student wanting it. This testing time demand will be much greater because we will need to administer the ACT for at least another

two years so that we can establish continuity with our current school evaluation efforts, and so that PARCC can prove itself as a viable college admission test. If it is not accepted as an admissions test, students will take it much less seriously, thus reducing its validity. Finally, and not surprisingly given its greater length, PARCC is significantly more expensive than the ACT tests. This issue alone may give the state an out from its commitment to PARCC - we can simply say that we cannot find the extra money required to use it.

Incidentally, both ACT and PARCC will be primarily computer-administered, which will allow easier scoring and the potential for some interactive items. A paper version of the test will cost significantly more. Unfortunately, unlike Smarter Balanced, the consortium for the rest of the U.S., neither ACT or PARCC will be computer-adaptive. Computer-adaptive tests, when done well, can tailor the item difficulty to the student's abilities, providing for more accurate measurement. So neither suite of tests is ideal in this respect.

Additional information on the PARCC versus ACT issues can be found in a report on my website.

20. Standardized tests are the Swiss Army Knife of tests. One test can contain several subject tests just like the knife can contain several tools (a blade, a screwdriver, etc.) in one package. However, just like there are better screwdrivers than the one on such a tool, there are better measures of subject performance than current standardized tests. For example, an essay test graded by a qualified educator can give much greater insights into student understanding than a few multiple choice questions. PARCC, and to a lesser extent, ACT, are trying to incorporate such greater measures of authenticity.

However, the increased subjectivity, as well as the skilled grader time required to grade such tests is the primary reason that most high-stakes standardized tests have been exclusively multiple choice. The reliability of subjective tests is often much lower than multiple choice or other objective tests, and the grading is much more expensive. It is noteworthy that SAT's recently announced redesign did not include a move to such more subjective items, and made their essay optional. In my opinion, these were wise moves.

AP tests often have had such subjective elements. But AP tests, at $89 for a single subject test, are more than triple the cost of an entire ACT or PARCC test. They are graded by skilled educators, not lower-paid clerical staff with an often simplistic rubric, as have been done with some attempts at more subjective testing in the past. And subjective grading by a computer currently is totally unproven. So properly administering more authentic tests on a regional/national scope for all students probably will require a greater investment of resources that we as a society are willing or able to commit. Until we are willing to make such an investment, we would do better investing in quality multiple choice items, which can measure critical thinking and a depth of understanding to a reasonable extent, and leave the deeper probing to local assessments.

Incidentally, an examination of some of the new ASPIRE items that require short answers have revealed that they simply are old EXPLORE and PLAN questions with the multiple choices removed. While this may provide slightly greater authenticity, they are a relatively small step forward.

21. We already have tests that can compare across the nation and the globe. For example, the NAEP is presented as "The Nation's Report Card," and considerable press is given to annual announcements of test score results, usually presented in the negative as a lack of growth in scores. The PISA and TIMMS examine our performance relative to other nations, again with an emphasis on relatively low U.S. performance

22. We do need to improve our academic performance across the nation. Using measures more sensitive to student performance, such as the ACT, we have seen how much improvement has been possible. For example, Illinois growth on the ACT demonstrates that we were not living up to our potential ten or twenty years ago. But the attention paid to these national and international tests is largely unwarranted, for the reasons listed.

The sampling procedures alone render the findings invalid. They often beg schools to participate, and even provide "incentives" like trips to conferences for administrators. Our universal education system is not found in many other countries, and the percentage of impoverished children in our sample often is much higher than other nations participating in the testing.

Most importantly, our students, already experiencing test burnout, have little motivation to perform on the test. Little feedback may be given on the test (e.g., NAEP provides none to student or school). There is no motivation for students to perform. We see students who make Christmas tree patterns and spell out often unrepeatable words on the test.

The importance of motivation cannot be overstated. For example, in District 214, we have made huge increases in AP and ACT performance, upon which we have focused. Yet our growth on the PSAE, even though it contains the ACT, has been very small. Motivation is very critical to student performance.

23. As mentioned earlier, several media outlets produce their own ratings. Three focus on AP (one exclusively), some also focus on ACT, some incorporate state tests, some incorporate the critical factor of demographics, and a variety of other considerations may be included. The ratings for all of these outlets generally include higher performing schools, but many lower performing schools also can be found in the rankings. This is apparent from a look at the ACT performance for the Illinois schools included. Despite its flaws, the ACT remains the best measure of school performance in Illinois and other universal ACT-testing states.

Also, schools that earn a high ranking in one report may be totally excluded from another. Additionally, some rankings include private schools and selective enrollment schools. These cannot be considered comparable to open admissions public schools.

Finally, some have unquestioningly used National Center for Educational Statistics (NCES) data (the same organization behind NAEP and TIMMS (and, to a lesser degree, PISA). NCES data have been inaccurate, and AP data keepers have not developed filters for bad data. In fact, the U.S. News and World Report study in 2012 led to a federal investigation of NCES, which provided some grossly inaccurate data (e.g., 0% poverty levels for Maine Township schools that had as high as 45% poverty levels).

So the rankings, while they provide schools with bragging rights, should be considered a mix of useful academic data and media fluff. There is no accurate national test as yet which can be used to rate schools. The lack of consistency between their ratings demonstrate that in the absence of a national test that is administered equitably and taken seriously (i.e., not the NAEP), the idea that schools can be accurately ranked across states with one measure, regardless of the complexity of the data that went into the measure, is unrealistic. Even Chicago Magazine, which dealt only with Chicago and suburban schools and had ACT scores and demographics available for them, produced rankings that could not be cross-validated.

24. There are other sources of school rating data. We already have mentioned the ACT College Readiness Benchmarks as an unfortunate marketing manipulation of the media. ACT also provides the aforementioned High School to College Success Report, which is supposed to provide a comprehensive overview of college success. Inasmuch as it only included 37% of District 214 college attendees, and was missing a high percentage of public Illinois post-secondary attendees who should have been included, its data are of very limited use.

The original ISBE school report cards, which soon will be replaced, have included ISAT scores with overly lax cutoff scores for meeting standards, overly strict PSAE cut-off scores for meeting standards, and PSAE tests which include subtests that have not been validated (as mentioned above). However, there is much useful ACT and demographic information in the report cards which, when contrasted across schools, can be very useful in identifying school performance. The new on-line version of the report card makes the data even more accessible. The traditional Illinois Interactive Report Card has made detailed data access possible for years, and has made cross-district and school comparisons possible. All school report card data are contained in a single database available from ISBE, but each of the almost 4000 state schools has 10,000 pieces of data, so finding the useful information can be highly challenging.

Please note that these are only the reports that are widely publicized. Other national reports on educational quality (e.g., Education Week's Quality Counts) are compiled but are not as strongly marketed and therefore do not reach a mass market.

25. So despite all of the negatives discussed above, there are useful data which, when used carefully, can provide meaningful evaluations of school performance. While national and international comparisons as of yet cannot be made accurately, contrasts within Illinois can be made. Using the English, math, and composite scores from ACT provide us with data that do measure overall school quality rather well. ACT may have much room for improvement as a test, but it remains very useful.

The scores can be used to gauge improvement over time. AP performance, particularly the number of passing AP tests per student or % of graduates passing at least one AP test, provide us with a good gauge of the rigor provided by a school. Additionally, growth from the EXPLORE test to the PSAE ACT test has provided a measure of what progress individual students have made while attending a given high school.

MAP tests can similarly provide such information at the elementary and middle school level.

As mentioned in slide 3, test scores are not as potentially authentic a measure as post-secondary information, particularly college data for areas with high levels of college attendance. But currently, such postsecondary data remain limited.

26. Here are some sources of college data. We already have discussed the limitations of the ACT High School to College Success Report. Many schools already provide exit surveys which solicit college attendance data. But such data are inaccurate due to self-reporting bias and self-selection. Students who fail to respond to such surveys usually are disproportionately less engaged in the school and less successful than those who do. Even among respondents, college plans have been found to be consistently more positive than subsequent attendance patterns.

Professional follow-up surveys face the same selective response issues, limiting their effectiveness. They can provide useful information on more successful students who are at least sufficiently connected to their high schools to respond, but cannot be considered to provide a representative picture of all students.

Illinois has been working on a tracking system across all grades for many years, but it has not yet been put into operation.

This leaves the National Student Clearinghouse, with a claimed tracking rate for 96% of the students attending post-secondary institutions. The information provided has attendance and graduation data (e.g., even degree earned and major), although GPA and other details are not yet provided.

Hopefully, the Clearinghouse will continue to evolve and more data will be gathered. However, assessing college success becomes very complex, since a consideration of the quality of the institutions attended also needs to be incorporated into any evaluation information (e.g., is a two-year 3.7 GPA at Harper better than a 2.8 at U of I?). Also,

college attendance takes years to assess, since students do not always attend right after high school and may not complete for at least four to six years.

27. Related to all of the above considerations is the current plan to use of test scores to evaluate teachers. It is perfectly logical to expect teaching quality to impact test performance, but creating a system for evaluating such performance is nearly impossible. The main issues are finding an instrument appropriate to the class content, having enough classes, teachers, and students to have meaningful sample sizes, factoring in student demographic characteristics that would impact performance and growth, and accounting for other related courses that might impact performance in the targeted courses. For most classes in most schools, this is not possible at the high school level. It might be slightly more feasible at the grade school level (my experience at this level is limited), but would remain a daunting task. Such scores should be used as confirmatory evidence in a more subjective evaluation of a teacher by an experienced administrator. However, there is more potential to use test scores to evaluate administrators. If two schools have comparable demographics, and one significantly outperforms the other, an accurate data-driven evaluation is much more feasible.

A paper detailing the intricacies of teacher evaluation using test data also can be found on my website.

28. In conclusion, we need to pick our data carefully, use it carefully, take any headline with a large grain of salt, and remember the adage "Not everything that counts can be counted, and not everything that can be counted counts." Educators have plunged into the era of data-driven accountability unprepared for the politics and test-company marketing that has reduced the discussion from school improvement to often questionably accurate test scores and ratings.

We cannot go back to the days where schools were free to have a self-image that may have had no basis in reality and may have been hurting the learning potential of its students. Data can be intelligently used to measure students, teachers, and schools. But there need to be multiple data points, and such data should consider all of the factors that might impact student performance. No student or institution can be reduced to a single number.