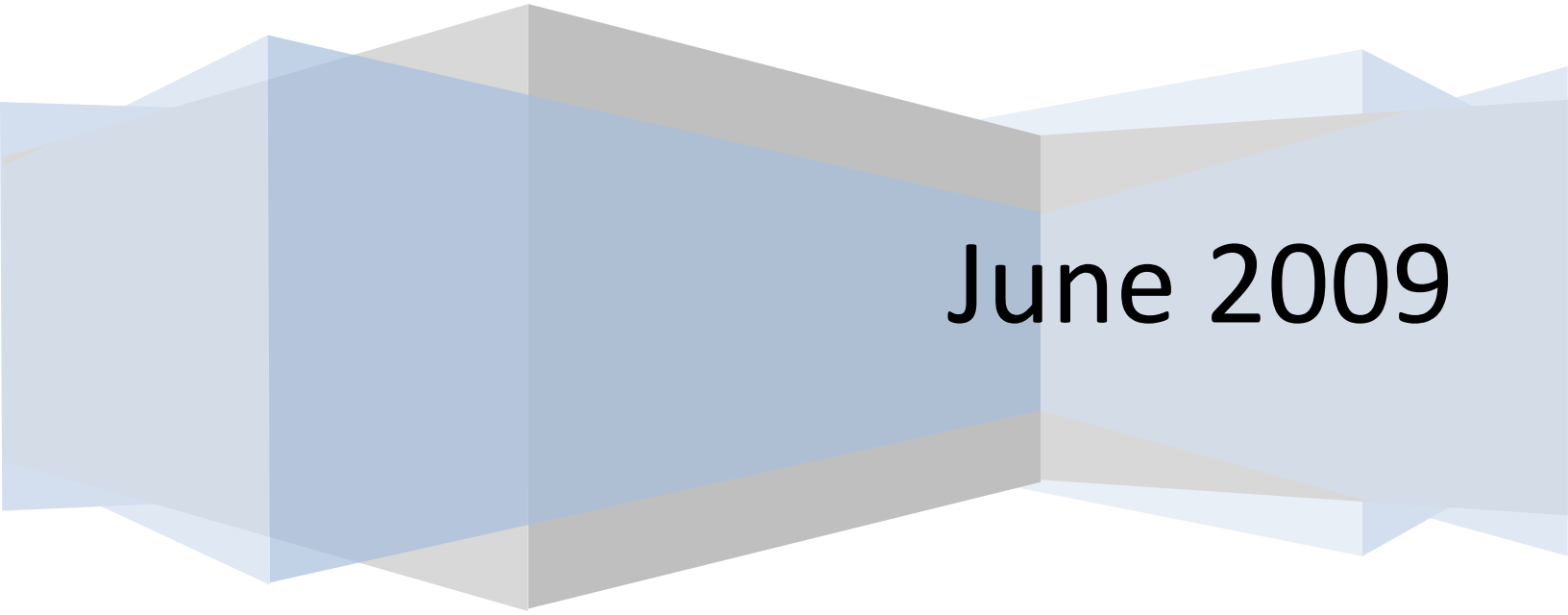
Urban School Leadership Program
University of Illinois-Chicago

Something's Wrong with Illinois Test Results

Sources and Remedies for the Growing Pains of Standards-Based Assessment

Paul Zavitkovsky
pzavit@uic.edu

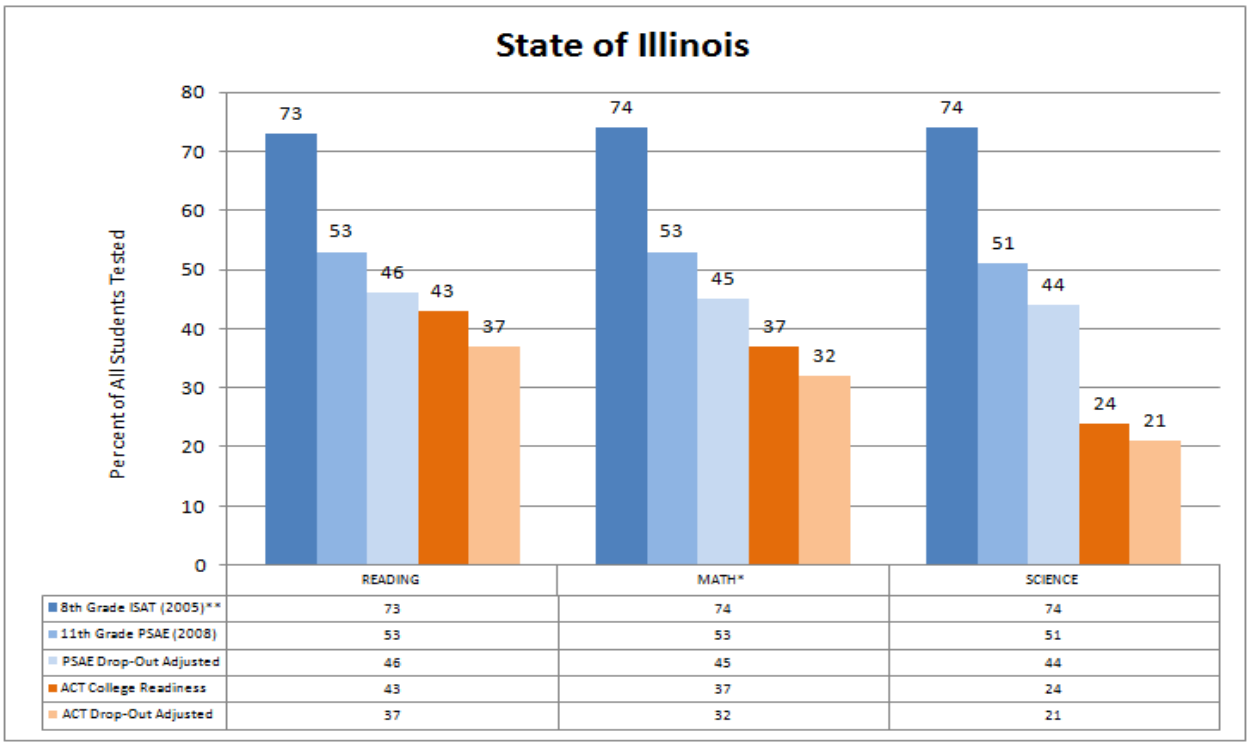


June 2009

Something's Wrong with Illinois Test Results

After close to a decade of annual "standards-based" testing in the State of Illinois, many school boards are perplexed by the contradictory test results that are being reported for their elementary and high school programs. Figure 1 shows how this problem looks statewide for the high school graduating Class of 2009.

FIGURE 1: ISAT/PSAE/ACT Comparisons for the Class of 2009



	READING	MATH*	SCIENCE
8th Grade ISAT (2005)**	73	74	74
11th Grade PSAE (2008)	53	53	51
PSAE Drop-Out Adjusted	46	45	44
ACT College Readiness	43	37	24
ACT Drop-Out Adjusted	37	32	21

*8th grade math scores for 2005 have been adjusted to reflect changes in Illinois math benchmarks that were adopted in 2006
**ISAT does not test for science in 8th grade; science numbers reflect 7th grade scores from 2004

In 2005, well over 70% of the 155,000 8th graders from the Class of 2009 who took the ISAT exam met or exceeded Illinois Learning Standards in reading, math and science. Three years later, just a little more than 50% of these same students were meeting or exceeding standards on the 11th grade PSAE. Adjusted to include likely results for the 22,000 students who dropped out of school between 2005 and 2008, these numbers drop to the mid-40s[1].

Discrepancies between middle school ISAT and high school ACT results are even more disconcerting. In the worst case, less than a third of the 74% of students from the Class of 2009 who met or exceeded Illinois science standards in middle school were able to meet or exceed ACT college readiness standards just three years later.

[1] "Drop-Out Adjusted" numbers show rough estimates of what PSAE and ACT College Readiness percentages *would have been* if all students who were tested in grade 8 were also tested in grade 11. These estimates are calculated by dividing the number of 11th graders who met PSAE and ACT benchmarks by the total number of students who were tested as 8th graders (or as 7th graders in science)

FIGURE 2: ISAT/PSAE/ACT Comparisons for the Class of 2009

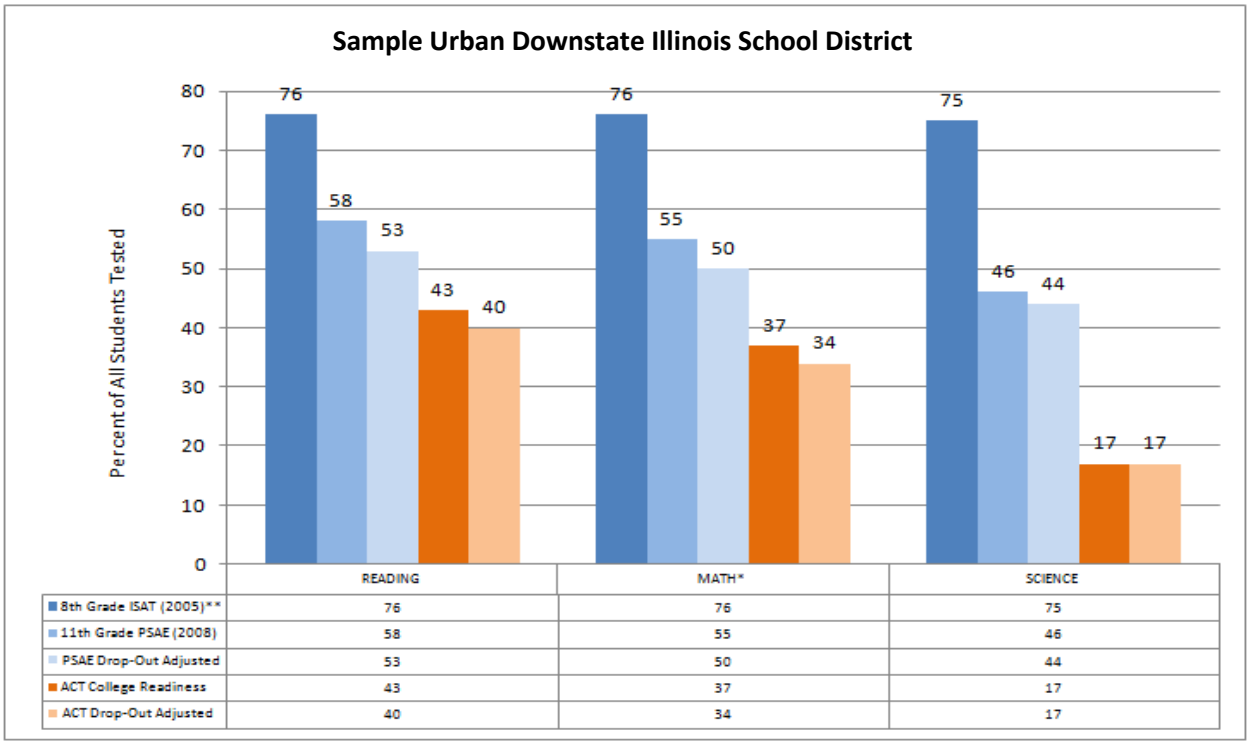


Figure 2 shows comparable outcomes for the 500+ students from the Class of 2009 who will be graduating this year from a sample urban downstate Illinois school district.

TAKING A CLOSER LOOK

Shown below are comparisons of Illinois proficiency ratings for students in this district from the Class of 2009 during middle school and high school:

FIGURE 3A: Illinois Standards Proficiency for 8th and 11th Grade Reading

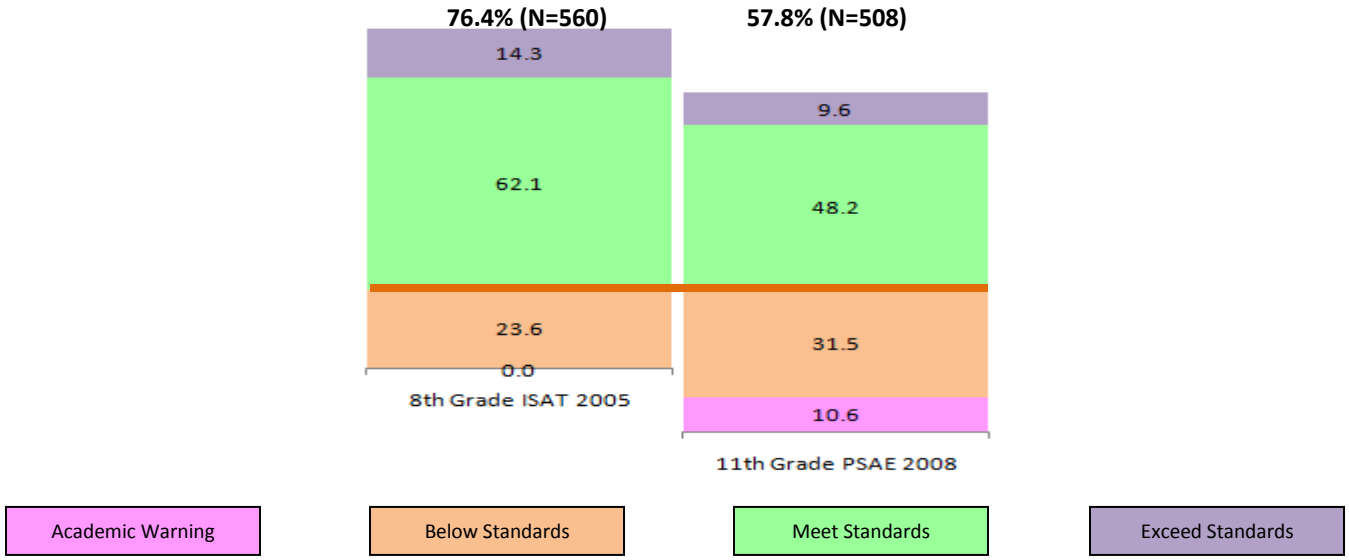


FIGURE 3B: Illinois Standards Proficiency for 8th and 11th Grade Math

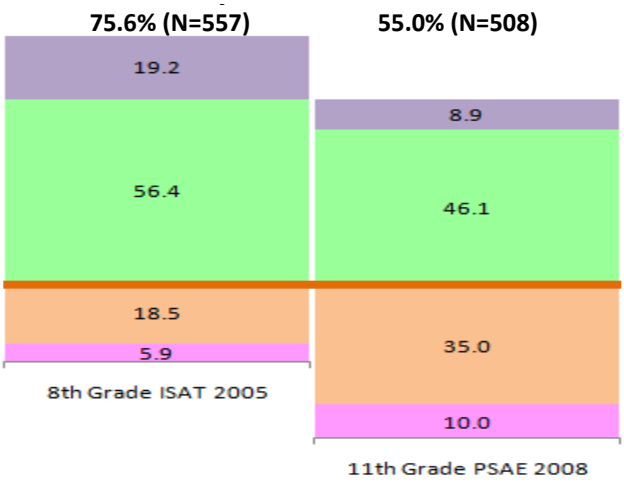
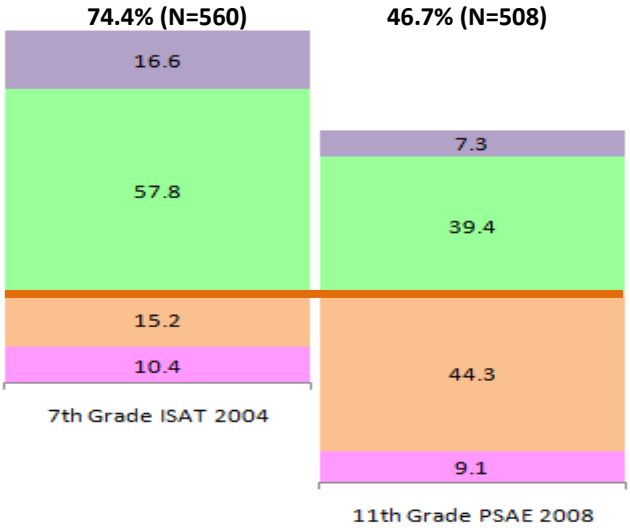


FIGURE 3C: Illinois Standards Proficiency for 7th and 11th Grade Science



On their face, results like these suggest that elementary and middle schools in this district and across Illinois are doing a solid job of preparing most of the state's young people but that high school programs are badly squandering their advantage. An alternative explanation is that relatively low proficiency benchmarks on the ISAT are creating inflated results that set up large numbers of students to under-achieve or fail once they get to high school.

A recent study[2] of same-student, ISAT-ACT scores in the Chicago Public Schools supports this view. That study showed that Chicago 8th graders whose math and reading scores just met Illinois Learning Standards in 2004 had less than a 5% likelihood of meeting college readiness benchmarks as high school juniors in 2007.

[2] Easton, John Q., Ponisciak, Stephen and Luppescu, Stuart (October 2008) *From High School to the Future: The Pathway to 20*, Consortium on Chicago School Research, University of Chicago Urban Education Institute http://ccsr.uchicago.edu/content/index.php

What's To Be Done?

This brief has two purposes:

1. To summarize key factors that are causing alignment problems among the ISAT, PSAE and ACT exams
2. To illustrate some promising ways to report ISAT, PSAE and ACT data that alleviate these problems

ISAT CHANGES IN 2006

In 2006, the State of Illinois made major revisions to the ISAT testing system for grades 3 through 8 in order to comply with the requirements of No Child Left Behind. The 11th grade PSAE exam remained unchanged.

A key feature of the revised ISAT system was the adoption of a common metric for all grades tested. No comparable provisions were made to link the metrics of the ISAT and PSAE.

Adoption of a common metric in mathematics required the State Board of Education to decide whether to raise benchmarks in grades 3 through 7 to meet the existing 8th grade standard, or to lower the 8th grade benchmark to match existing standards in prior grades. The Board opted to lower the 8th grade benchmark.

The 11th Grade PSAE and ACT College Readiness Benchmarks

The PSAE has two parts. Part I is the ACT sequence of English, math, reading and science reasoning exams that many colleges require for admissions. Part II includes an Illinois test of science content knowledge and two workplace-oriented tests of applied reading and math skills called Work Keys.

ACT, Inc. has identified college readiness benchmarks for each of the four tests in the ACT sequence. These benchmarks are 18 in English, 21 in reading, 22 in math and 24 in science reasoning.

College readiness benchmarks are based on the actual achievement of college freshmen at the 900+ colleges and universities that require ACT scores for admission. Meeting ACT college readiness benchmarks in 11th or 12th grade means that a student has a 50% likelihood of doing B-level work or better in freshman courses . . . and a 75% probability of doing C-level work or better.

DIFFERENT RULES FOR REPORTING NORM-REFERENCED AND STANDARDS-BASED RESULTS

In recent years, a lot of public attention has been given to the difference between "standards-based" tests like the ISAT and PSAE, and more conventional "norm-referenced" tests like the Stanford 10 (SAT-10) and the ACT. In practice, design differences between these two types of tests are actually pretty small. Both use similar questions and each produces a normal distribution, or bell curve, of student scale scores[4] each time the test is administered. The place where differences become pronounced is in the way that ***proficiency benchmarks are set*** and the way that ***results are reported.***

NORM-REFERENCED RULES

Results for conventional, norm-referenced tests like the ACT and SAT-10 are defined and reported ***in comparison to all other students who took the test***. Typical report categories include:

Percentile: On a scale of 1 to 100, how does your score compare with the scores of all other students tested? A 40th percentile score is as good or better than 40% of all other scores.

Quartile: When all scores are rank-ordered and split into four equal groups, which group does your score fall into? A top quartile score is in the top 25% of all scores, or somewhere between the 75th and 99th percentile.

Grade equivalents: How does your score compare to the average scores that most students achieve at different grade levels? A grade equivalent of 5.5 means that your score is the same as the average score for students who are in the 5th month of 5th grade.

Percent At/Above Grade Level: What percentage of students at your school had scale scores that were at or above the average scale score for their grade level?

Stanine: A fixed range of percentile scores that is statistically significant.[5] Shifts of a stanine or more in the average scores of a group are likely to reflect real growth or decline in the normal pace of learning. Shifts of less than a stanine often reflect real learning changes but may also be due to testing error, changes in testing procedure or other factors.

STANDARDS-BASED RULES

The disadvantage of norm-referenced scoring strategies is that they do not report how well students have mastered particular kinds of knowledge and skill. A student who scores in the bottom quartile of a highly proficient population may still have mastered a fairly wide range of important knowledge and skills. But no matter how proficient students are, 50% of a total test population will always score above average and the remaining 50% will always score below average.

[4] Scale scores are based on the number of correct answers that a student gets. Some scale scores are calculated by weighting selected questions more than others. Other scale scores (like those for the ACT) simply reflect the total number of correct answers.

[5] The technical definition of a stanine is a range of percentile scores that equals half a standard deviation from the mean of a normal distribution. Stanine ranges are larger in the middle of a distribution than on either extreme. For example, stanine 5 includes scores from the 40th to 59th percentile. Stanine 8 only includes percentiles from the 89th to the 95th percentile.

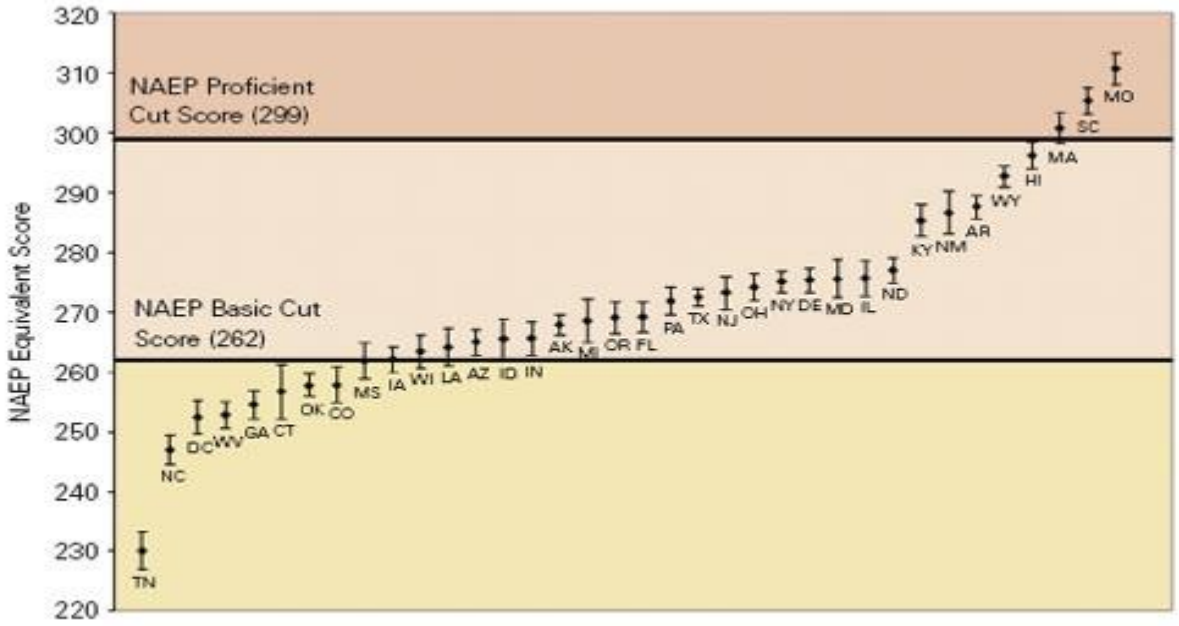
Standards-based tests have attempted to resolve this problem by setting ***fixed proficiency benchmarks*** that purport to represent a particular level of mastery. Once a student obtains a scale score that exceeds this benchmark, a positive result is reported regardless of how that score compares to the scores of other students.

RIGOROUS STANDARDS VERSUS RIGOROUS SCORING

While sensible in theory, the practical effect of the shift from norm-referenced to standards-based scoring has been that many elementary and middle school benchmarks are ***no longer grounded in the high-stakes outcomes of college readiness and international competitiveness*** that they were originally created to support. Taken together with the real-world pressures of NCLB accountability, the net effect in most states has been a substantial lowering of expectations for elementary and middle school benchmarks.

Figure 4 illustrates how the proficiency benchmarks for 8th grade math in 35 states and the District of Columbia compared with 8th grade proficiency benchmarks on the 2005 National Assessment of Educational Progress (NAEP). Unlike most state benchmarks, NAEP cut scores are closely aligned to TIMSS, PISA and other international studies of student achievement.

FIGURE 4



Source: National Center for Educational Statistics [printed in *Education Week*, June 7, 2008]

Note that all but three of the state benchmarks shown in Figure 4 fall below the NAEP benchmark for “proficiency" and nine fall below the NAEP benchmark for “basic” mastery. It is also worth noting that the comparatively high benchmark shown for Illinois was lowered

substantially by the State Board of Education in 2006. The current Illinois benchmark for 8th grade math sits at or below the NAEP benchmark for basic mastery (see also footnote 12 on p.15.)

HIGH STAKES ASSESSMENT LITERACY

All these problems are compounded by the fact that most parents, teachers and school administrators are not familiar with the technical distinctions between norm-referenced and standards-based reporting practices. In everyday school conversations, these distinctions often get blurred. As a result, Illinois' four proficiency categories of Academic Warning, Below Standards, Meet Standards and Exceed Standards are regularly equated with norm-referenced quartiles. It is also fairly common to hear teachers, parents and administrators describe the percentage of students who meet or exceed standards as students who are scoring "at or above grade level."

Figure 5 illustrates how far off the mark these common-sense connections actually are.

FIGURE 5

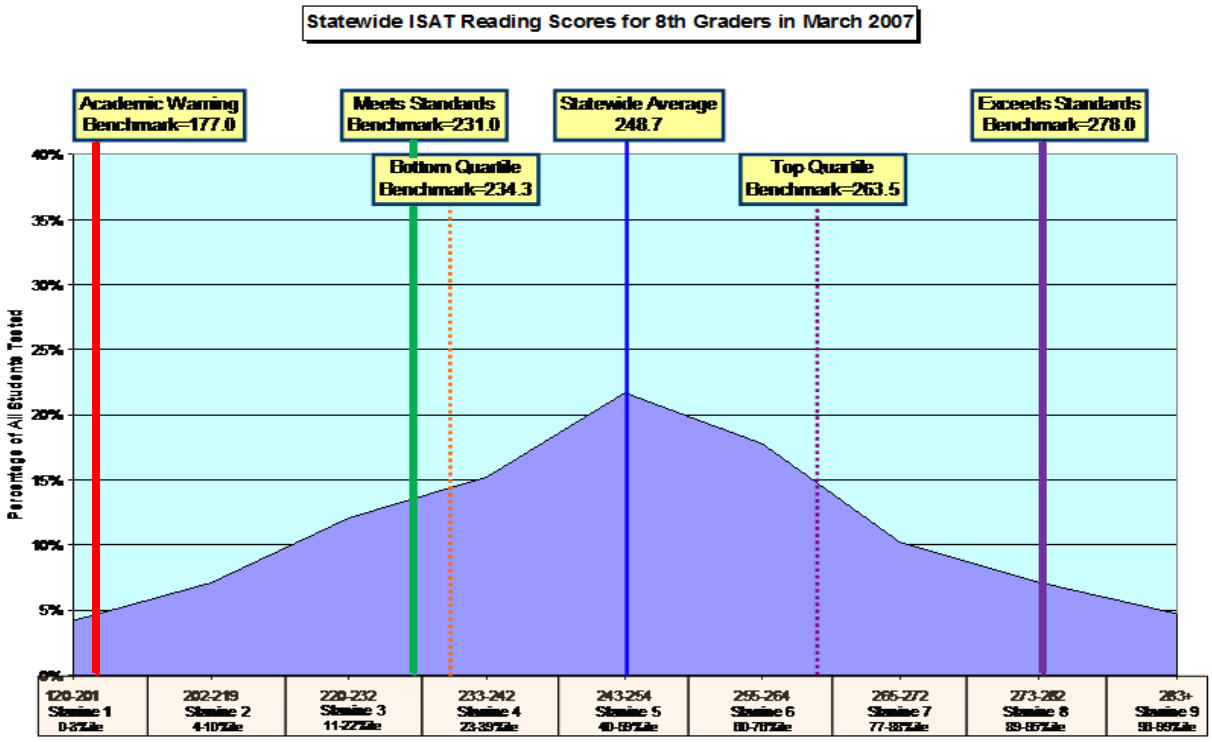


Figure 5 shows the distribution of scale scores for 8th grade reading on the 2007 ISAT. Notice that the green Meets Standards benchmark is at about the 20th percentile (stanine 3) of the full distribution. That places it in the upper range of the ***bottom quartile . . .*** about 30 points shy of the statewide, grade-level average. The red Academic Warning benchmark is below the 2nd percentile. Meanwhile the purple Exceeds Standards benchmark is at the 92nd percentile in the upper third of the top quartile.

Some Promising Ways to Resolve Alignment and Reporting Problems Using Existing Test Scores

The Urban School Leadership Program at the University of Illinois—Chicago has developed a promising set of report strategies that are designed to ***tighten the relationship*** between ISAT, PSAE and ACT scores and ***increase the value*** of those scores for policy and practice. A sampling of these strategies is outlined below using data from school district described on pages 3 and 4.

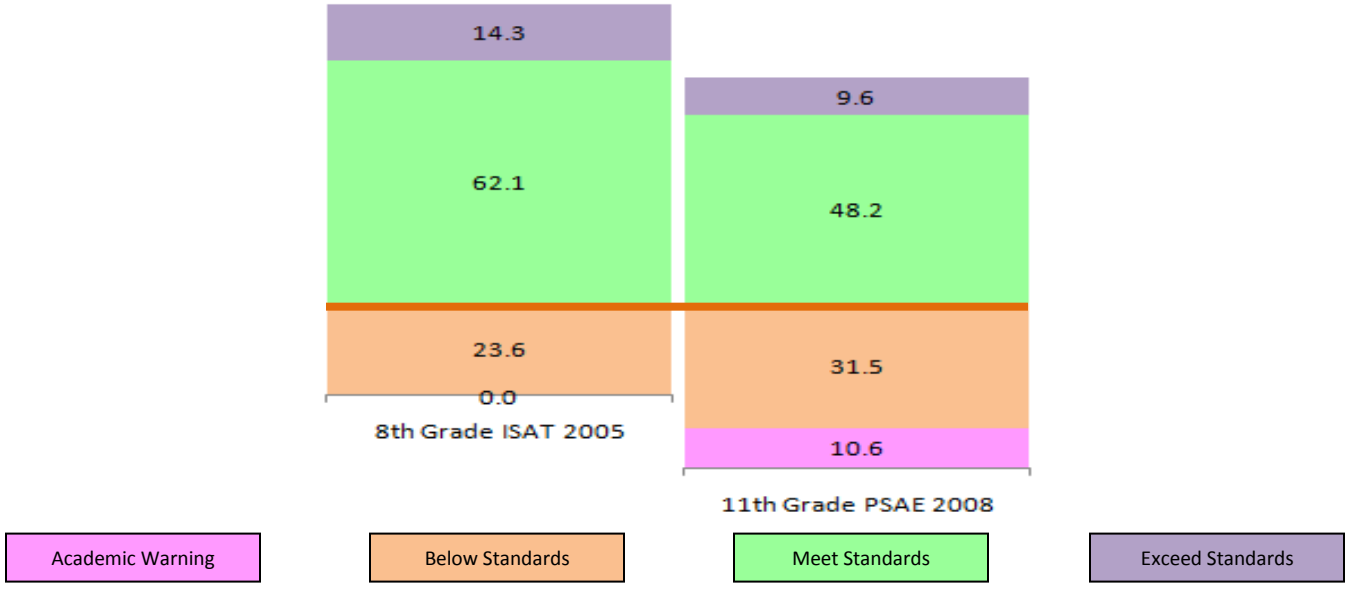
TIGHTENING THE RELATIONSHIP BETWEEN ISAT AND PSAE REPORTS

The only fully reliable way to determine the relationship between elementary and high school test scores in any school district is to match the actual scores for several cohorts of students who take both tests[6]. The problem with this approach is that many years of same-student data are required to complete a full cycle of match-ups.

With these limitations in mind, a simpler way to tighten the relationship between the ISAT and PSAE is to report the results of each test in norm-referenced as well as well as standards-based terms. This approach has two major advantages for policy and practice:

1. It dramatically increases the predictive power of ISAT results for later achievement on the PSAE
2. It creates a more reliable tool for measuring value-added between middle school and high school

Consider the earlier example from Figure 3A on page 3 that showed very dramatic declines in reading proficiency between middle school and high school:



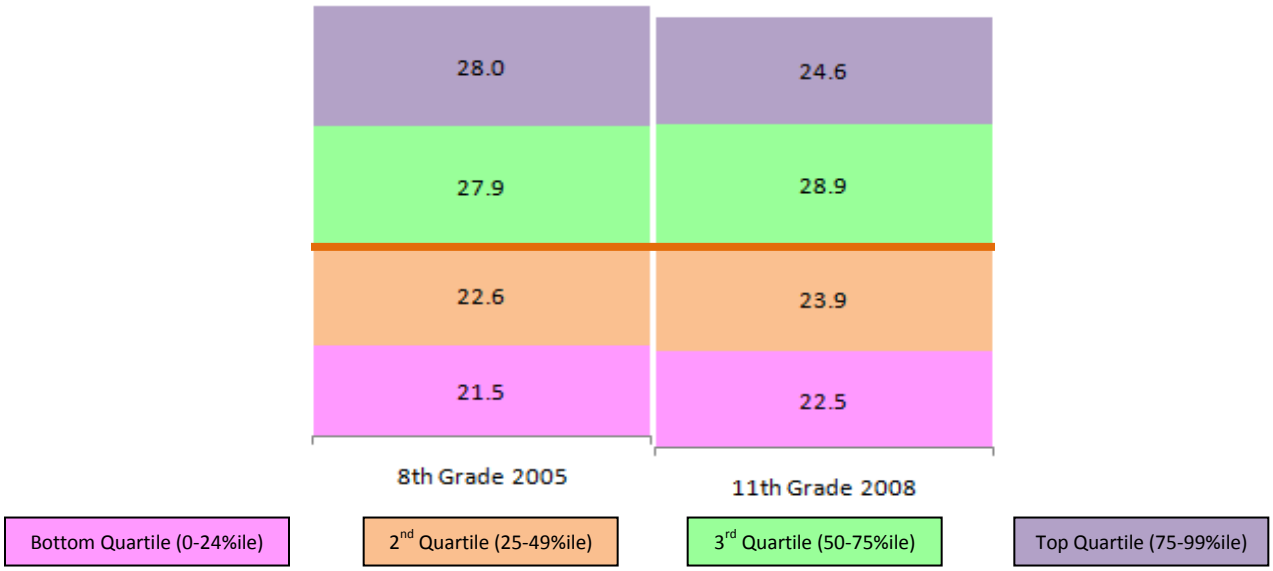
[6] This is what the Chicago Consortium did for the City of Chicago in *From High School to the Future* (see footnote 2). Because the State of Illinois adopted statewide student identification numbers in 2006, similar studies will be possible on a state-wide basis at the completion of the 2009 testing cycle.

When the exact same data are reported as Illinois ***quartile distributions***, the slip in reading achievement between 8th and 11th grade takes on a far less catastrophic look:

FIGURE 6A: Score Distributions in Illinois Quartiles

At/Above Illinois Grade Level Average

8th Grade: 55.9% **11th Grade: 53.5%**



A similar picture emerges when achievement is reported in Illinois ***stanine distributions***.

Why Benchmark Stanine 5?

The cut point between stanine 4 and stanine 5 is the 40th percentile.

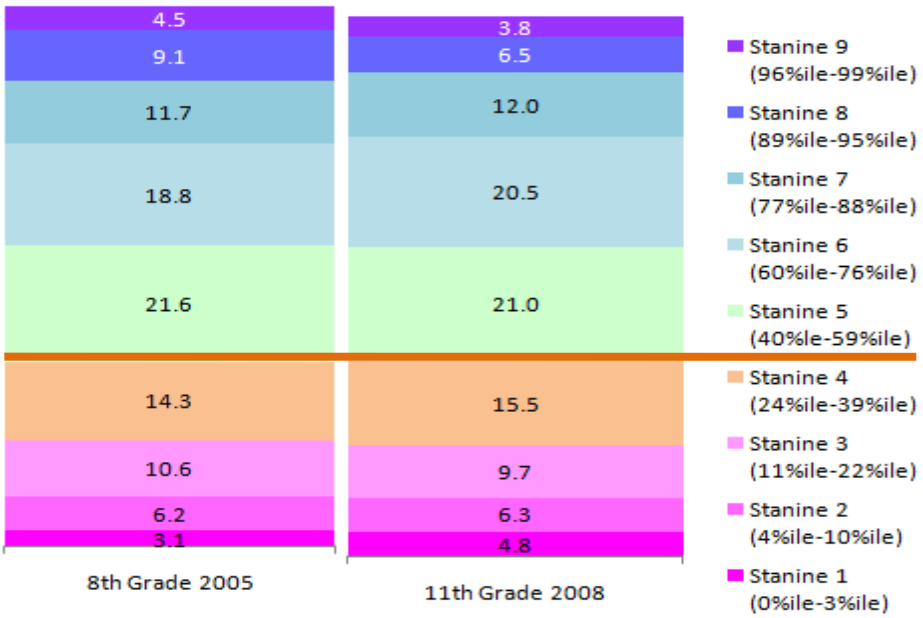
In 2006, the Illinois State Board of Education opted to peg the Meets Standards benchmarks for all ISAT subjects and grade levels to the 38th percentile of the SAT-10 national achievement exam.

The technical rationale for this decision is detailed in, "Report on the ISAT/SAT-10 Bridge Study and Development of the 2006 ISAT Reporting Scales" January 6, 2006 http://isbe.net

FIGURE 6B: Score Distributions in Illinois Stanines

Stanine 5 and Above

8th Grade: 65.7% **11th Grade: 63.8%**

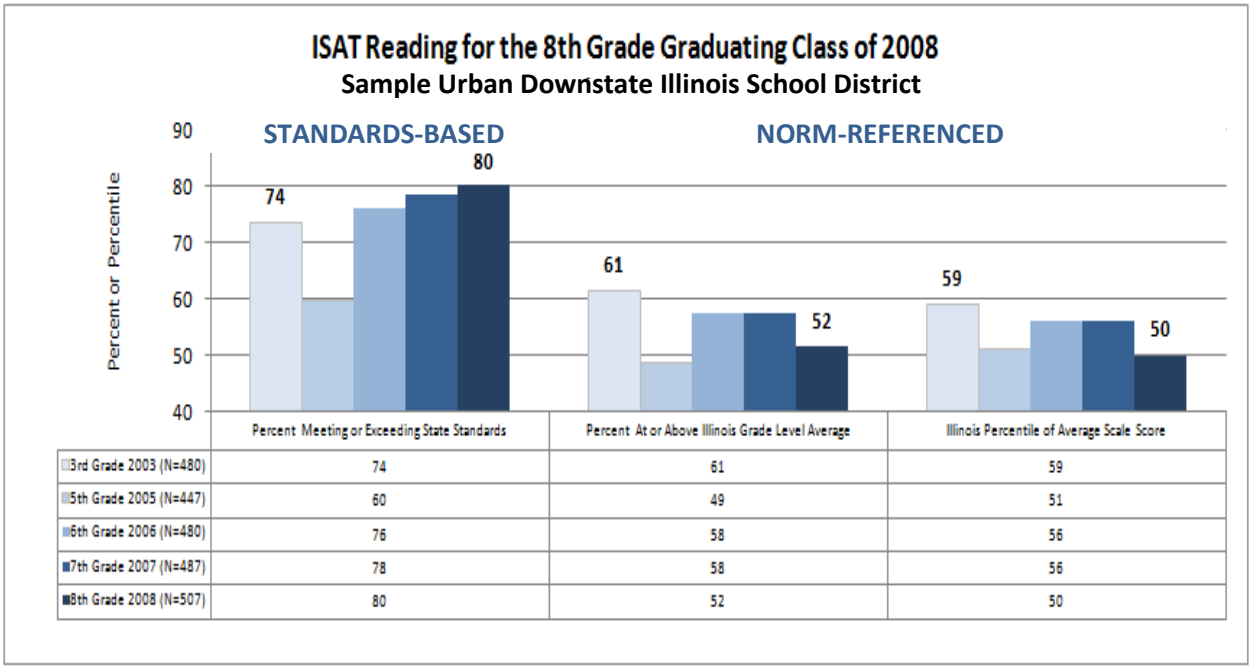


USING NORM-REFERENCED MEASURES TO TRACK VALUE-ADDED OVER TIME

"Value-added" describes the amount of progress made by an individual or group compared to the average amount of progress that all students make during the same period.

Figures 7A and 7B track ISAT reading progress for the 8th grade graduating Class of 2008 in the school district described above across five of the six years between 3rd grade (2003) and 8th grade (2008)[7]. These Figures illustrate that standards-based benchmarks produce a ***very different picture*** of overall achievement and progress than norm-referenced measures.

FIGURE 7A



	Percent Meeting or Exceeding State Standards	Percent At or Above Illinois Grade Level Average	Illinois Percentile of Average Scale Score
3rd Grade 2003 (N=480)	74	61	59
5th Grade 2005 (N=447)	60	49	51
6th Grade 2006 (N=480)	76	58	56
7th Grade 2007 (N=487)	78	58	56
8th Grade 2008 (N=507)	80	52	50

Figure 7A uses three different metrics to describe how students did during each year shown:

1) Percentage of students who met or exceeded state standards (standards-based)
2) Percentage of students who scored at or above the statewide, grade-level average (norm-referenced)
3) The Illinois percentile of the average scale score for the group[8] (norm-referenced)

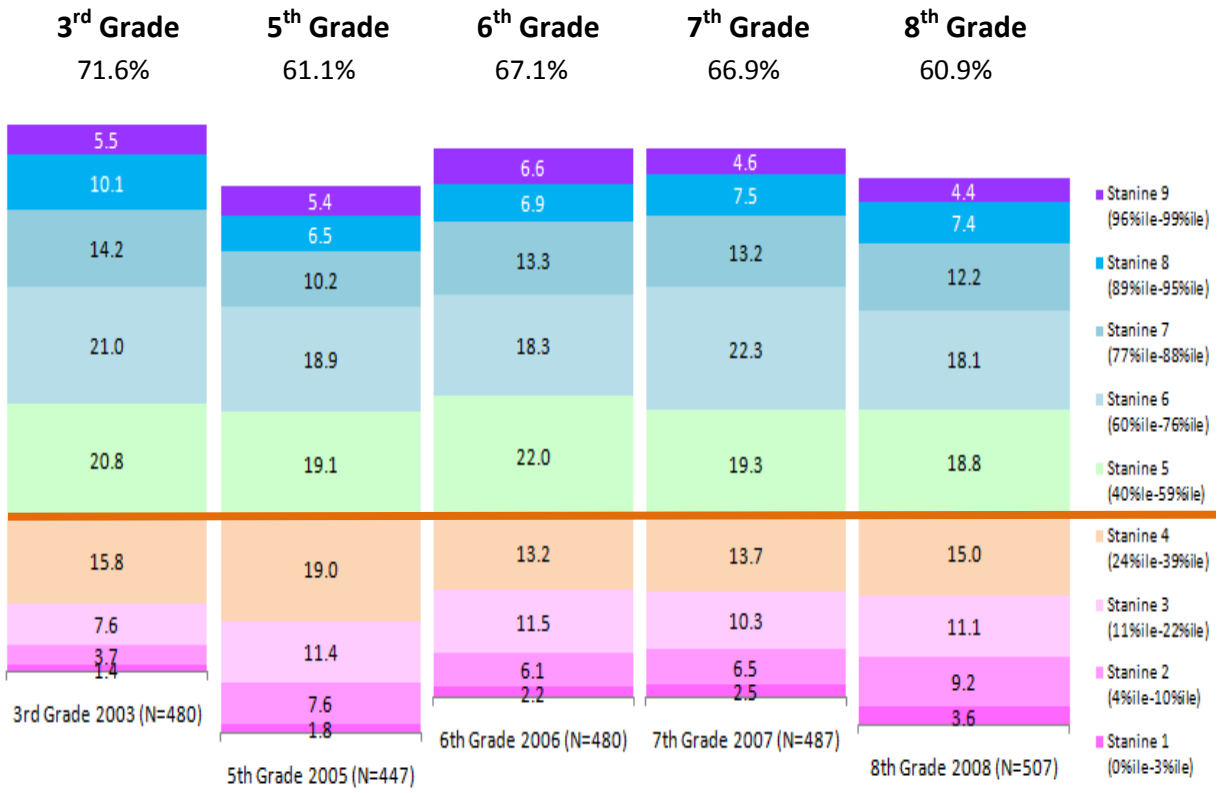
All three metrics show declining value-added between grade 3 and grade 5 and increasing value-added between grades 5 and grade 6. But between grade 6 and grade 8, the standards-based metric shows continuing growth. The norm-referenced metrics show a sharp decline between grades 7 and 8. Note also that, on average, standards-based measures report scores that are 15 to 20 points higher than the norm-referenced measures.

[7] ISAT reading tests were not administered to 4th grades before spring 2006

[8] For example, 8th graders in 2008 scored at the 50th percentile for Illinois based on average scale score of 248. This means that a score of 248 was as good or better as scores achieved by 50% of all the students statewide who took the 8th grade ISAT reading test in 2008.

Stanine distributions provide a convenient and statistically significant way to track value-added across the full spectrum of student achievement. Figure 7B uses stanine distributions to track reading progress for the 8th grade Class of 2008 in this school district.

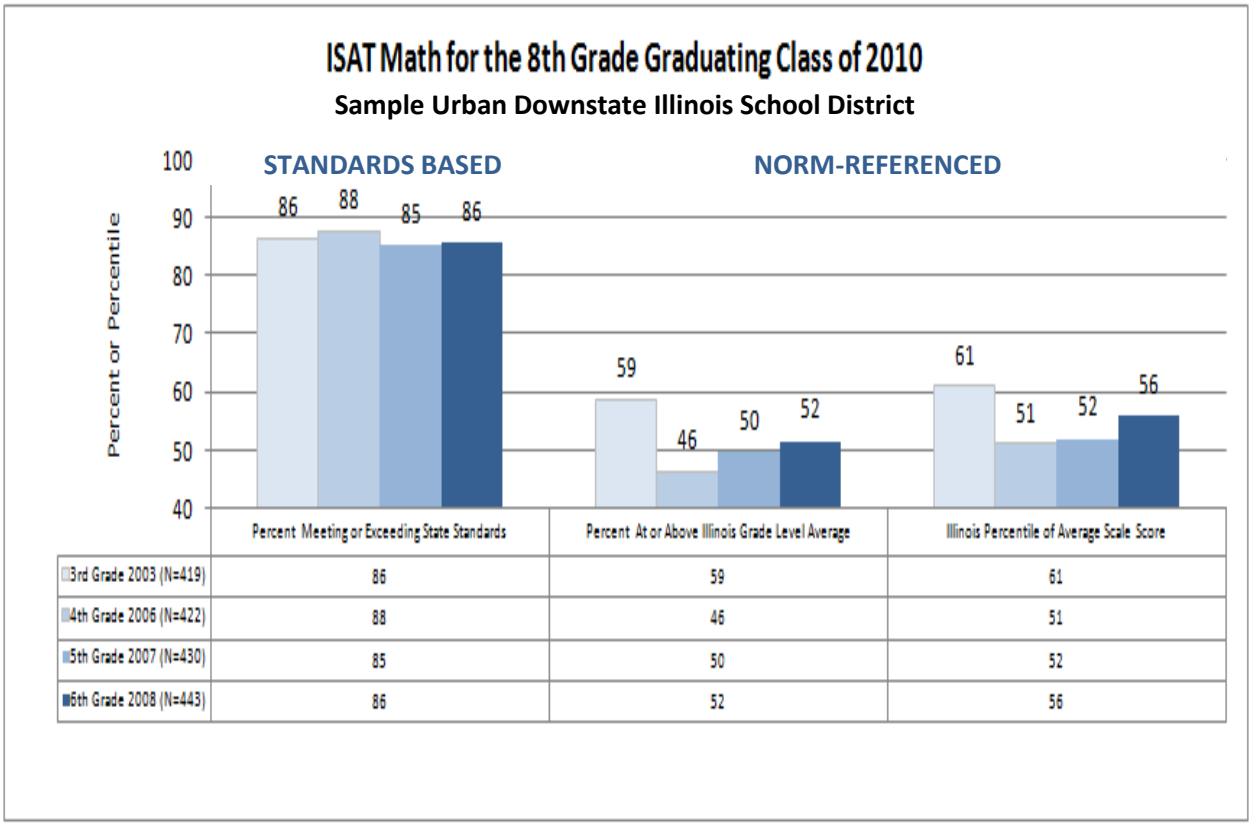
FIGURE 7B: Stanine 5 and Above



Patterns worth noting in Figures 7A and 7B include the following:

- Standards-based meet/exceed percentages shown in Figure 7A suggest that overall student achievement at grades 6, 7 and 8 improves over its 3rd grade base, but . . .
- Norm-referenced data show that the overall distribution of achievement never fully recovers from the ground it loses in the transition from primary grade 3 to intermediate grade 5.
- Standards-based meet/exceed percentages show that the portion of students who are academically at risk declines from 3rd and 8th grade to just a little over 20% of the total population, but . . .
- Figure 7B illustrates that the percentage of students who score at stanine 4 or below actually increases from 28.5% in 3rd grade to 38.9% in grade 8. The percentage of students scoring at stanine 3 or below increases from 12.7% to 23.9%.
- Figure 7B also shows that top-end achievement which predicts ACT college readiness in reading (stanine 6 and above—see page 14) declines from 50.5% in grade 3 to 41.0% in grade 5, and from 47.6% in grade 7 to 42.1% in grade 8.

FIGURE 8A



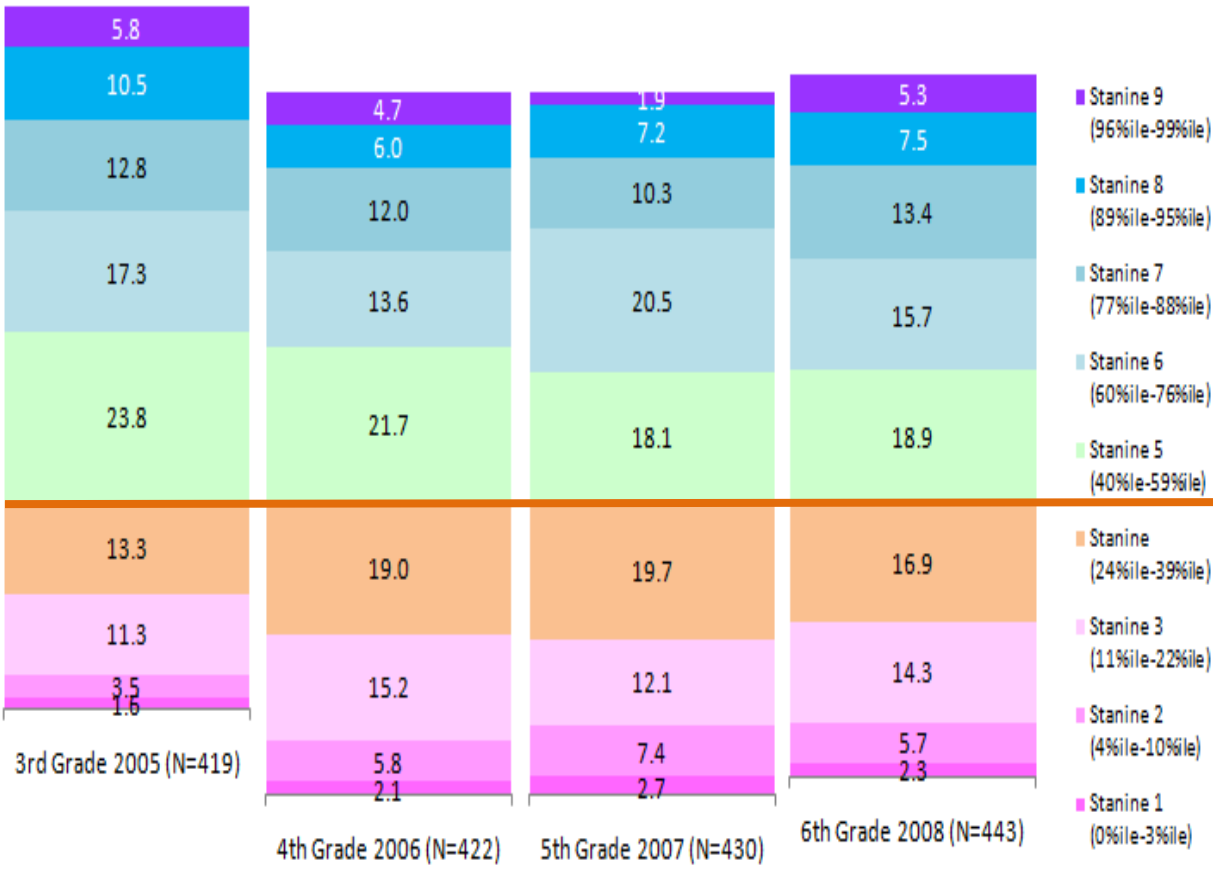
ANOTHER EXAMPLE

Figures 8A and 8B track ISAT math progress from 3rd grade (2005) through 6th grade (2008) for the 8th graduating Class of 2010. Some patterns worth noting include the following:

- The standards-based meet/exceed metric in Figure 8A indicates high and stable levels of math achievement across all of the four years shown. It also shows that a small increase in student achievement occurred as students transitioned from the districts eight, K-3 primary schools to a single intermediate school in 4th grade. But . . .
- Norm-referenced measures show that a big drop in math achievement occurred ***across the entire performance distribution*** as the group transitioned from 3rd to 4th grade.
- Norm-referenced measures show that important gains did occur between grades 4 and 6 but those gains were not enough to make up for the ground that was lost in the one-year transition from primary to intermediate school.
- Stanine distributions in Table 8B show that between grade 3 and grade 6, students scoring at stanine 4 and below increased from 29.7% to 39.2%. Students scoring at stanine 3 and below increased from 16.4% to 22.3%.
- Table 8B also shows that top-end achievement which predicts ACT college readiness (stanine 6.5[11] and above—see page 14) declined from 31% in grade 3 to 28% in grade 6.

FIGURE 8B: Stanine 5 and Above

3rd Grade	4th Grade	5th Grade	6th Grade
70.2%	58.0%	58.0%	60.8%



TIGHTENING THE RELATIONSHIP BETWEEN THE ISAT and ACT COLLEGE READINESS BENCHMARKS

The single most important job of a standards-based testing system is to match year-to-year proficiency benchmarks with the continuum of progress that students need to make to reach high-stakes outcomes by the end of grade 12.

One high-stakes outcome measure that is currently embedded in the Prairie State Achievement Exam (PSAE) is the ACT college entrance exam. Across grades and subject areas, there appears to be a consistent relationship between particular sets of ISAT scores and the ACT scores that students need to achieve to reach college benchmarks at the end of grade 11.[9]

[9] Currently, the only years for which statewide ACT college readiness scores are publicly available are 2006, 2007 and 2008 [available at http://iirc.niu.edu]

At the district level, ISAT score ranges that predict college readiness are relatively narrow. Statewide, these ranges are larger but are generally limited to a single stanine[10].

Figures 9A, 9B and 9C present examples from our sample Illinois school district that show how well ISAT stanine scores at elementary and middle school predict ACT college readiness when students reach 11th grade. A strong, common-sense implication of these data is that preparation for ACT college readiness begins early and that improved achievement at elementary and middle schools is a necessary condition for meaningful gains on the ACT.

FIGURE 9A: ACT College Readiness Predictions from ISAT READING Scores at Grade 5 and 8

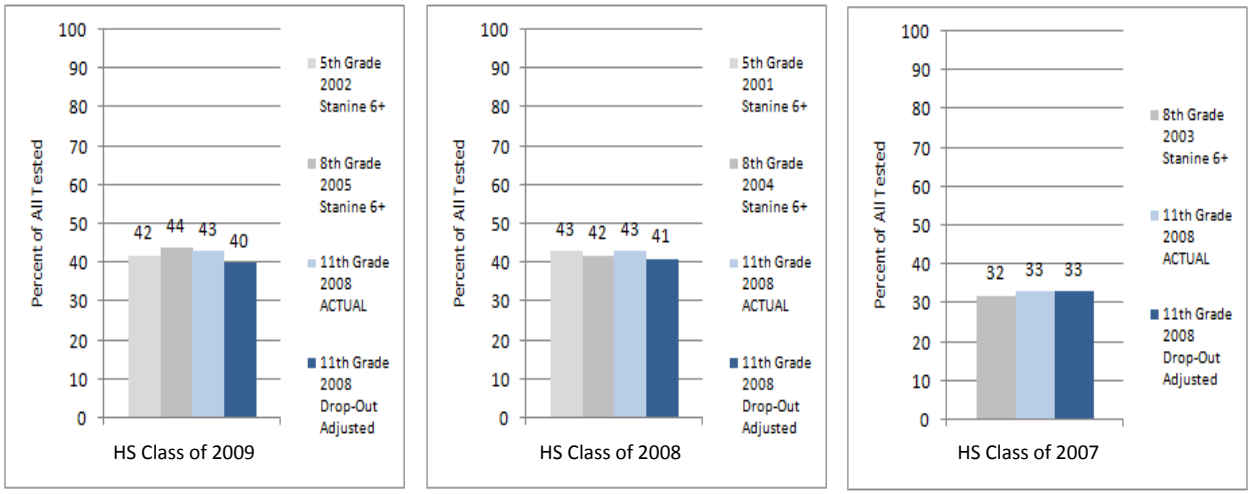
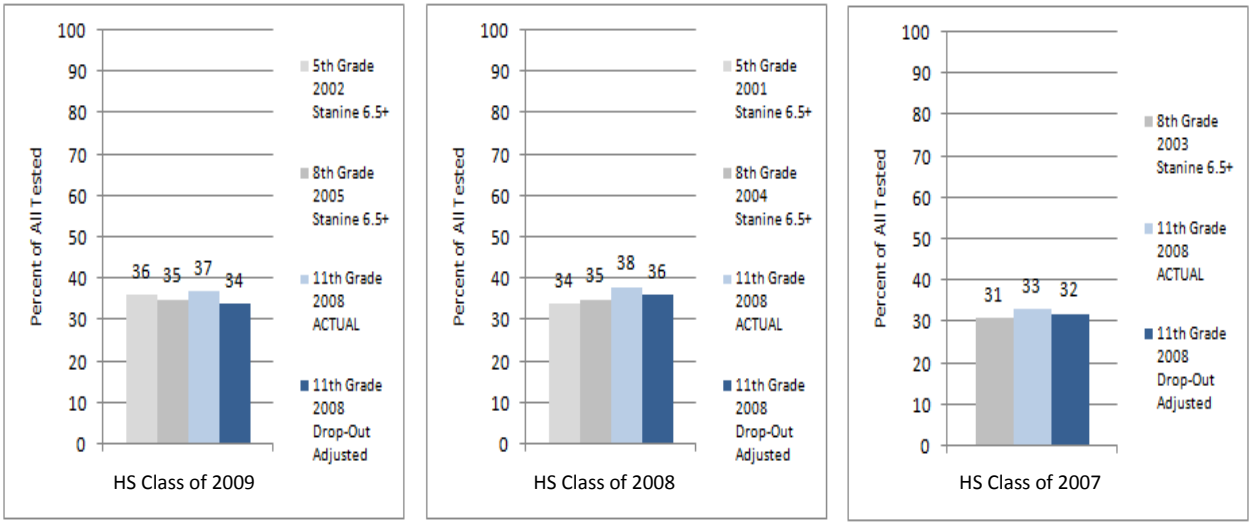


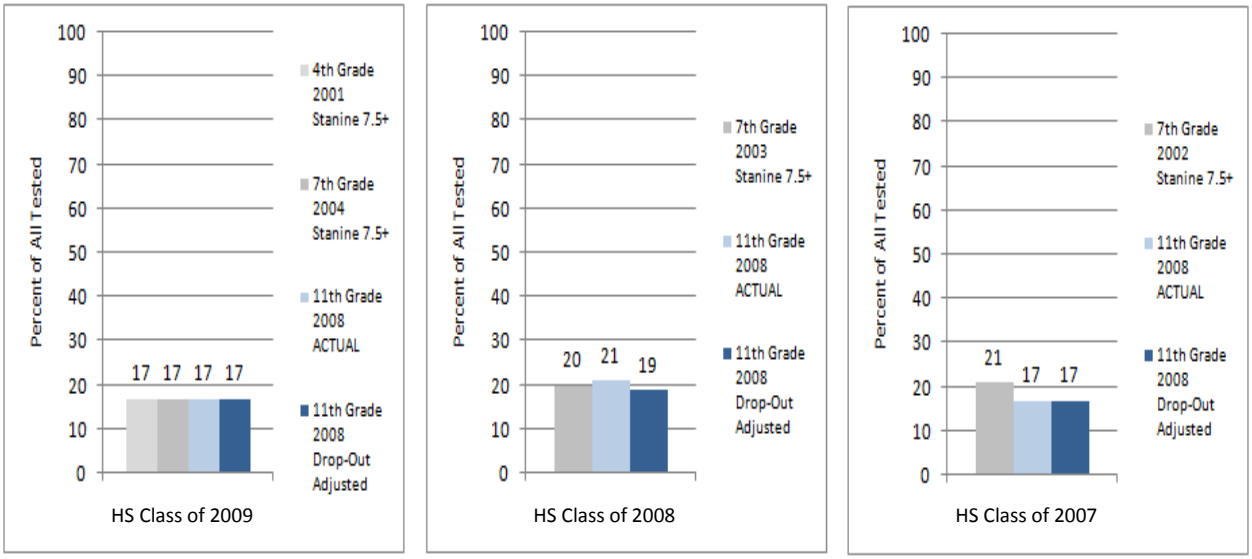
FIGURE 9B: ACT College Readiness Predictions from ISAT MATH[11] Scores in Grades 5 and 8



[10] To date, bottom-end ISAT ranges which appear to be good predictors of college readiness are: ACT Reading-- stanine 6; ACT Math--the range between mid-stanine 6 and mid-stanine 7[11]; ACT Science Reasoning-- the range between mid-stanine 7 and mid-stanine 8.[11]

[11] The method used to calculate ISAT college readiness estimates for math in Figure 9B is to take one half of the total number of students scoring at stanines 6 and 7 and add that to the number of students scoring at stanines 8 and 9. In Figure 9C, one half of the total number of students scoring at stanines 7 and 8 is added to the number of students scoring at stanine 9.

FIGURE 9C: ACT College Readiness Predictions from ISAT SCIENCE[11] Scores at Grades 4 and 7



At each grade level, the statewide range of scale scores for each stanine is relatively stable from one year to the next.[12] Absent more specific local analysis, the approach illustrated above gives every Illinois district the ability to make meaningful estimates of the portion of students at each ISAT grade level who are on-track to meet ACT college readiness standards at the end of 11th grade.

ISAT Scale Scores and ACT College Readiness

The specific ISAT scores that predict ACT college readiness vary from district to district depending on the demographics of the student population and the academic expectations of high schools students attend. For example, in 2005, an 8th grader who entered high school in west suburban Naperville District 203 with ISAT scores of 250[13] in reading and 280[13] in math was typically on-track[14] to meet college readiness benchmarks in 2008. By contrast, the average 8th grader in the Chicago Public Schools needed a reading score in the low 260s and a math score in the low 290s to be on-track.

Despite these differences, it is still possible to make general predictions about ACT college readiness. That's because most district-to-district differences fall within a single stanine range, or ½ standard deviation unit (see footnotes 10 and 11). In general, the percentage of students who score in that range or above are likely[14] to meet college readiness standards in 11th grade.

[12] A major exception to this stability occurred when the ISAT was revised in 2006. Across all grades and subjects tested, scale score ranges for most stanines made an abrupt upward shift. For example, between 2003 and 2005, the range of reading scale scores at 5th grade stanine 5 was 214 to 230. Since 2006, the range of reading scale scores at 5th grade stanine 5 has been 222-234. Since cut scores stayed the same for all but one of the ISAT proficiency benchmarks, the practical effect of these shifts was that it became substantially easier to meet state standards after 2006 than it was before 2006.

[13] Scores shown are 2006 conversions of actual scores from the pre-2006 ISAT metric

[14] On-track is defined here as having a 50% or better probability of meeting college readiness benchmarks in 11th grade.

SUMMARY

Sound decision making begins with good metrics. At best, current ISAT reporting strategies present a confusing picture of student learning, school effectiveness and value-added over time. At worst, they flat out misrepresent what is actually going on.

Standards-based curriculum reform began with the promise of bringing greater depth and clarity to what we expect students to know and be able to do. The purpose of this brief is to underscore the importance of bringing greater depth and clarity to the ways we report student progress.

Elementary proficiency standards can be readily back-mapped from ***existing*** high stakes metrics for college readiness (ACT) and workplace preparedness (Work Keys). This work needs to be supported by wider application of norm-referenced reporting strategies that provide students, parents, school personnel and the community at large with more accurate and usable information about student progress over time.