TITLE: Taking Stock: How Standardized Test Reports Let Us Down Under No Child Left Behind… And How We Can Fix What is Wrong

AUTHORS: Zavitkovsky, P., Roarty, D., & Swanson, J.

PUBLICATION DATE: Spring 2016

ABSTRACT:

This study clarifies achievement trends that occurred under NCLB and explains why NCLB reporting practices made those trends so hard to see. It concludes by describing important contributions that new PARCC exams can make and warns of new reporting problems that threaten to squander those contributions before they see the light of day.

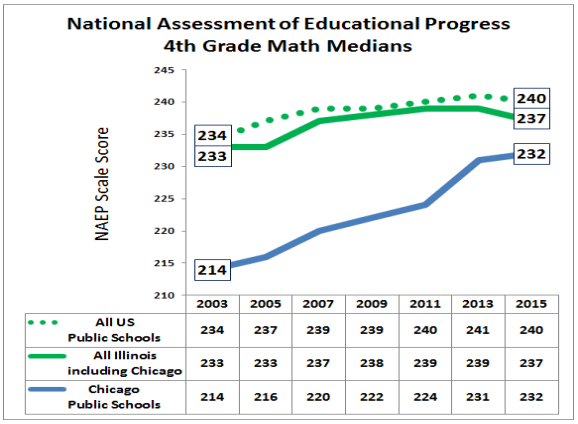
Taking Stock

Spring 2016

Paul Zavitkovsky
Denis Roarty
Jason Swanson

Center for Urban Education Leadership
University of Illinois at Chicago

HOW STANDARDIZED TEST REPORTS LET US DOWN UNDER NO CHILD LEFT BEHIND . . . AND HOW WE CAN FIX WHAT'S WRONG



	2003	2005	2007	2009	2011	2013	2015
All US Public Schools	234	237	239	239	240	241	240
All Illinois including Chicago	233	233	237	238	239	239	237
Chicago Public Schools	214	216	220	222	224	231	232

UIC
Making Good on the Promise of Public Education

CENTER for URBAN EDUCATION LEADERSHIP

Executive Summary

The promise of standards-based assessment under No Child Left Behind (NCLB) was that it would make test information more meaningful and useful for parents, educators and the public at large. But arbitrary grading and shoddy reporting practices destroyed the credibility of the Illinois Standards Achievement Test (ISAT) and created deep confusion about what standardized tests actually assess. In the end, reporting practices under NCLB made it harder than ever . . . even for insiders . . . to get a clear picture of what was actually going on.

This study clarifies achievement trends that occurred under NCLB, and explains why NCLB reporting practices made those trends so hard to see. It concludes by describing important contributions that new PARCC exams can make, and warns of new reporting problems that threaten to squander those contributions before they see the light of day.

Part 1 describes achievement trends in Illinois' elementary and middle school test population from 2001 through 2015:

- *Section 1 documents flattening achievement statewide and rising achievement in Chicago under NCLB, and illustrates why common explanations for both do not hold water.*
- *Section 2 describes regional differences in how achievement shifted under NCLB*
- *Section 3 provides evidence that, on average, the transition to middle school is having a negative impact on the achievement of early adolescents outside of Chicago*
- *Section 4 describes changes in third grade achievement in and out of Chicago among Illinois' three largest racial groups.*

Key findings elaborated in Part 1 include the following:

- *During most of the NCLB era, achievement growth in Chicago exceeded growth outside of Chicago among all racial sub-groups. Within each sub-group, achievement levels in Chicago now match or exceed those of comparable sub-groups in the rest of Illinois at all grade levels tested*
- *Regional gains in composite reading and math achievement at grades 3-8 were strongest in Chicago and the 6-county metropolitan area surrounding Chicago, and weakest in central and southern Illinois*
- *In Chicago, average growth over time proceeds fairly evenly from grade three through eight. Average achievement in the rest of Illinois slows markedly as students transition from intermediate grades 3-5 to middle school grades 6-8*

- *Statewide, the student populations that benefited least from improvements in instructional effectiveness under NCLB were black and white students from low-income households*
- *Recent stagnation of overall, statewide achievement has mostly resulted from decreasing enrollments and flattening achievement among white students from middle and upper income households*
- *Achievement growth among Latino students not identified as English Language Learners (ELL) consistently outpaced that of black and white students. Failure to disaggregate students temporarily classified as ELL from Latino achievement reports masked and under-reported actual growth rates.*

Part 2 explores the alternative universe of reporting practices that distorted how test results were communicated under NCLB:

- *Section 5 shows how oversimplified reporting practices reinforced old stereotypes and missed important changes in achievement gaps that are commonly associated with race, family income and English language proficiency*
- *Section 6 describes how arbitrary "standard setting" obscured the close match between ISAT results and results of more highly regarded tests like the Measures of Academic Progress (MAP), National Assessment of Educational Progress (NAEP), ACT and, more recently, PARCC*
- *Section 7 looks more closely at what standardized test items actually assess and examines how very different tests end up producing close-to-identical results*
- *Section 8 explains why common NCLB diagnostic reports like "content strands," "item analysis" and "power standards" are mostly just packaging gimmicks that misrepresent and under-report what standardized tests actually assess*

Part 3 describes why PARCC assessments are better equipped than their predecessors to report meaningful, standards-based information, but warns of early evidence that this information may once again get squandered by a new generation of deeply inadequate reporting practices.

CONTENTS

PART 1

RAISING THE PROFILE OF STATEWIDE ACHIEVEMENT TRENDS

Under the radar, evidence has been accumulating for close to a decade that standardized achievement is flattening statewide while achievement in Chicago has been steadily increasing.

Created in 1969, the National Assessment of Educational Progress (NAEP) is widely recognized by researchers, educators, policy makers and legislators as the "gold standard" for standards-based assessment in the United States. In October 2015, results from the NAEP generated a little more attention than usual in the national media. For the first time in 25 years, national averages dropped on three of the four tests reported. And average growth in the country's largest cities flattened after exceeding national growth rates for more than a decade.

For the most part, Illinois' major newspapers covered NAEP results with a single release from the Associated Press that focused on nationwide results. One exception was the Chicago *Tribune.* It used a Sunday editorial to congratulate Chicago students and teachers for bucking national trends and making stronger gains than statewide averages.

Springfield

The State Journal-Register

Suburban Chicagoland

Daily Herald

October 28, 2015

Math, reading scores slip for nation's school kids

Jennifer C. Kerr, The Associated Press

How did Illinois fare?

Grade 4, math—37% at or above proficient
Grade 4, reading—35% at or above proficient
Grade 8, math—32% at or above proficient
Grade 8, reading—35% at or above *proficient*

Washington—It's a not-so-rosy report card the nation's schoolchildren. Math scores slipped for fourth and eighth graders of the last two years and reading were not much better, flat for fourth graders and lower for eighth graders, according to the 2015 Nation's Report Card.

Peoria

JournalStar

October 28, 2015

School report shows dip in math scores for 4th and 8th grade; reading slips for 8th, flat for 4th

By Jennifer C. Kerr of the Associated Press

Washington—It's a not-so-rosy report card the nation's schoolchildren. Math scores slipped for fourth and eighth graders of the last two years and reading were not much better . . .

Chicago Tribune

November 1, 2015

CPS makes the grade … but the nation's schools slip

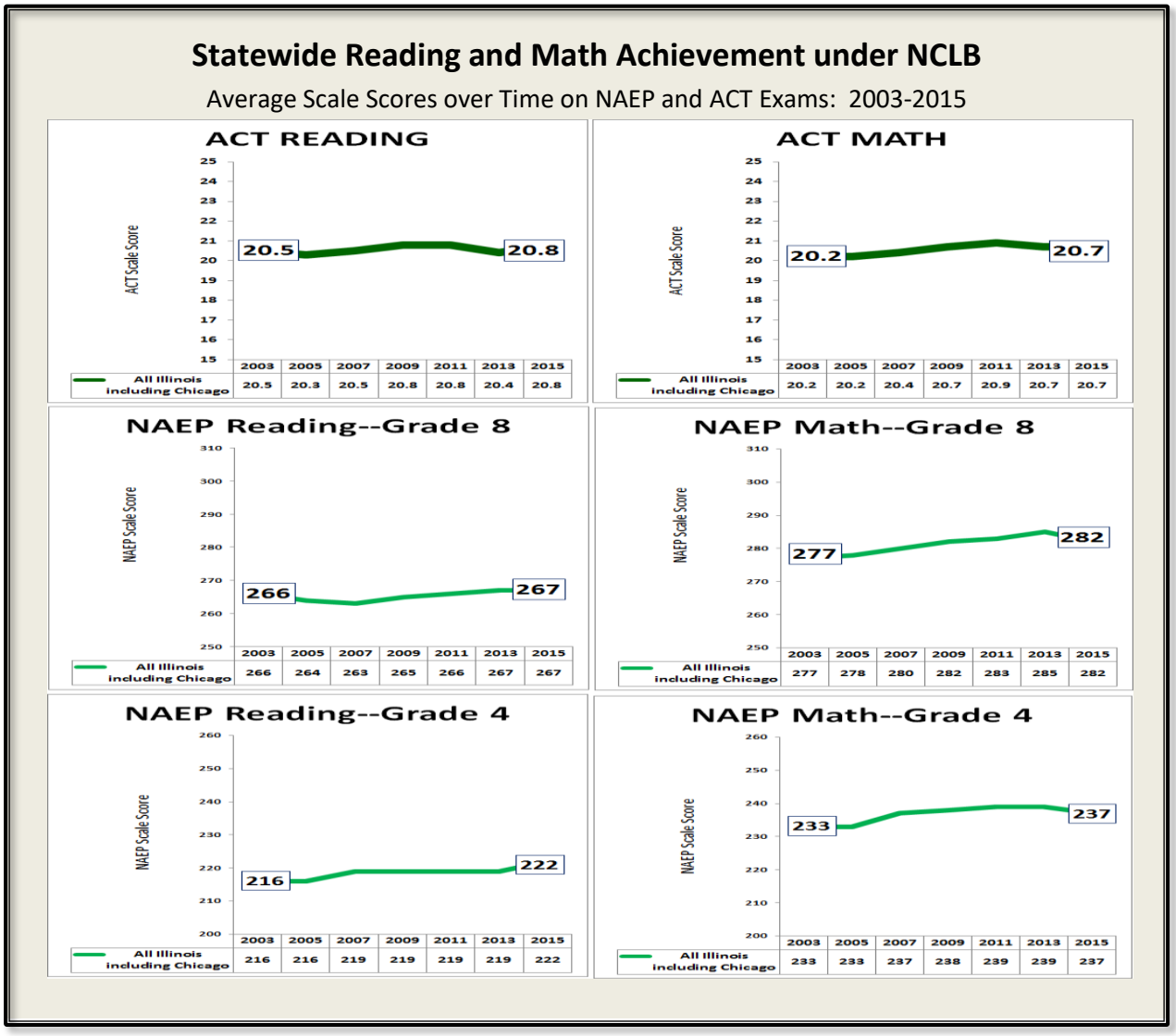
The Nation's Report Card dished out encouraging news for Chicago Public Schools last week. CPS fourth- and eighth-graders are now performing on par or nearly so with many of their peers in math and reading on the benchmark national assessment test.

Overall the nation's students are still behind much of the rest of the industrialized world in academics. And it's getting worse: Some U.S. scores slipped on the test, known as the National Assessment of Educational Progress, or NAEP, for the first time for the first time since 1990. But at least Chicago's children are catching up with the national pack.

Missing from most public descriptions of 2015 NAEP results was an unsettling fact. Growth in statewide achievement was statistically flat in 2015 . . . just like it was in 2013, 2011, 2007 and 2005. Statistically-flat means that small changes in statewide scoring between 2003 and 2015 could easily have been caused by normal testing variations and random errors.

Part 1 of *Taking Stock* takes a closer look at the factors that have contributed to flattening achievement in Illinois:

- Section 1 draws on achievement trends in Chicago and the six-county area surrounding Chicago to illustrate why common explanations do little to explain what has actually been going on.
- Section 2 describes regional difference in achievement trends that occurred in Illinois during the NCLB era
- Section 3 shows evidence that the transition to middle school is having a negative impact on the achievement of many early adolescents outside of Chicago
- Section 4 describes changes in third grade achievement in and out of Chicago among Illinois' three largest racial groups



SECTION 1

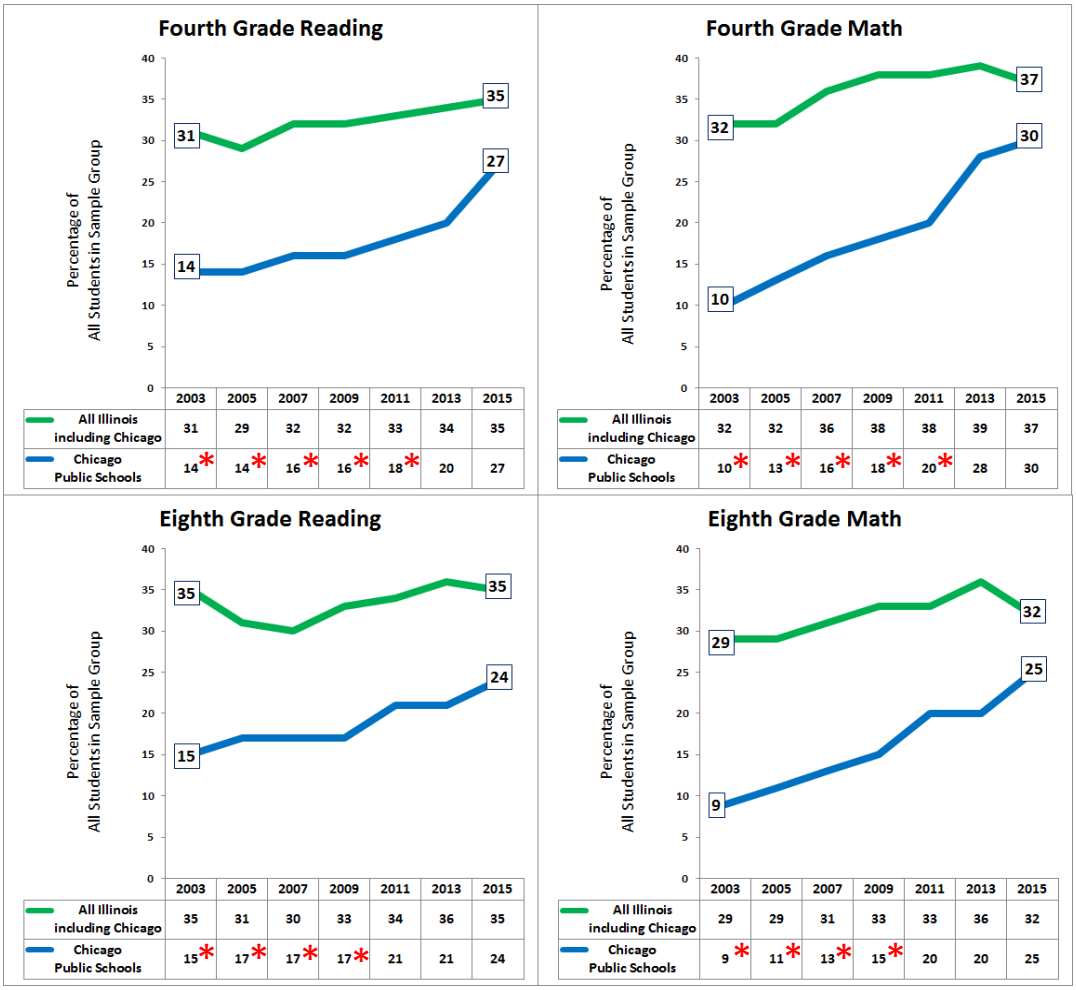
Statewide Achievement in Illinois: Statistically Flat since 2003

Growth in Illinois achievement was statistically flat in 2015 . . . for the 12th year in a row. More disturbing still, NAEP results in 2015 offered further evidence that the only thing keeping statewide trends from outright decline was sustained growth in Chicago, which accounts for close to 20% of all statewide scoring. This was particularly true of fourth grade scores which are strong predictors of future achievement in middle school and high school.

Figure 1.1 shows the percentage of students in Chicago and statewide who scored "proficient or advanced" on fourth and eighth grade NAEP exams between 2003 and 2015. The solid green lines show statewide trends that include Chicago. The solid blue lines show trends for Chicago alone. Red asterisks in the tables below each chart identify results in earlier years that were significantly lower than those posted in 2015.

Figure 1.1

Percentage of Students Scoring Proficient or Advanced on the NAEP: 2003-2015



*Significantly lower than 2015 (p=<0.05)

Source: National Center for Educational Statistics http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx

There are a number of common explanations for why lower-achieving populations like those in Chicago are learning at a faster rate than higher-achieving populations statewide. But none of them provide satisfactory answers for what's been going on.

Explanation #1: High NAEP Cut Scores Under-report Statewide Achievement

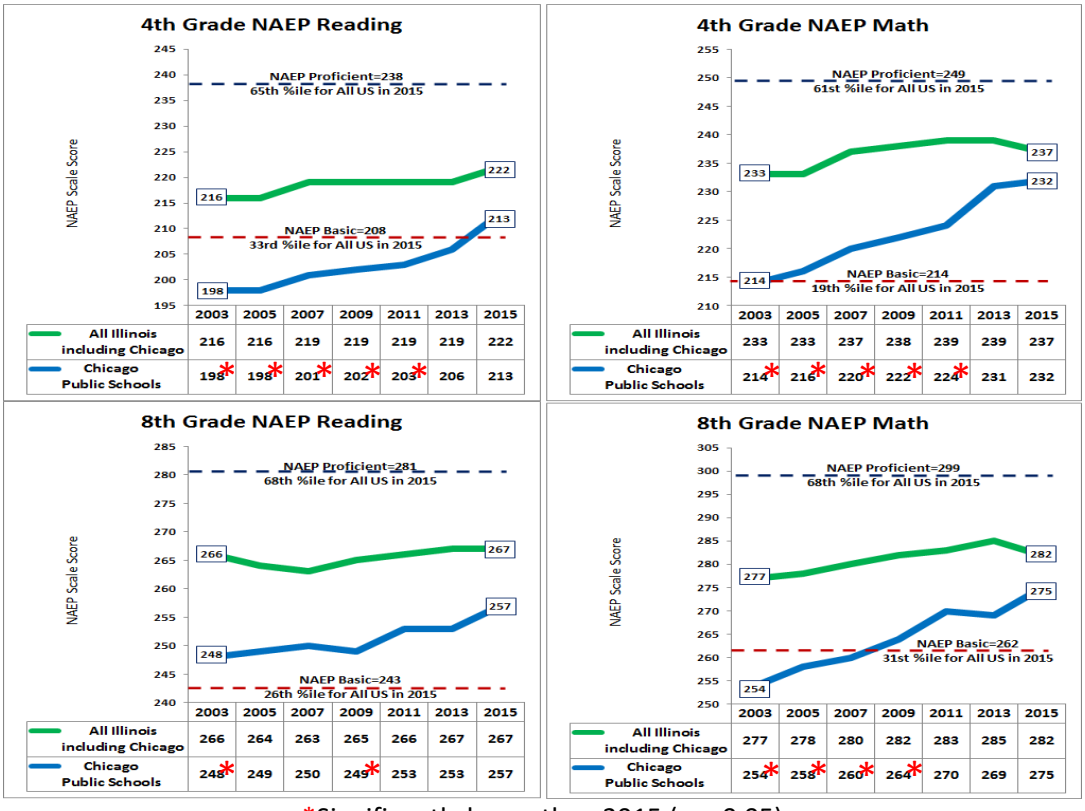
Cut scores are the locations on standardized test scales that policy makers use to define different levels of academic competence. They create the basis for grading standardized achievement in much the same way teachers use less technical criteria to distinguish As, Bs and Cs on conventional report cards.

Cut scores on the NAEP have a national reputation for being rigorous and demanding. In Figure 1.2, the blue-dashed lines at the top of each chart mark the boundary between "basic" and "proficient" on 4th and 8th grade NAEP exams. In 4th grade reading, for example, the cut score for proficiency is 238; in 4th grade math, it is 249.

One possible explanation for flat statewide achievement on the NAEP is that NAEP cut scores for proficiency have been set too high to capture changes that may be occurring among average achievers. To test this explanation, Figure 1.2 uses median scores rather than percentages of students scoring at or above proficient to represent achievement on the NAEP. Median scores describe the achievement of students who score right in the middle of each year's achievement range. What Figure 1.2 shows is that scores in the middle of statewide distributions flattened in exactly the same way they did for higher-achieving students who scored proficient and above.

Figure 1.2

Median Scores for NAEP Reading and Math in Grades 4 and 8: 2003-2015



4th Grade NAEP Reading	2003	2005	2007	2009	2011	2013	2015
All Illinois including Chicago	216	216	219	219	219	219	222
Chicago Public Schools	198*	198*	201*	202*	203*	206	213

4th Grade NAEP Math	2003	2005	2007	2009	2011	2013	2015
All Illinois including Chicago	233	233	237	238	239	239	237
Chicago Public Schools	214*	216*	220*	222*	224*	231	232

8th Grade NAEP Reading	2003	2005	2007	2009	2011	2013	2015
All Illinois including Chicago	266	264	263	265	266	267	267
Chicago Public Schools	248*	249	250	249*	253	253	257

8th Grade NAEP Math	2003	2005	2007	2009	2011	2013	2015
All Illinois including Chicago	277	278	280	282	283	285	282
Chicago Public Schools	254*	258*	260*	264*	270	269	275

*Significantly lower than 2015 (p=<0.05)

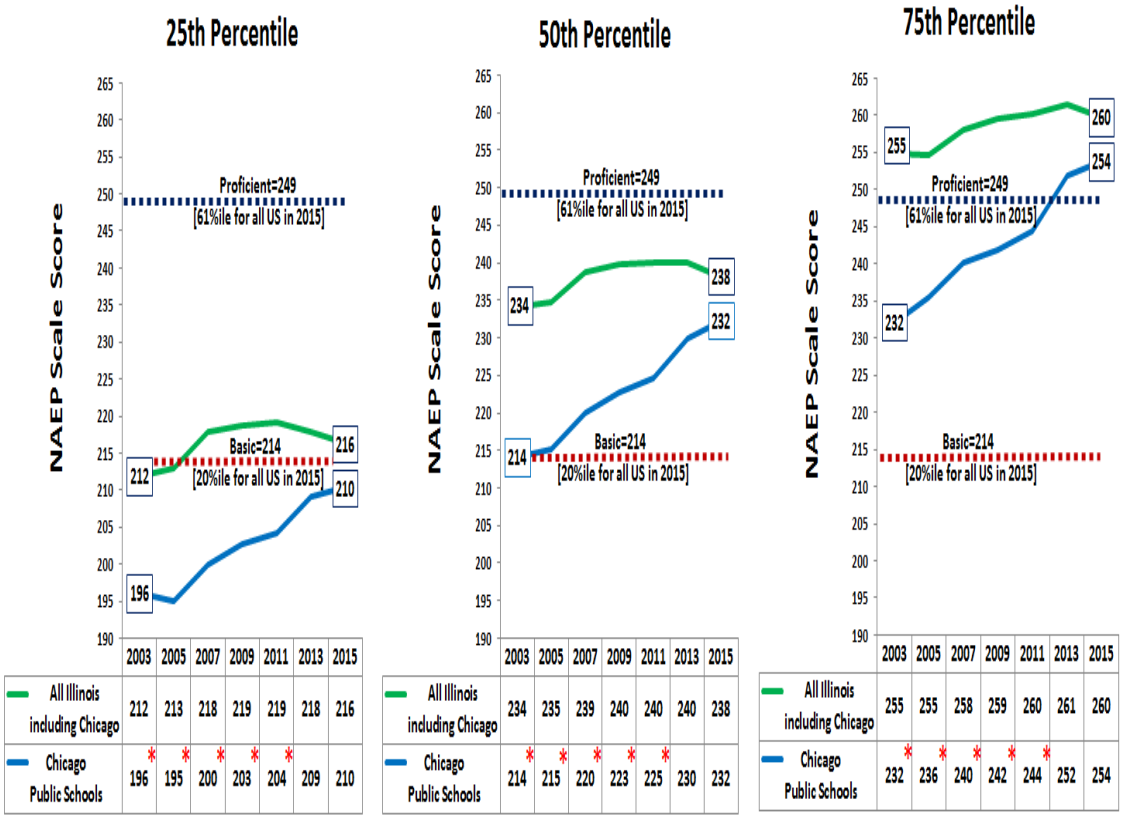
Explanation #2: It is Easier to Make Gains with Lower-Achieving Students

Another explanation for why Chicago scores have grown while statewide scores flattened is that gains might somehow be easier make when initial achievement levels are low. This explanation suggests that statewide scores may be "topping out" at middle and higher levels of the achievement spectrum while Chicago scores, which started at lower levels, had more room to grow before the climb became more difficult.

Figure 1.3 tests this explanation by showing changes over time in median scores at the 25th, 50th and 75th percentiles of Chicago and All Illinois scoring distributions. In the chart on the left, changes at the 25th percentile show gains made by lower-achieving students. In the chart on the right, changes at the 75th percentile show gains made by higher-achieving students. If growth is easier to obtain among lower-achieving students, growth rates at the 25th percentile should be substantially higher than growth rates at the 50th and 75th percentiles.

If anything, Figure 1.3 points to the opposite conclusion. It shows that long-term gains in Chicago grew larger as achievement levels rose . . . from 14 points at the 25th percentile, to 18 points at the 50th percentile, to 22 points at the 75th percentile. Meanwhile, gains for All Illinois including Chicago were only 4 to 5 points at each level. Other NAEP results showed similar patterns.

Figure 1.3
4th Grade Math Medians at the 25th 50th and 75th Percentile of Chicago and All Illinois Scoring Distributions



*Significantly lower than 2015 (p=<0.05)

Explanation #3: Increases in Poverty Account for Flattening Achievement

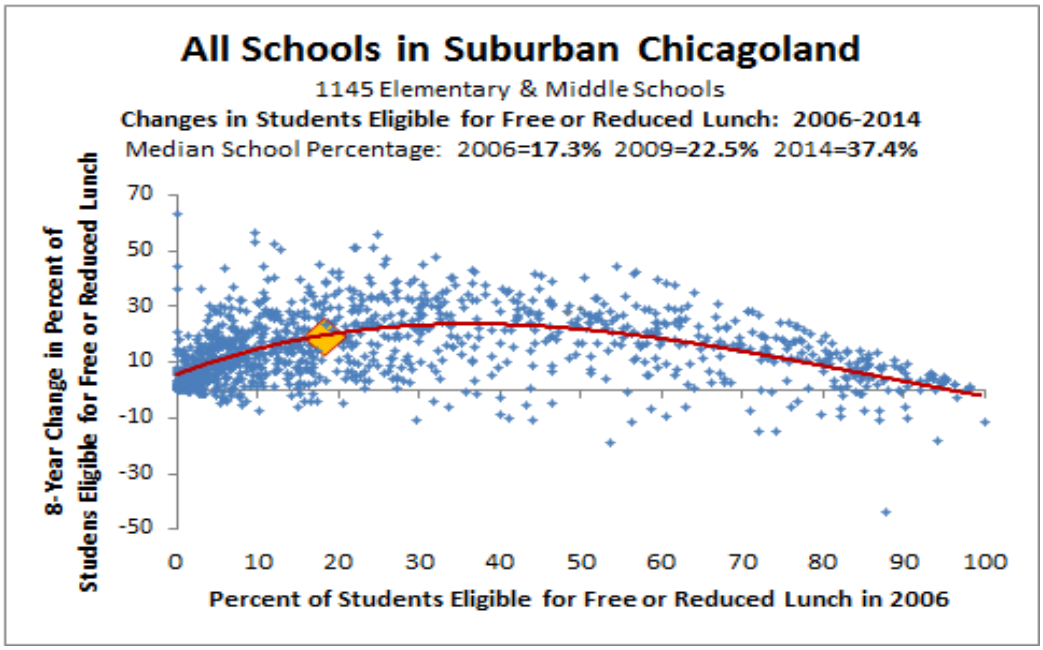
The most common explanation for flattening achievement statewide is that the percentage of Illinois students who come from low-income households has grown steadily throughout the NCLB era. In 2001, 37% of the students tested in Illinois were eligible for free or reduced lunch. By 2014, that percentage had increased to 52%.

One way to test this explanation is to track the connection between achievement and low-income enrollments in suburban Chicagoland. Suburban Chicagoland is the six-county region in northeast Illinois that surrounds (but does not include) the City of Chicago. It includes all of suburban Cook, DuPage, Kane, Lake, McHenry and Will counties and accounts for close to 50% of all students tested statewide.

Between 2006 and 2014, low-income enrollments in suburban Chicagoland grew at a faster rate than any other region in the state, more than doubling from a median of 17.3% in 2006 to a median of 37.4 % in 2014. The scatterplot in Figure 1.4 illustrates how this change was distributed across all 1,145 elementary and middle schools in the suburban Chicagoland region.

- Each blue dot in Figure 1.4 represents an individual school
- Each dot marks the coordinate between the percentage of students who were eligible for free or reduced lunch in 2006 (horizontal axis) and the change in free/reduced eligibility that occurred between 2006 and 2014 (vertical axis).
- The red trend line shows changes that were most typical of each starting point in 2006
- The orange diamond shows the individual school that was most typical of all schools in the region between 2006 and 2014.

Figure 1.4

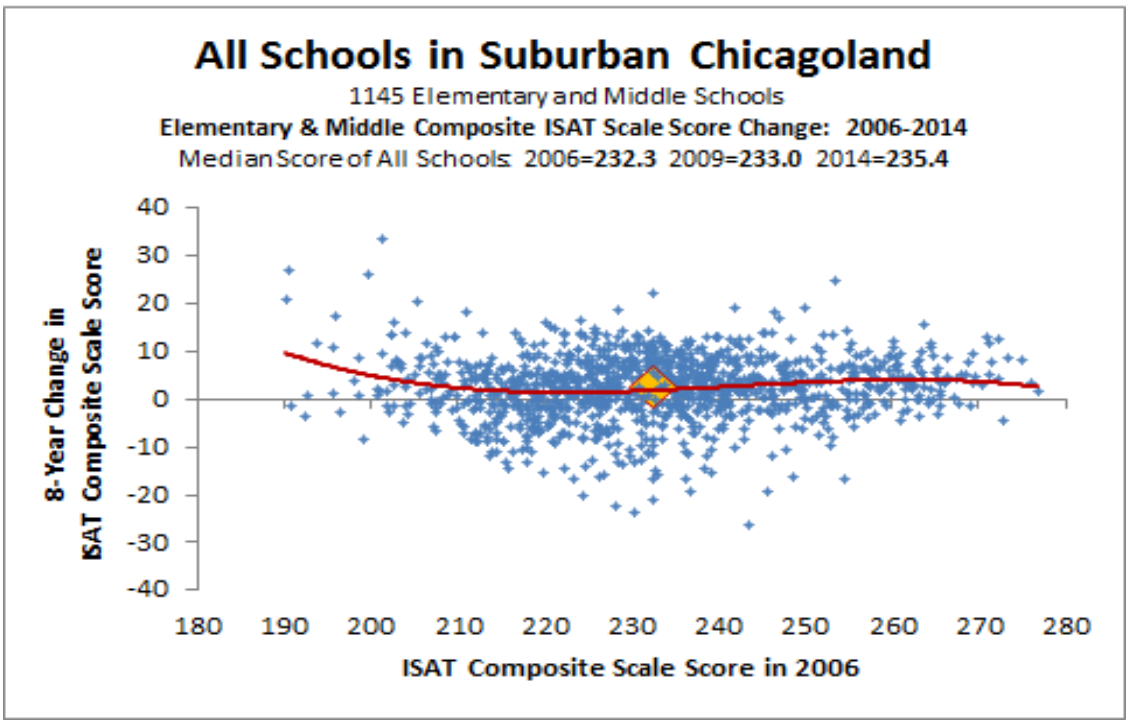


Source: Illinois State Board of Education ftp://ftp.isbe.net/SchoolReportCard/

If flattening achievement is an inevitable consequence of increasing low-income enrollments, recent achievement in suburban Chicagoland would surely reflect that impact. But actual changes in achievement point in the opposite direction.

Figure 1.4 above shows that low-income enrollments increased substantially at all but a handful of schools between 2006 and 2014. But Figure 1.5 below shows that composite reading, math and science achievement in grades three through eight actually increased at two thirds of the schools in the region. It also shows that schools where achievement growth occurred were fairly evenly distributed across the full range of school achievement levels.

Figure 1.5

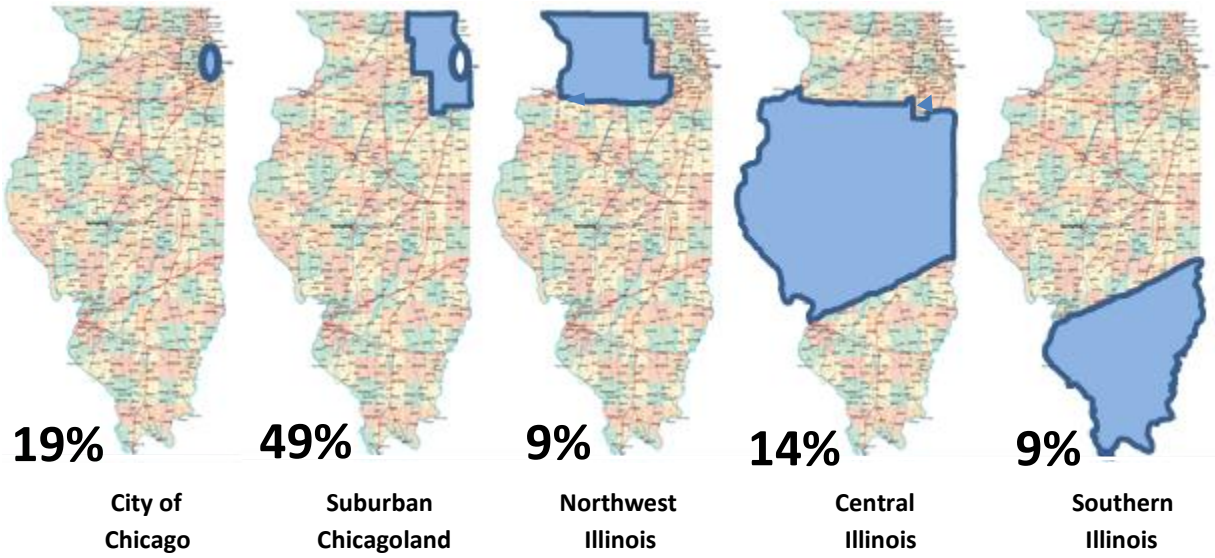


The data in Section1 illustrate that flattening achievement outside of Chicago defies simple, statewide explanations. Section 2 elaborates on this theme by describing how changes in achievement varied across different regions of the state during the NCB era.

SECTION 2

Regional Differences in Demographics and Achievement under NCLB

Regional gains in composite reading math and science achievement under NCLB were strongest in Chicago and the 6-county metropolitan area surrounding Chicago, and weakest in central and southern Illinois



This section summarizes changes in achievement which occurred between 2006 and 2014 in five geographic regions:

- City of Chicago: Serving **19%** of all students tested statewide in 2014
- Suburban Chicagoland (DuPage, Kane, Lake, McHenry, Will and suburban Cook counties): Serving **49%** of all students tested statewide in 2014
- Northwest Illinois (schools north of Interstate 80 other than those located in the City of Chicago or Suburban Chicagoland): Serving **9%** of all students tested statewide in 2014
- Central Illinois (schools located between Interstate 80 and Interstate 70): Serving **14%** of all students tested statewide in 2014
- Southern Illinois (schools located south of Interstate 70): Serving **9%** of all students tested statewide in 2014

Like the examples presented in Section 1, scatterplots used in this section use blue dots to represent individual schools that are located in each region. For example, Figure 2.1 shows changes in composite ISAT scale scores between 2006 and 2014 at the 542 elementary and middle schools in the central Illinois region.

In Figure 2.1

- Blue dots mark the coordinate between each school's average composite score in 2006 (horizontal axis) and the change in that score between 2006 and 2014 (vertical axis).
- The red trend line shows changes that were most typical of each starting point in 2006. The trend line in Figure 2.1 shows that, on average, most schools in central Illinois saw little or no change in composite scores between 2006 and 2014. The slight upward bowing at each end of the trend line means that the lowest and highest scoring schools in 2006 were slightly more likely to show positive growth than schools scoring closer to the middle of the pack.
- The orange diamond shows the individual school that was most typical of all schools in central Illinois. In both 2006 and 2014, its average composite score was about 230.

Figure 2.1
Composite Scores at Typical Elementary and Middle Schools in Central Illinois Were Mostly Unchanged between 2006 and 2014
8-Year Changes in ISAT Composite Scores at Elementary and Middle Schools in Central Illinois

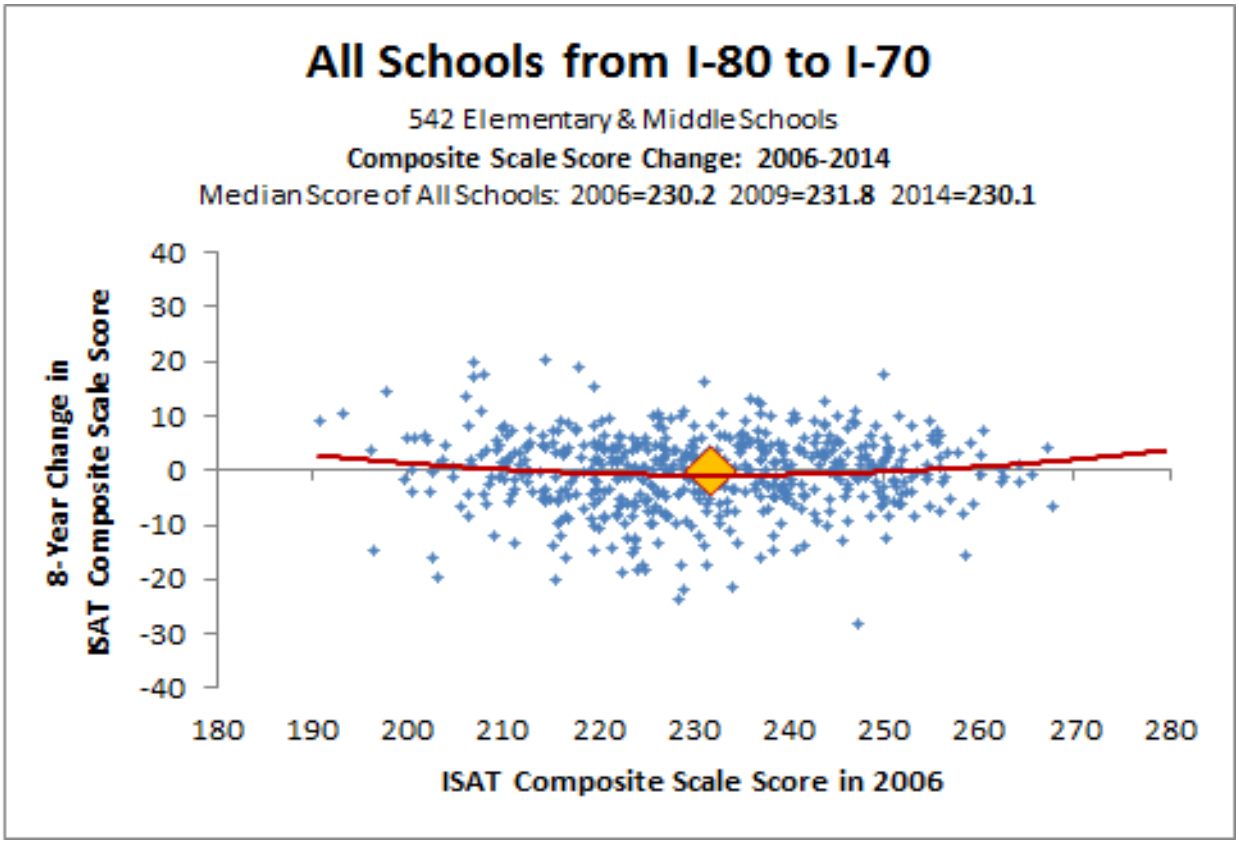
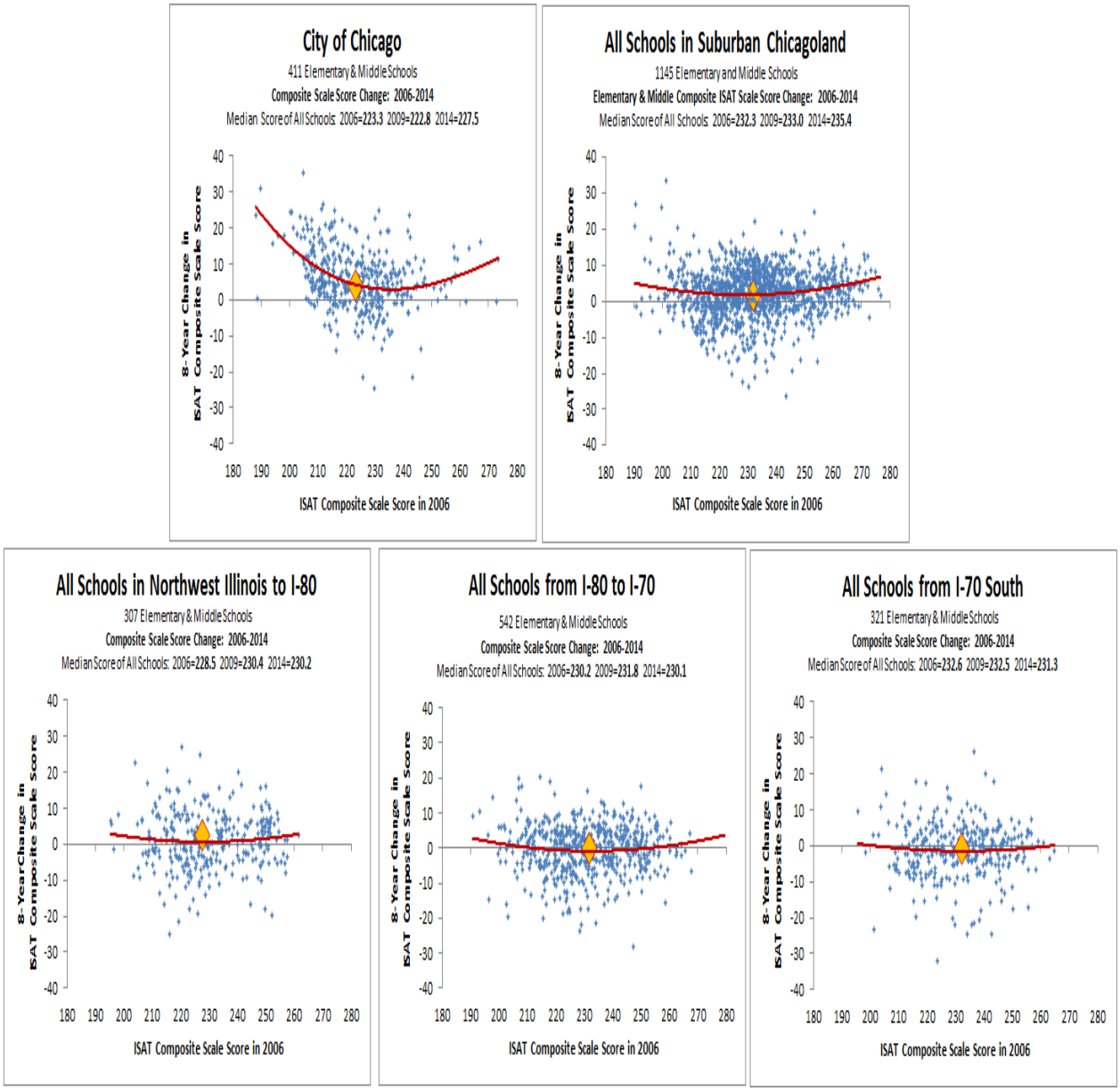


Figure 2.2 compares 8-year changes in composite scores across all five regions of the state. It illustrates that, on average, growth was strongest in Chicago and suburban Chicagoland and weakest in central and southern Illinois. Composite scores at the typical Chicago school grew by a little over four points from 223.3 to 227.5. The typical school in southern Illinois dropped a little more than a point from 232.6 to 231.3.

Figure 2.2

Changes in Composite Scores between 2006 and 2014 Were Highest in Northern Illinois and Lowest in Central and Southern Illinois

Changes in ISAT Composite Scores at Elementary and Middle Schools in Five Illinois Regions



Another helpful way to assess relative achievement growth across regions is to track changes in the percentage of students at each school who score at or above statewide averages. Since statewide averages rose between 2006 and 2014, this measure provides a rough estimate of how scoring distributions in each school and region shifted in relation to rising scores statewide.

Figure 2.3 shows what this looks like for the 542 elementary and middle schools in central Illinois. Blue dots for each school show the coordinate between percentage of students who scored at or above state averages in 2006 (horizontal axis) and the change in that percentage between 2006 and 2014 (vertical axis).

Figure 2.3 illustrates that scoring distributions at most schools in central Illinois lost ground against statewide distributions between 2006 and 2014. The red trend line shows that the higher achieving a school was in 2006, the more ground it was likely to lose relative to other schools in the state between 2006 and 2014. For example, a school which had 30 percent of students scoring at or above state averages in 2006 typically saw little or no change between 2006 and 2014. By contrast, schools with 65% of students scoring at or above statewide averages in 2006 lost an average of five percentage points between 206 and 2014. The school most typical of the region (gold diamond) lost about three percentage points between 2006 and 2014.

Figure 2.3

Percentages of Students Scoring At or Above State Averages Declined at Most Elementary and Middle Schools in Central Illinois between 2006 and 2014

8-Year Changes in School-Level Percentages of Students Scoring At/Above State Averages

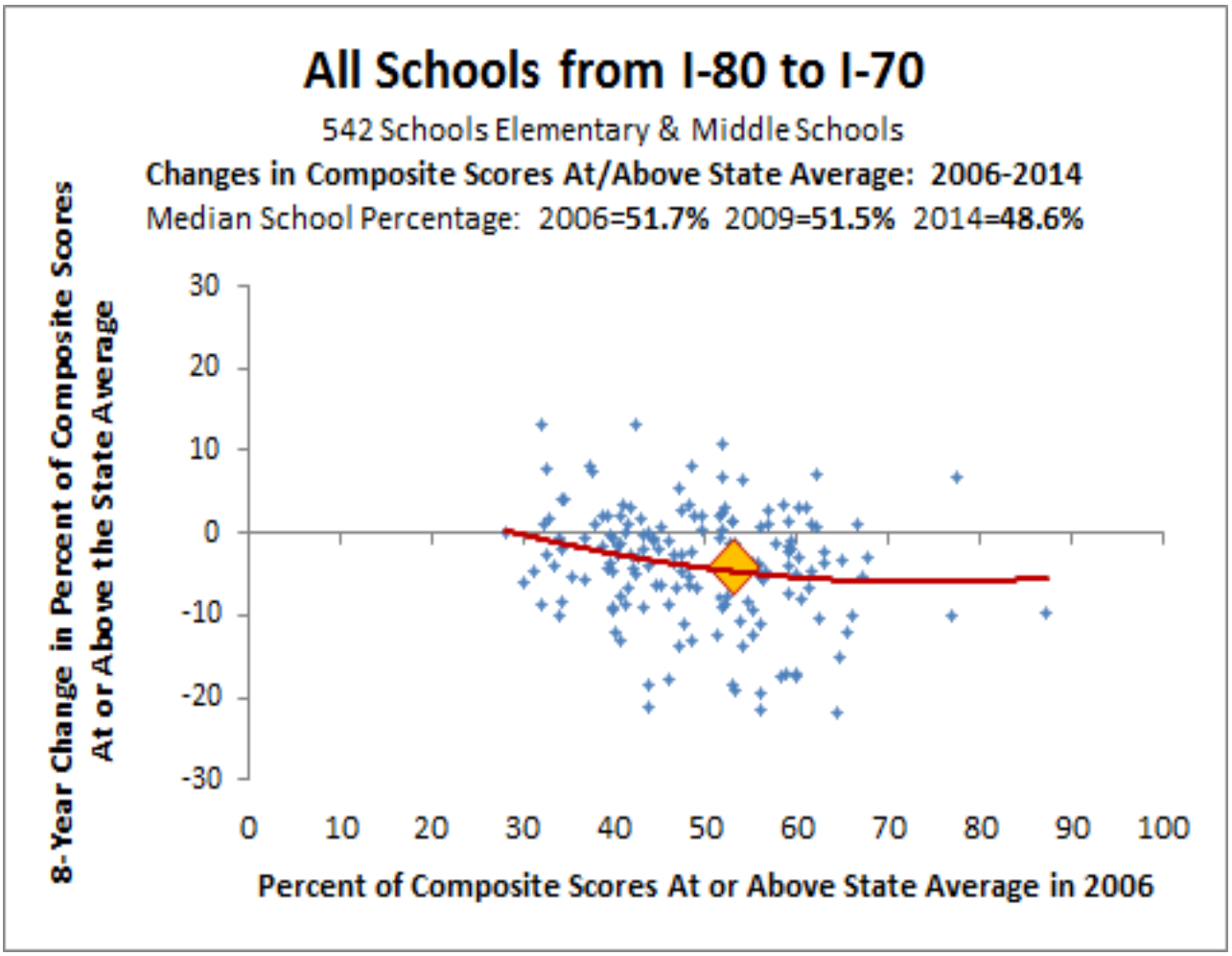


Figure 2.4 compares changes in students scoring at or above statewide averages across all five regions of the state. It illustrates that percentages of students scoring at or above statewide averages grew by an average of one percentage point in Chicago schools, with larger percentages likely at schools that were at the lower and upper ends of the achievement spectrum in 2006.

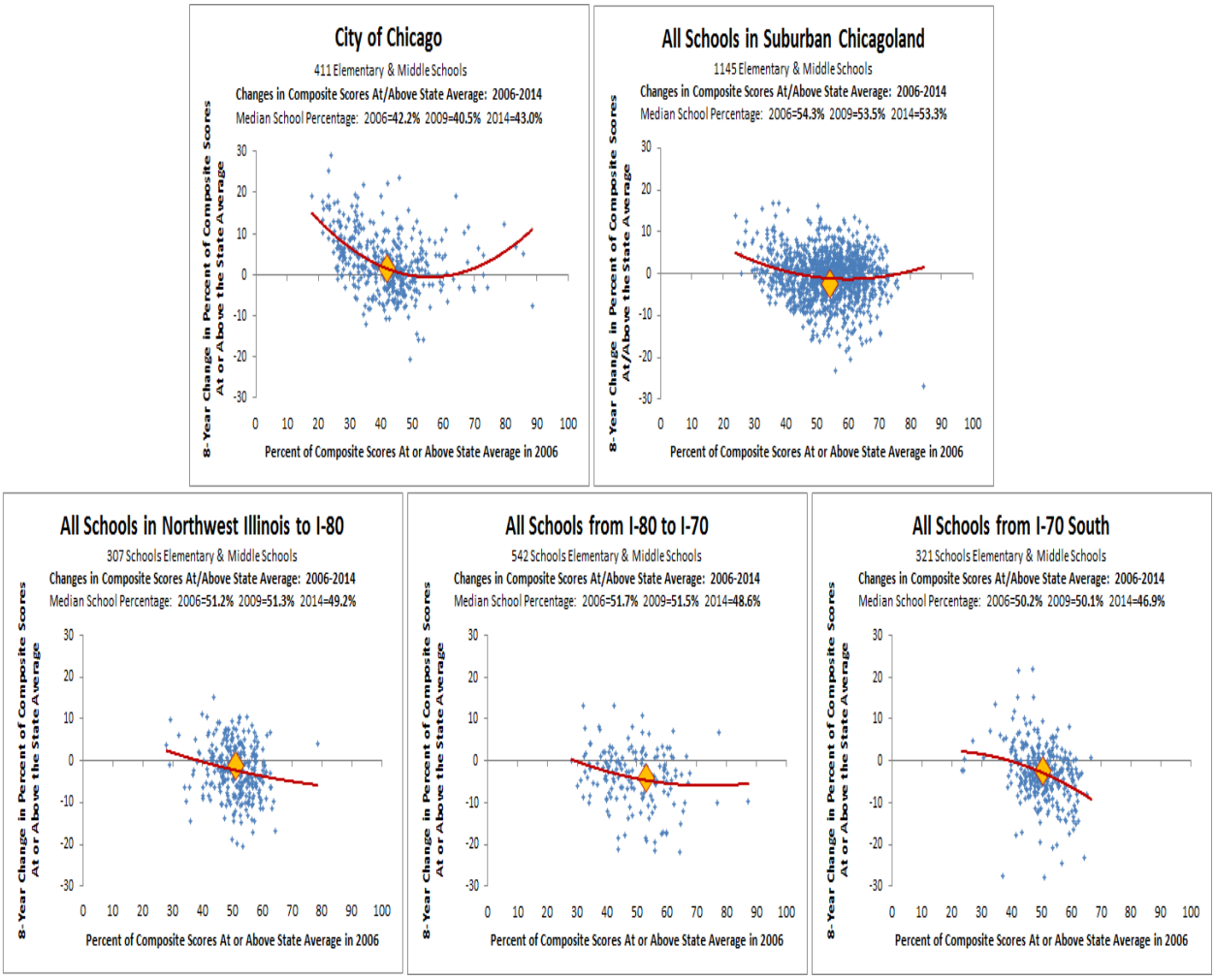
Schools in suburban Chicagoland showed a pattern similar to Chicago but, on average, lost about a percentage point compared with achievement statewide.

Schools in northwest, central and southern Illinois typically lost from two to four percentage points. Trend lines for those regions show that declines were most likely to occur at schools which, in 2006, were the highest-achieving schools in each region.

Figure 2.4

The Percentage of Students Scoring At or Above State Averages Declined at Most Elementary and Middle Schools Outside of Chicago. In Northwest, Central and Southern Illinois, Schools That Lost the Most Ground were Their Region's Highest Achieving Schools in 2006

8-Year Changes in School-Level Percentage of Students Scoring At/Above State Averages



SECTION 3

The Transition to Middle School In and Out of Chicago

In Chicago, average growth over time proceeded fairly evenly from grade three through eight during the NCLB era. By contrast, average achievement in the rest of Illinois slowed markedly as students transitioned from intermediate grades 3-5 to middle school grades 6-8

Slowed acquisition of new knowledge in the intermediate and middle school grades has long been characteristic of achievement growth in American schools. In the late 1980's the National Council of Teachers of Mathematics (NCTM) reported that the division between new learning and review in typical American classrooms flipped from 75%-new/25%-review in grade 1, to 30%-new/70%-review by grade 8.

Standardized test scales reflect slow-downs in new learning as students move through the grades. Depending on the scale being used, typical growth in primary achievement is 15 to 20 scale points per year. Average annual growth in higher grades often slows to 5 points or less. Figure 3.1 uses the reading scale from the Measures of Academic Progress (MAP) to illustrate the point. The MAP is widely used in school districts throughout Illinois to measure achievement and growth against national norms.

Figure 3.1

MAP Scale Score Growth Slows Dramatically as Students Progress through the Grades

Median Scale Scores by Grade on the MAP Reading Exam (2011 norms)

Slow-downs in new learning as students move through the grades are reflected in most other standardized tests as well. Figure 3.2 describes changes in average achievement across the grades on seven, widely-used standardized tests. Changes are shown in standard deviations.

Figure 3.2

All Major Standardized Tests Show Slowing Growth as Students Progress through the Grades

Average Achievement Growth on Seven Standardized Tests Measured in Standard Deviations

Grade Transition	Reading	Math	Science	Social Studies
Grade K - 1	1.52	1.14	--	--
Grade 1 - 2	0.97	1.03	0.58	0.63
Grade 2 - 3	0.60	0.89	0.48	0.51
Grade 3 - 4	0.36	0.52	0.37	0.33
Grade 4 - 5	0.40	0.56	0.40	0.35
Grade 5 - 6	0.32	0.41	0.27	0.32
Grade 6 - 7	0.23	0.30	0.28	0.27
Grade 7 - 8	0.26	0.32	0.26	0.25
Grade 8 - 9	0.24	0.22	0.22	0.18
Grade 9 - 10	0.19	0.25	0.19	0.19
Grade 10 - 11	0.19	0.14	0.15	0.15
Grade 11 - 12	0.06	0.01	0.04	0.04

NOTES: Adapted from Bloom, Hill, Black, and Lipsey (2008). Spring-to-spring differences are shown. The means shown are the simple (unweighted) means of the effect sizes from all or a subset of seven tests: CAT5, SAT9, Terra Nova-CTBS, Gates-MacGinitie, MAT8, Terra Nova-CAT, and SAT10.

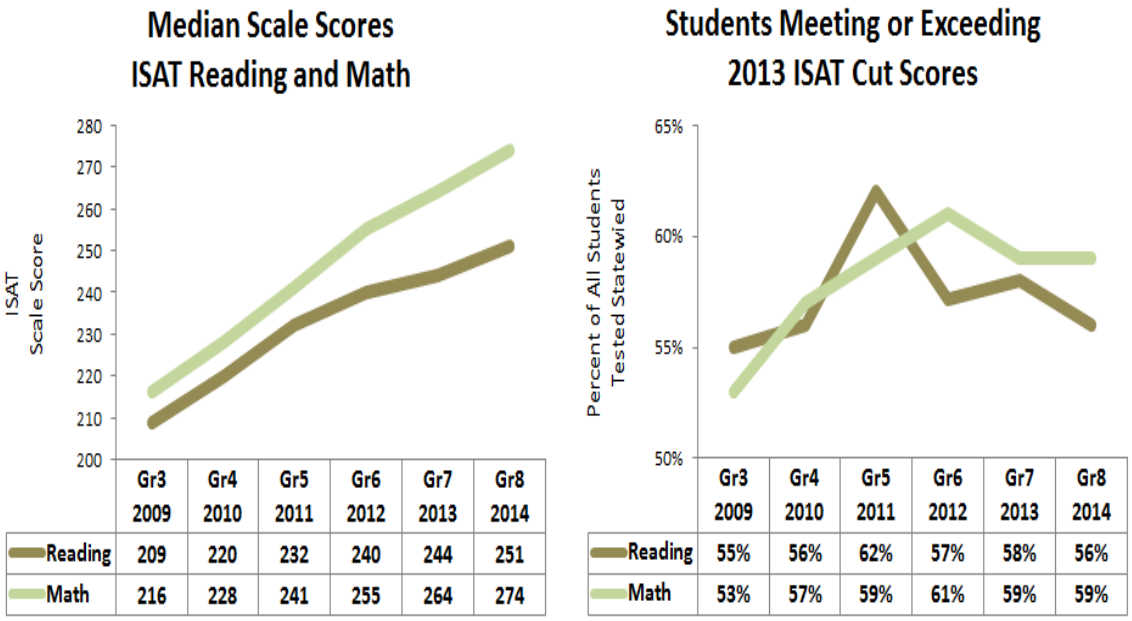
Source: Lipsey, Mark et. al. (2012) *Translating the Statistical Representation of Effects of Education Interventions into More Readily Interpretable Forms* NC SER 2013-3000, Institute for Education Sciences, US Dept. of Education

Figure 3.3 shows how decreasing scale score growth from one grade to the next affected reading and math achievement statewide for the cohort of students that graduated from eighth grade in 2014. The chart on the left shows how reading and math growth began to slow for this cohort in fifth and sixth grade. The chart on the right uses 2013 cut scores back-mapped to 2009 to show changes across he grades in the percentage of students statewide who met or exceeded. Flattening and declining percentages after grade five are worth noting because, unlike earlier cut scores, 2013 cuts were closely aligned and had roughly comparable, statewide percentile values across grade levels.

Figure 3.3

In the 8th Grade Graduating Class of 2014, Statewide Growth Declined after Grades 5 and 6

Changes in Median ISAT Scale Scores and Percentages of Students in the 8th Grade Class of 20014 Who Met or Exceeded 2013 Cut Scores from Grade 3 in 2009 through Grade 8 in 2014



	Gr3 2009	Gr4 2010	Gr5 2011	Gr6 2012	Gr7 2013	Gr8 2014
Reading	209	220	232	240	244	251
Math	216	228	241	255	264	274

	Gr3 2009	Gr4 2010	Gr5 2011	Gr6 2012	Gr7 2013	Gr8 2014
Reading	55%	56%	62%	57%	58%	56%
Math	53%	57%	59%	61%	59%	59%

The rise of college and career readiness as a state and national priority has brought renewed attention to slowing achievement in middle school. In 2008, an ACT study called *The Forgotten Middle* showed a strong predictive relationship between middle school achievement and the likelihood of meeting ACT college readiness benchmarks in grade 11. In 2011, the Hamilton Project of the Brookings Institution summarized studies from New York City and the State of Florida which showed that, by the end of 8th grade, achievement among students who attended PK-8 elementary schools was typically 0.10 to 0.15 standard deviations higher than achievement among students who attended consolidated middle schools.

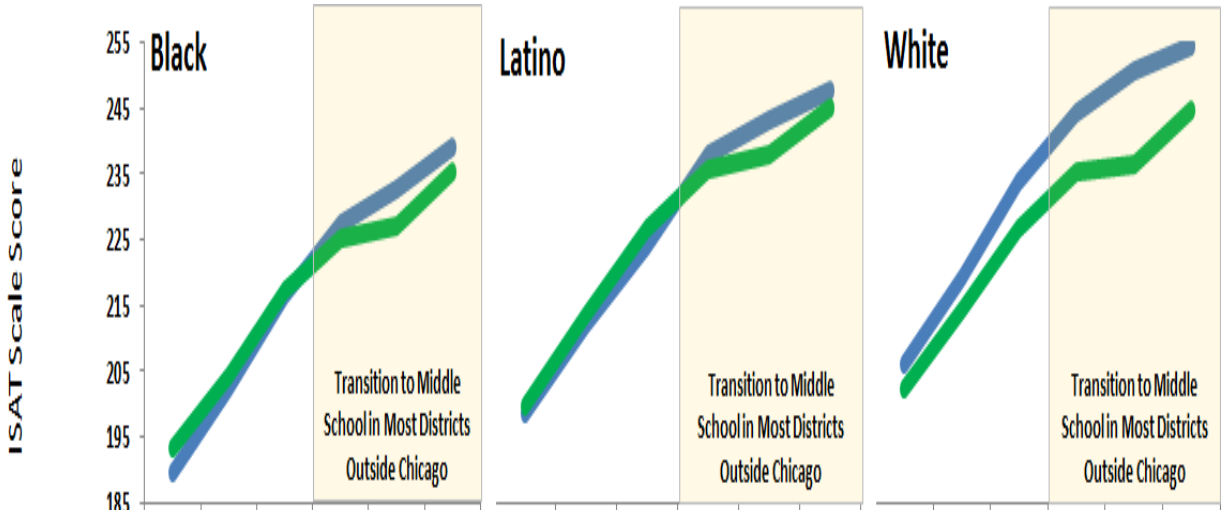
In Illinois, surprisingly little policy attention has been paid to how the transition to middle school affects the achievement of early adolescents. But big differences in school organization in and out of Chicago offer in an interesting opportunity to explore the question. While most school districts outside of Chicago move students to middle schools somewhere between fifth and seventh grade, almost all Chicago students remain at neighborhood elementary schools through the end of eighth grade.

ISAT scoring trends over time for the seven most recent cohorts of Illinois 8th graders offer compelling evidence that the transition to middle school is, on average, having a negative impact on early adolescent achievement outside of Chicago. Figure 3.4 illustrates how this impact was reflected in the average reading achievement of Black, Latino and White students from low-income households in the 8th grade graduating class of 2014. Blue lines show Chicago trends. Green lines show trends in the rest of Illinois.

Figure 3.4

Achievement Slows More Outside of Chicago as Students Transition to Middle School

Changes in Median ISAT Reading Scores for Low-Income 8th Graders in the Graduating Class of 2014



	Grade 3 2009	Grade 4 2010	Grade 5 2011	Grade 6 2012	Grade 7 2013	Grade 8 2014
Black						
City of Chicago	190	202	216	227	233	239
Rest of Illinois	193	204	217	225	227	236
Latino						
City of Chicago	199	212	224	238	243	248
Rest of Illinois	200	214	227	236	238	245
White						
City of Chicago	206	219	234	244	251	254
Rest of Illinois	202	214	227	235	236	245

The patterns shown in Figure 3.4 accurately represent differences in median scale scores in and out of Chicago. However, they underreport the full magnitude of those differences. The reason they do is that scale score differences have different meanings from one grade level to the next. As illustrated in Figure 3.1, a one-point difference in 8th grade MAP scores represents one third of the total expected gain during 8th grade. In 3rd grade, one point is only about a tenth of the expected gain for the year.

Figure 3.5 controls for differences in expected reading gains at different grade levels by converting numerical differences into standardized differences that are measured in standard deviations:

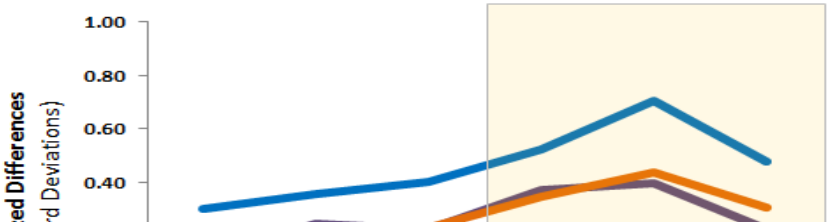
- Positive changes in standardized differences . . . about 0.20 standard deviations among Black and Latino students, more among Whites . . . reflect higher achievement in Chicago than in the rest of Illinois
- Upward shifts in the pitch of lines after grade five reflect a sudden widening of achievement differences; the steeper the pitch, the wider the difference.

Figure 3.6 shows that differences in math achievement also widen after grade five.

Figure 3.5

Standardized Differences in Average Reading Scores Increase between Chicago and the Rest of Illinois as Most Students Outside of Chicago Transition to Consolidated Middle Schools

Changes in ISAT Reading Achievement for Three Groups of Low-Income Students In and Out of Chicago

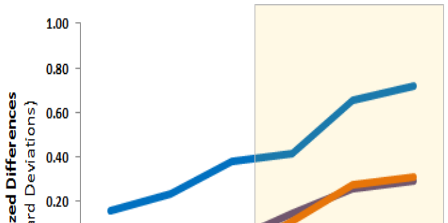


	Grade 3 2009	Grade 4 2010	Grade 5 2011	Grade 6 2012	Grade 7 2013	Grade 8 2014
Black	0.13	0.25	0.23	0.37	0.40	0.22
Latino	0.22	0.18	0.23	0.35	0.44	0.31
White	0.30	0.36	0.40	0.52	0.71	0.48

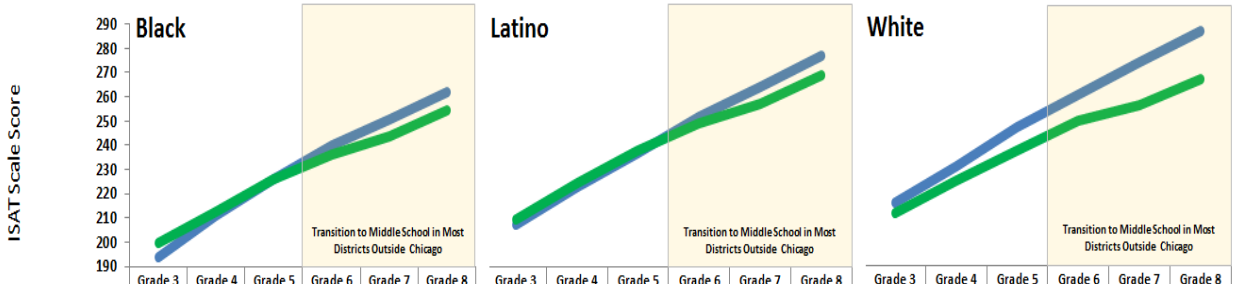
Figure 3.6

Changes in Median Scores and Standardized Differences Follow the Same Basic Pattern in Math as They Do in Reading

Changes in ISAT Math Achievement for Three Groups of Low-Income Students In and Out of Chicago



	Grade 3 2009	Grade 4 2010	Grade 5 2011	Grade 6 2012	Grade 7 2013	Grade 8 2014
Black	-0.20	-0.06	0.02	0.14	0.25	0.29
Latino	-0.08	-0.06	-0.04	0.11	0.28	0.31
White	0.16	0.24	0.38	0.41	0.65	0.72



	Grade 3 2009	Grade 4 2010	Grade 5 2011	Grade 6 2012	Grade 7 2013	Grade 8 2014
City of Chicago	194	211	226	240	250	262
Rest of Illinois	200	213	226	236	244	255

	Grade 3 2009	Grade 4 2010	Grade 5 2011	Grade 6 2012	Grade 7 2013	Grade 8 2014
City of Chicago	207	223	237	252	264	277
Rest of Illinois	210	224	238	249	257	269

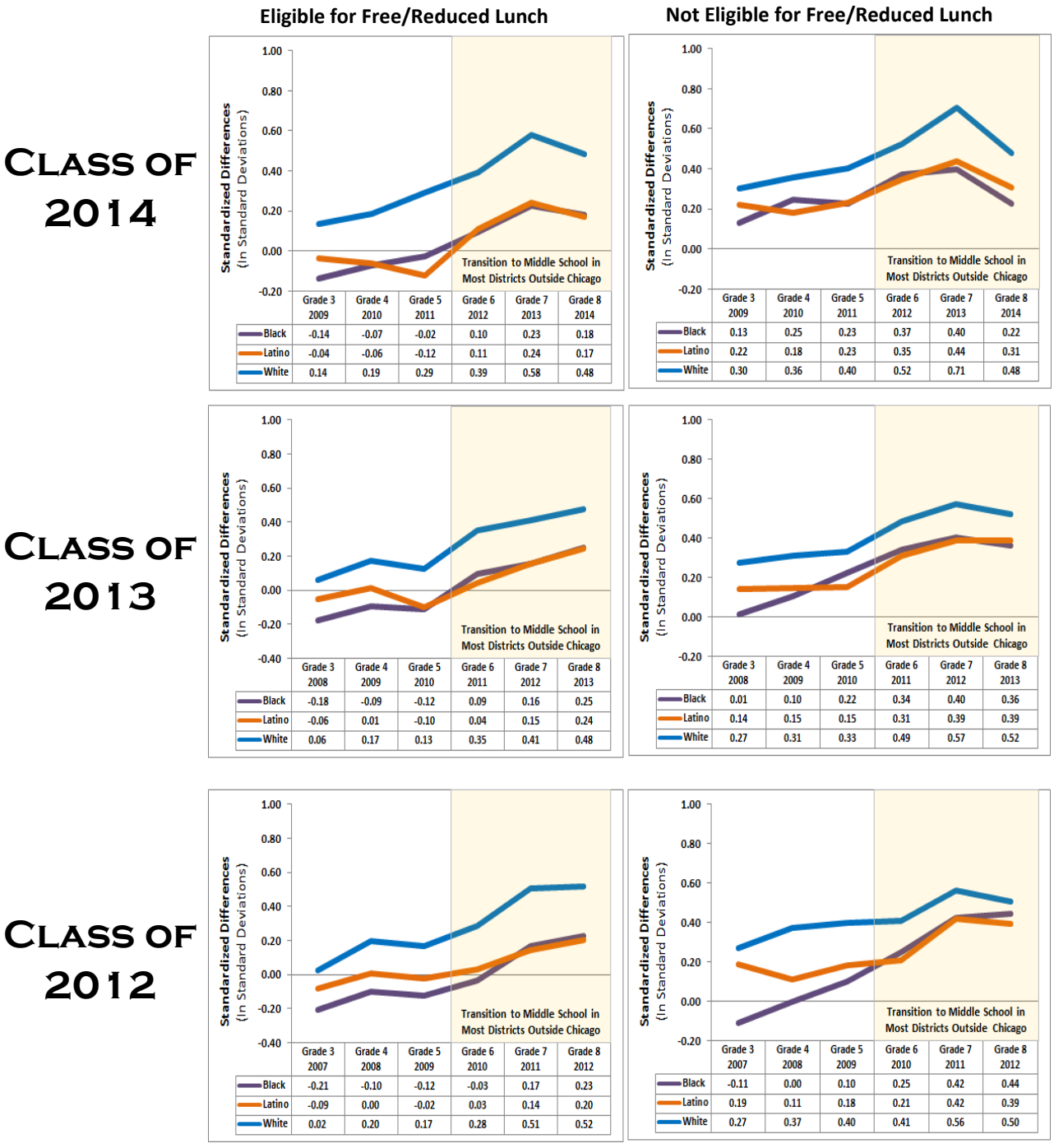
	Grade 3 2009	Grade 4 2010	Grade 5 2011	Grade 6 2012	Grade 7 2013	Grade 8 2014
City of Chicago	216	231	247	261	274	286
Rest of Illinois	212	225	237	250	256	267

Figure 3.7 illustrates that widening achievement differences after grade five have been a consistent feature of reading achievement in and out of Chicago across consecutive cohorts of students.

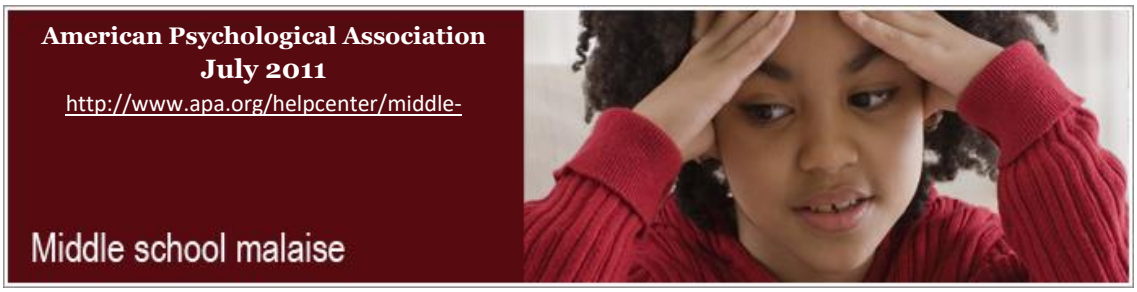
Figure 3.7

Gaps in Instructional Effectiveness Widen between Chicago and the Rest of Illinois As Students Transition from Grades 3-5 to Grades 6-8

Standardized Differences in Average **ISAT Reading Scores** for Three Recent 8th Grade Graduating Classes



Recent Research on the Transition to Middle School



"The switch from elementary to junior high school coincides with several major changes for young adolescents. Most are in the throes of puberty; they're becoming more self-aware and self-conscious, and their thinking is growing more critical and more complex. At the same time, adolescents are often "in a slump" when it comes to academic motivation and performance.

"Researchers at the University of Michigan have studied the transition from elementary to middle school and have found that:

- *On average, children's grades drop dramatically during the first year of middle school compared to their grades in elementary school.*
- *After moving to junior high school, children become less interested in school and less self-assured about their abilities.*
- *Compared to elementary schools, middle schools are more controlling, less cognitively challenging and focus more on competition and comparing students' ability.*

"Through this and other similar research, psychologists have discovered a "developmental mismatch" between the environment and philosophy of middle schools and the children they attempt to teach. At a time when children's cognitive abilities are increasing, middle school offers them fewer opportunities for decision-making and lower levels of cognitive involvement, but a more complex social environment. At the same time, numerous teachers have replaced the single classroom teacher and students often face larger classes and a new group of peers.

"These factors all interact to make the transition to junior high school difficult for many youngsters. Studies find the decreased motivation and self-assuredness contribute to poor academic performance; poor grades trigger more self-doubt and a downward spiral can begin."

*****

Schwerdt, G., & West, M. R. (2011). *The impact of alternative grade configurations on student outcomes through middle and high school.* Cambridge, MA: Institute for Economic Research, Harvard University and Harvard Graduate School of Education.

"*We use statewide administrative data from Florida to estimate the impact of attending public schools with different grade configurations on student achievement through grade 10. Based on an instrumental variable estimation strategy, we find that students moving from elementary to middle school suffer a sharp drop in student achievement in the transition year. These achievement drops persist through grade 10. We also find that middle school entry increases student absences and is associated with higher grade 10 dropout rates. Transitions to high school in grade nine cause a smaller one-time drop in achievement but do not alter students' performance trajectories."*

Available from: http://www.edweek.org/media/gradeconfiguration-13structure.pdf

For additional studies, see also Regional Education Laboratory REL Central

https://www.relcentral.org/what-does-the-research-say-about-sixth-grade-placement-should-they-be-in –an-elementary-school-or-a-middle -school/

SECTION 4

Primary Achievement In and Out of Chicago

For over a decade, reading and math gains in Chicago have substantially outpaced gains in the rest of Illinois. But until recently, primary achievement in Chicago lagged behind primary achievement in the rest of Illinois. By 2015, however, Black, Latino and White achievement at all grade levels tested was the same or higher in Chicago than it was the rest of the state.

Chicago has always had a special status in statewide achievement reportage because it accounts for close to 20% of the entire statewide test population. It also has much higher concentrations of low-income students of color than most other areas of the state. But while Chicago achievement has always been reported separately from statewide achievement, the same has not been true for aggregate, statewide achievement outside of Chicago. This omission has made it difficult for most of the public to see how achievement among racial sub-groups outside of Chicago compared with that of comparable sub-groups in the city.

In 2007, and again in 2011, the Consortium on Chicago School Research reported that achievement in Chicago actually surpassed achievement in the rest of Illinois after controlling for racial differences in each group. But these studies also showed that, on average, Chicago students in lower grades continued to achieve at lower levels than their counterparts in the rest of the state.

This section takes a fresh look at third grade achievement in and out of Chicago after controlling simultaneously for race, family income and English language proficiency. Third grade achievement patterns have special significance because:

- they reflect the cumulative effect of all primary and early childhood instruction
- they are strong predictor of future achievement and paint a clear picture of challenges that lie ahead for improving instructional effectiveness in Illinois schools

Figures 4.1 and 4.2 describe achievement changes in two ways. The pair of green and blue lines at the top of each cluster shows aggregate achievement changes in and out of Chicago from 2001 through 2014. The four charts at the bottom of each cluster break down aggregate achievement by race and family income level. Numbers on the right side of each chart reflect estimated scale score gains from 2001 through 2014. For ease of comparison, scale scores for 2001 through 2005 have been converted into values which closely match those used from 2006 onward

The upper charts in Figures 4.1 and 4.2 illustrate that overall achievement outside of Chicago (green lines) was close to flat throughout the NCLB era. But all of the major groups that contributed to

achievement outside of Chicago made modest to strong gains under NCLB. As described in Section 5, the explanation for this paradox is that changes in the size of each group changed the contribution that each group made to overall gains (see Figure 5.5). The net effect was that aggregate statewide gains were far smaller than gains made by each contributing group.

Figures 4.1 and 4.2 show that most Chicago sub-groups grew at faster rates than their counterparts in the rest of Illinois. By 2014, all sub-groups in Chicago were achieving at levels that matched or exceeded those in the rest of Illinois.

Figure 4.1

Third Grade ISAT READING Medians in Chicago and the Rest of Illinois

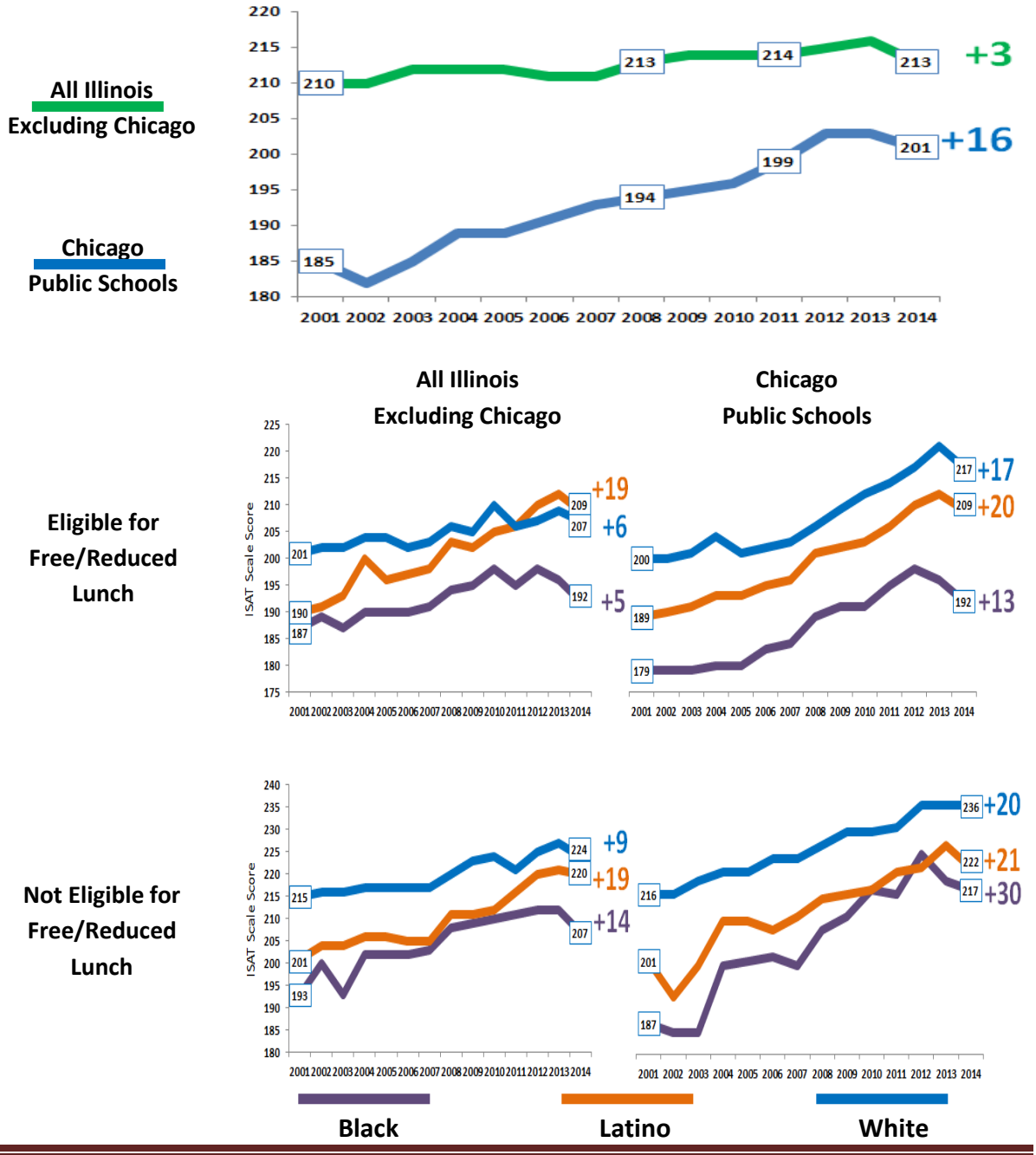
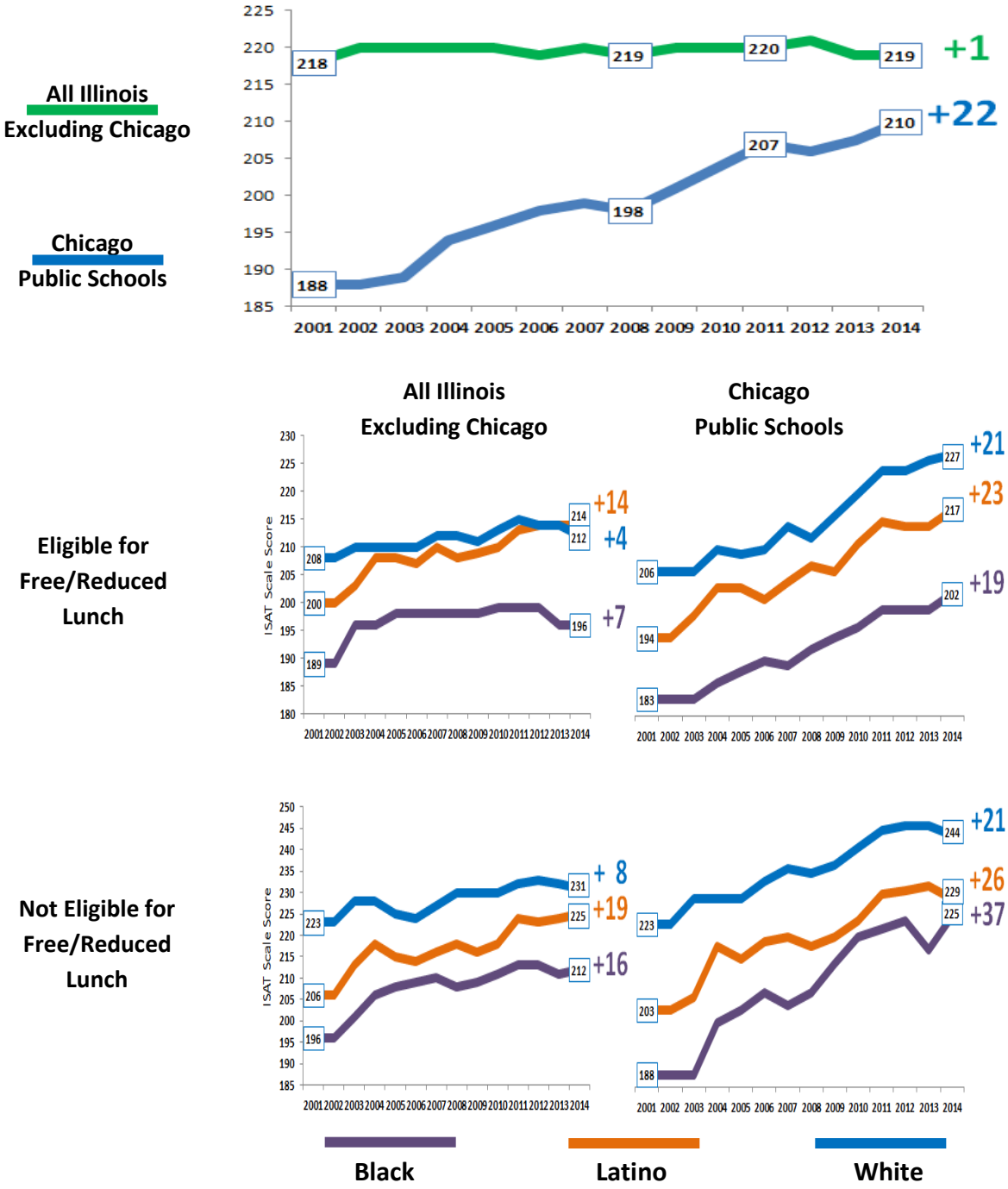


Figure 4.2

Third Grade ISAT MATH Medians in Chicago and the Rest of Illinois



Recent Trends

Figures 4.3 and 4.4 report the percentage of students who scored at or above statewide reading and math medians on recent ISAT and PARCC exams. Because third grade achievement is a strong predictor of future achievement, they offer a glimpse of what achievement patterns are likely to look like in and out of Chicago during the decade ahead.

Figure 4.3 shows continuing growth in reading achievement across all Chicago sub-groups and flat or declining achievement among most sub-groups in the rest of Illinois.

Figure 4.3
Percentages of 3rd Graders in Chicago and the Rest of Illinois Who Scored At or Above Statewide Medians on Recent ISAT and PARCC READING Exams

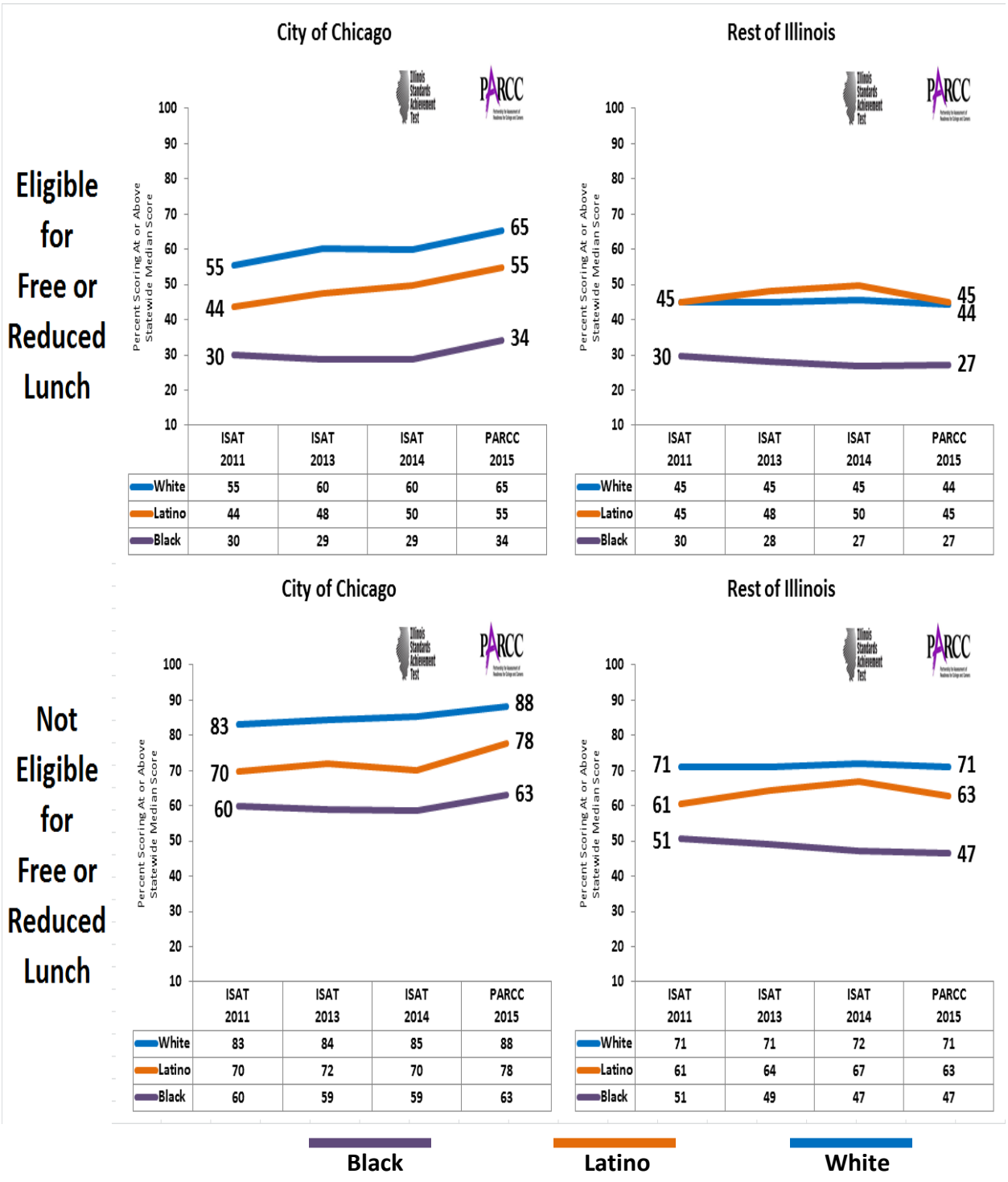
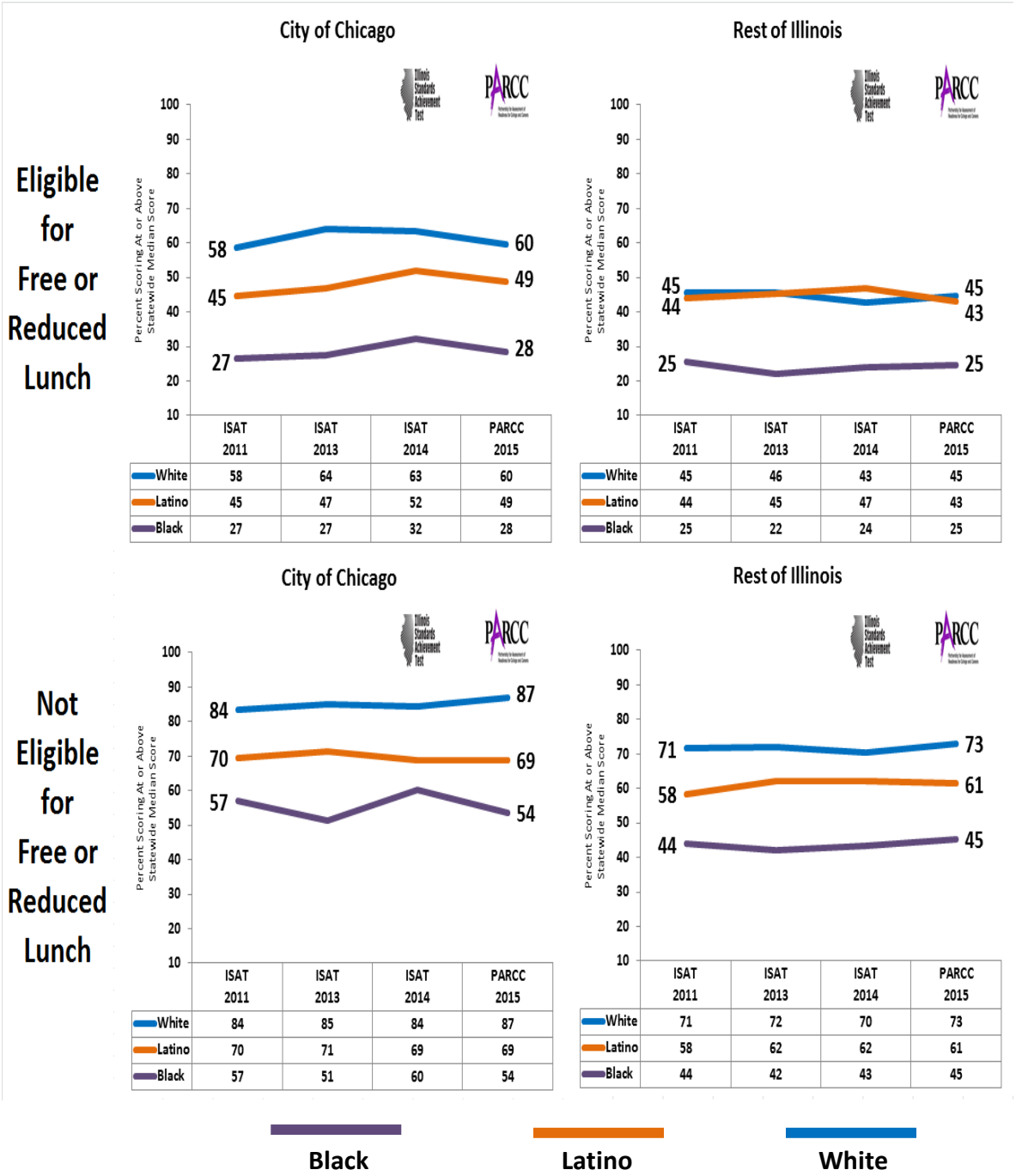


Figure 4.4 shows more mixed patterns of growth in math achievement in and out of Chicago. Differences in levels of achievement in and out of Chicago are also smaller in math than they are in reading.

Figure 4.4
Percentages of 3rd Graders in Chicago and the Rest of Illinois Who Scored At or Above Statewide Medians on Recent ISAT and PARCC MATH Exams



THE ANNIE E. CASEY FOUNDATION

Double Jeopardy

How Third-Grade Reading Skills and Poverty Influence High School Graduation

Annie E. Casey Foundation (2012)

"Educators and researchers have long recognized the importance of mastering reading by the end of third grade. Students who fail to reach this critical milestone often falter in the later grades and drop out before earning a high school diploma. Now, researchers have confirmed this link in the first national study to calculate high school graduation rates for children at different reading skill levels and with different poverty rates.

"Results of a longitudinal study of nearly 4,000 students find that those who do not read proficiently by third grade are four times more likely to leave school without a diploma than proficient readers. For the worst readers, those who could not master even the basic skills by third grade, the rate is nearly six times greater. While these struggling readers account for about a third of the students, they represent more than three-fifths of those who eventually drop out or fail to graduate on time.

"What's more, the study shows that poverty has a powerful influence on graduation rates. The combined effect of reading poorly and living in poverty puts these children in double jeopardy.

- *About **16** percent of children who are not reading proficiently by the end of third grade do not graduate from high school on time, a rate four times greater than that for proficient readers*
- *For children who were poor for at least a year and were not reading proficiently, the proportion failing to graduate rose to **26** percent*
- *For children who were poor, lived in neighborhoods of concentrated poverty and not reading proficiently, the proportion jumped to **35** percent*
- *Overall, **22** percent of children who lived in poverty do not graduate from high school, compared to 6 percent of those who have never been poor. The figure rises to **32** percent for students spending more than half of their childhood in poverty.*
- *Even among poor children who were proficient readers in third grade, **11** percent still did not finish high school. That compares to 9 percent of subpar third-grade readers who have never been poor.*
- *About **31** percent of poor African-American students and **33** percent of poor Hispanic students who did not hit the third-grade proficiency mark failed to graduate. These rates are greater than those for White students with poor reading skills. But the racial and ethnic graduation gaps disappear when students master reading by the end of third grade and are not living in poverty."*

http://www.aecf.org/resources/double-jeopardy/

PART 2

An Alternate Universe of Large-Scale Test Information

In all cases, the trends described in PART 1 run counter to widely held assumptions about achievement patterns in Illinois. The obvious question is how these things could have gone unnoticed in an era that generated more data about achievement than any prior era in public education history.

The answer lies in the way large-scale assessment information was packaged and reported under NCLB. That packaging misrepresented what tests actually assessed and failed to communicate meaningful information about student achievement over time.

What Happened?

Prior to the passage of No Child Left Behind, standardized testing had smaller ambitions than it does today. For the most part, standardized tests stuck to comparing large groups of people with each other on various measures of aptitude and achievement. They offered useful tools for assessing general knowledge and predicting future performance but made no pretense of being able to diagnose mastery of specific skills and content knowledge. Their strength lay in measuring performance relative to others being tested. Their weakness was that forty-nine percent of every test population always had to score "below average"

NCLB called for a very different kind of assessment. It expected states to develop large-scale tests that could assess achievement and growth against well-defined academic standards. In one, bold statutory swoop, NCLB required states to spell out clear standards for what students needed to learn, and to build assessments that 100% of students could potentially pass.

Illinois and most other states signaled the shift to standards-based assessment with a whole new palette of reporting strategies. Test reports no longer described the percentage of students who scored at or above grade level, or the percentage of students who scored in each quartile compared with state or national norms. Instead, they reported the percentage of students who scored at different "proficiency levels" and paid particular attention to the percentage of students who "met or exceeded state standards." To inform instruction, this information was supplemented with "content strands" and "item analysis" that purported to diagnose specific aspects of standards mastery. The message was clear. Unlike older tests that compared students with each other, new tests assessed mastery of specific skills and content knowledge that were spelled out in state standards.

"You Go to War with the Army You Have . . ."

In December 2004, a disgruntled American soldier challenged Secretary of Defense Donald Rumsfeld to explain why his unit had to rummage through trash heaps for scrap metal they could use to strengthen the armor of their old Humvees. Rumsfeld famously responded, "You go to war with the army you have . . . not the army you might want or wish to have at a later time."

In 2001, the "army we had" for revolutionizing large-scale assessment design was big banks of norm-referenced test items and close to a century of experience building tests that compared students with each other. It was mostly these resources that the testing industry relied on to build large scale, "standards-based" assessments. As a result, most of what came to be called standards-based testing under NCLB was just conventional, norm-referenced testing dressed up in standards-based clothing.

From the beginning, signs were clear that dressing up the "army we had" in standards-based clothing was not going to be a responsible strategy. The first alarm came from statisticians and measurement professionals (see below) They warned that cut scores and "meet/exceed" metrics ignored the mathematical properties of scoring distributions and introduced deep distortions into the results that tests produced. Then, a growing stream of studies reported that most states were finessing NCLB accountability requirements by setting very low thresholds for meeting state standards. More recently, studies have shifted their critique to standards and assessments themselves, asserting that both reflected very low levels of academic rigor.

Sections 5, 6, 7 and 8 describe some key ways that standards-based packaging violated public trust by misrepresenting what large-scale test results were actually measuring:

- Section 5 shows how oversimplified reporting practices reinforced old stereotypes and missed important changes in achievement gaps that are commonly associated with race, family income and English language proficiency
- Section 6 describes how arbitrary "standard setting" obscured the close match between ISAT results and results from more highly regarded tests like the Measures of Academic Progress (MAP), National Assessment of Educational Progress (NAEP), ACT and most recently, PARCC
- Section 7 looks more closely at what standardized test items actually assess and examines how very different tests end up producing close-to-identical results
- Section 8 explains why common NCLB diagnostic reports like "content strands," "item analysis" and "power standards" are packaging gimmicks that misrepresent and under-report most of what standardized tests actually assess.

For Every Complex Problem, There is an Answer that is Clear, Simple and Wrong

H.L Mencken

Cut scores were the tool that almost all NCLB-era tests used to grade and report "standards-based" test results. On their face, cut scores offered a simple, clearly-defined way to define proficiency levels and to identify the point on standardized test scales where students "met state standards."

It turns out that imposing cut scores on normal distributions of test results creates a raft of technical distortions that compromise validity and reliability In important ways. In the early years of NCLB, Andrew Ho, now Professor of Education at the Harvard Graduate School of Education, warned that cut scores fatally distort standardized measures of academic progress. That warning applied not only to state assessment systems like the ISAT and PSAE, but to more venerable systems like the NAEP as well

"The limitations [that are introduced by cut scores] are unpredictable, dramatic, and difficult to correct in the absence of other data. Interpretation of these depictions generally leads to incorrect or incomplete inferences about distributional change . . . [and can lead] to short-sighted comparisons between state and national testing results."

Andrew Dean Ho (2008) "The Problem with "Proficiency": Limitations of Statistics and Policy Under No Child Left Behind" *Educational Researcher*, Vol. 37, No. 6, p. 351

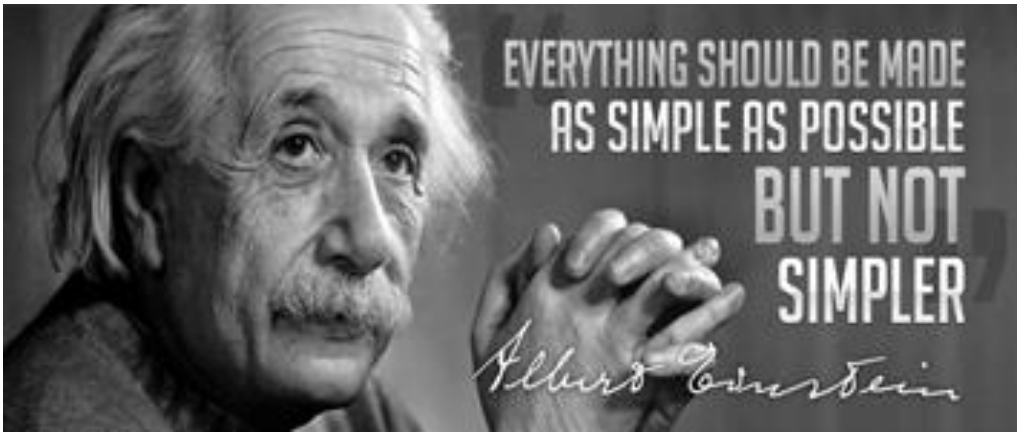
Cut score distortions were hard to confront in the early days of NCLB because they required technical explanations that can be difficult for non-statisticians to follow. To address this problem, Ho and the University of Iowa created a 44-minute video clip on the subject in 2006. This video can be viewed at: http://www2.education.uiowa.edu/html/tv/talent/stats/index.htm

For an early look at distortions that cut scores created in Illinois, see also, Zavitkovsky, Paul (2009). *Something's Wrong with Illinois Test Results,* Urban Education Leadership Program, University of Illinois--Chicago.

SECTION 5

Simple as Possible . . . but Not Simpler

Oversimplified reporting practices reinforced old stereotypes and missed important changes in achievement gaps that are commonly associated with race, family income and English language proficiency.



The central goal of NCLB was reduce chronic gaps in achievement and instructional effectiveness long associated with race, gender, family income and other demographic characteristics. But the metrics used to increase transparency and track progress toward this goal were too simplistic to do the job. Instead, they perpetuated stereotypes about race, class and academic achievement, and missed important shifts in instructional effectiveness that occurred under NCLB.

Figure 5.1 illustrates how the State of Illinois reported statewide shifts in third grade reading achievement. Following the formal requirements of NCLB, achievement was broken out by race, family income and English language proficiency. Across two big changes in cut scores in 2006 and 2013, this way of reporting still showed a consistent 30 to 40 point achievement gaps between:

- White students and their Black and Latino counterparts
- Students who were and were not eligible for free or reduced lunch
- Student who were and were not English proficient

The message of Figure 5.1 is that gaps in achievement and instructional effectiveness stayed more or less the same under NCLB. Black achievement moved marginally upward compared with White and Latino achievement. And achievement among English language learners declined relative to English-proficient students. But cut score changes in 2006 and 2013 made it unclear whether those differences were real, or were simply the result of slicing up scoring distributions in different ways.

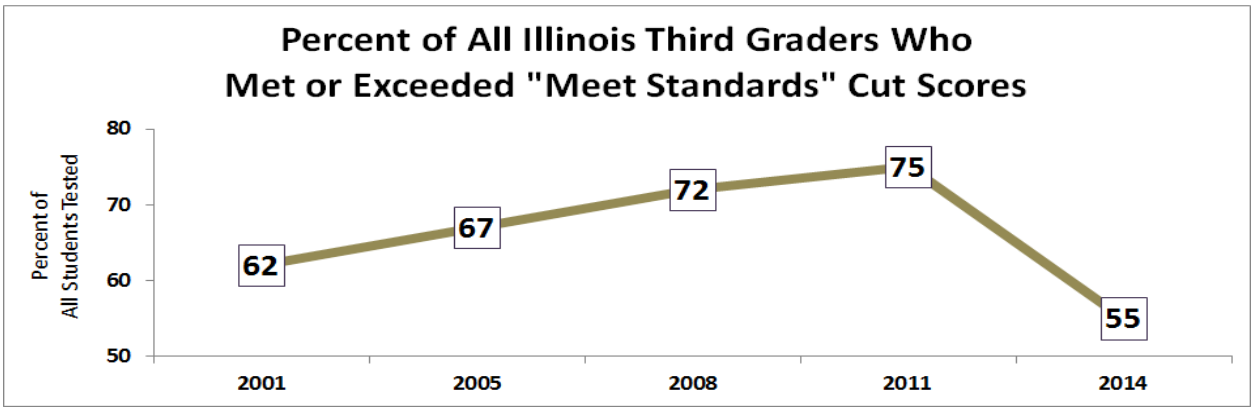
An obvious problem with describing achievement in this way is that racial and other demographic sub-groups are not homogeneous. Latino achievement among students from low-income households and who are identified as English Language Learners (ELL) is not really the same as Latino achievement among students who are English-proficient and come from middle income families. But statewide

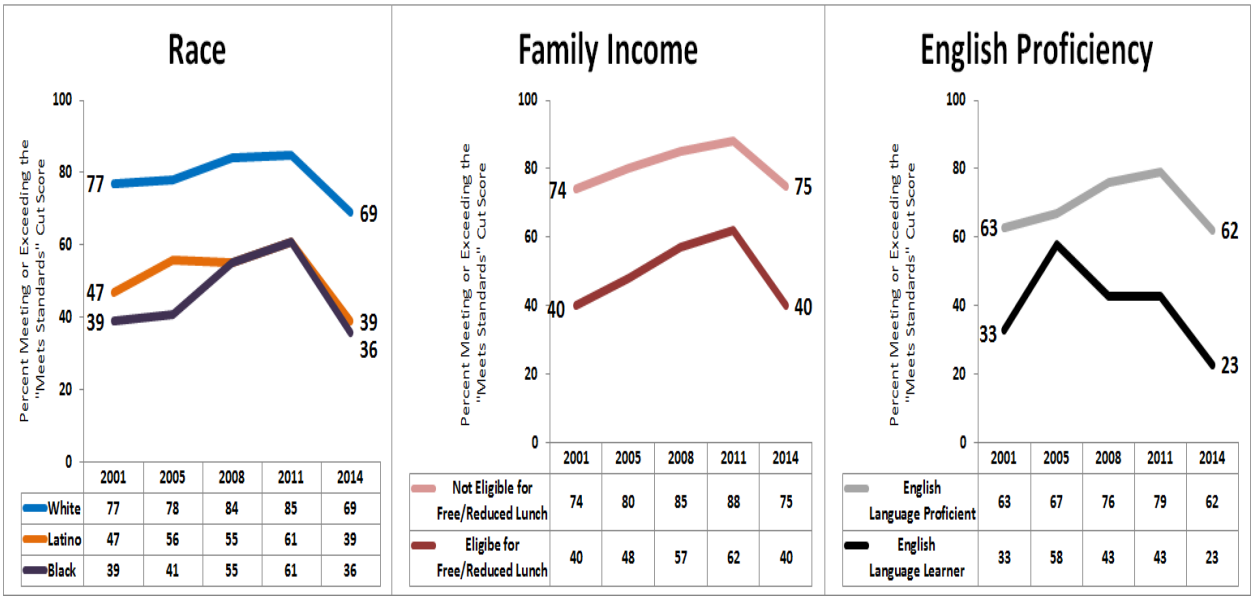
reporting procedures ignored these differences and lumped all Latino achievement into a single racial category. So doing, they met the letter of the law, but completely ignored its spirit.

Figure 5.1

Official Reportage Shows No Real Change in Achievement Gaps under NCLB

Third Graders Who Met or Exceeded ISAT Cut Scores for "Meeting State Standards" in Reading: 2001-2014





Walking and Chewing Gum at the Same Time

No meaningful picture of statewide achievement changes under NCLB is possible without controlling ***simultaneously*** for key demographic factors that affect achievement. Figure 5.2 shows what that looks like using exactly the same data that Figure 5.1 does. Unlike Figure 5.1, however, Figure 5.2 controls simultaneously for differences in race, family income and English language proficiency.

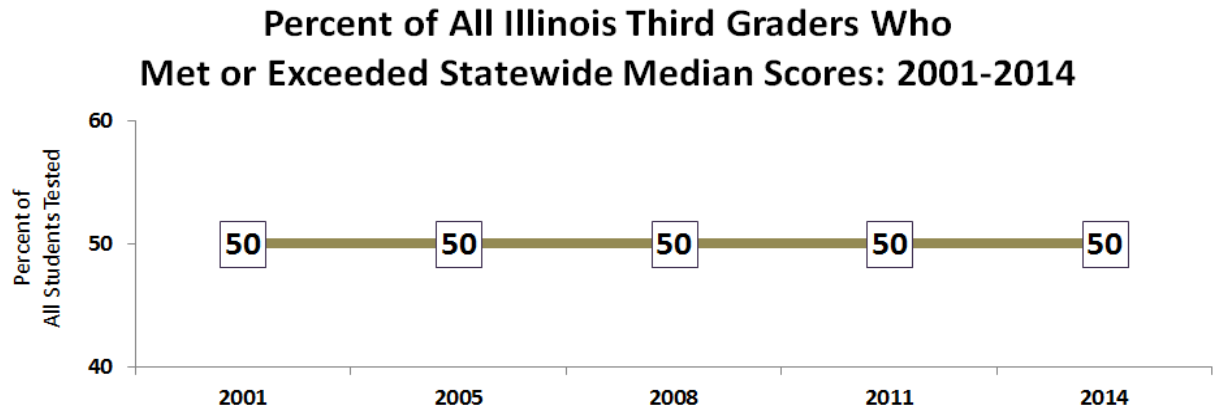
Figure 5.2 starts by excluding students who were temporarily identified as English language learners from each of the three racial groups shown. Thein It breaks each group down into students who were and were not eligible for free or reduced lunch. In addition, Figure 5.2 eliminates the distortions of shifting cut scores by reporting the percentage of students in each group who scored at or above the

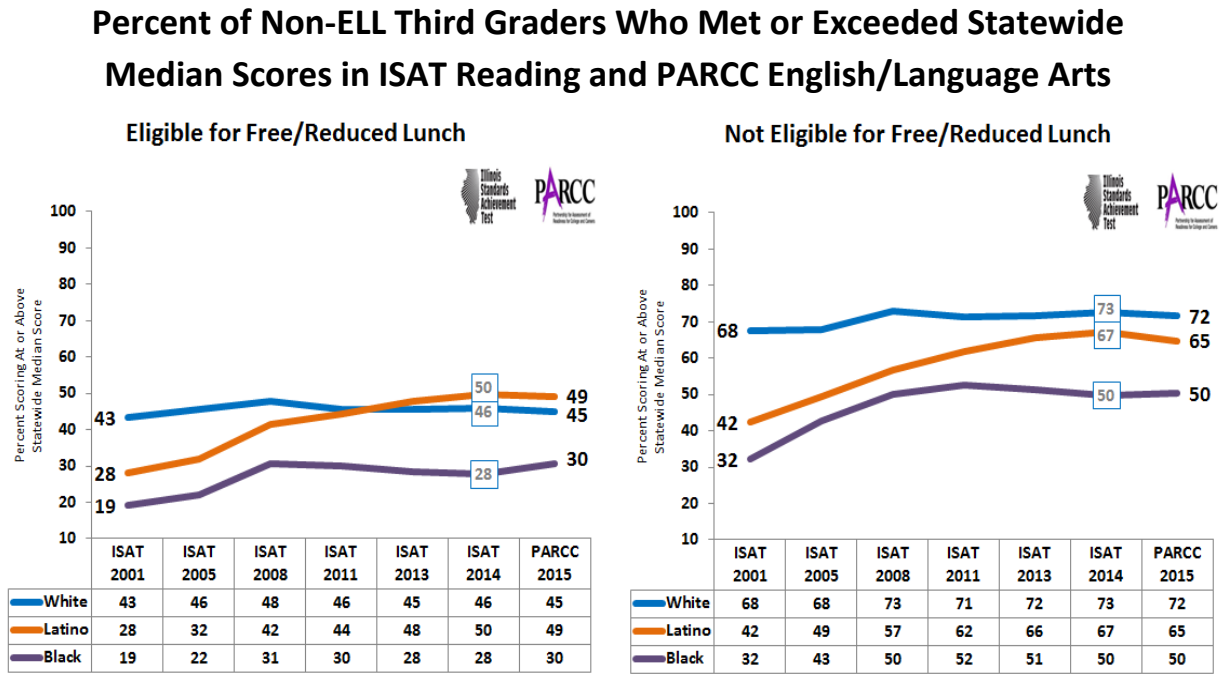
statewide median. These controls produce a very different picture of achievement changes under NCLB than the one that is painted by Figure 5.1.

Figure 5.2

Latino Achievement Grew Dramatically Relative to White and Black Achievement under NCLB

Illinois Third Graders Who Met or Exceeded Statewide Median Scores in Reading: 2001-2014





In Figure 5.2, both groups of White students (blue lines) made relatively small gains compared with those of Black and Latino students, and made no real gains at all between 2008 and 2014. Both groups of Black students (purple lines) made steady gains between 2001 and 2008 but then made no growth between 2008 and 2014. PARCC results in 2015 closely mirrored 2014 results on the 2014 ISAT.

By contrast, both groups of Latino students (tan lines) made sustained gains over the entire NCLB era. Among students not eligible for free or reduced lunch, differences between non-ELL White and non-ELL Latino students narrowed by 20 percentage points between 2001 and 2014. Differences

between non-ELL White and non-ELL Latino students from low-income households changed from a 15-point Latino deficit in 2001 to a 4-point Latino advantage in 2014. Once again, PARCC results in 2015 mirrored results on the 2014 ISAT

Lots of Moving Parts

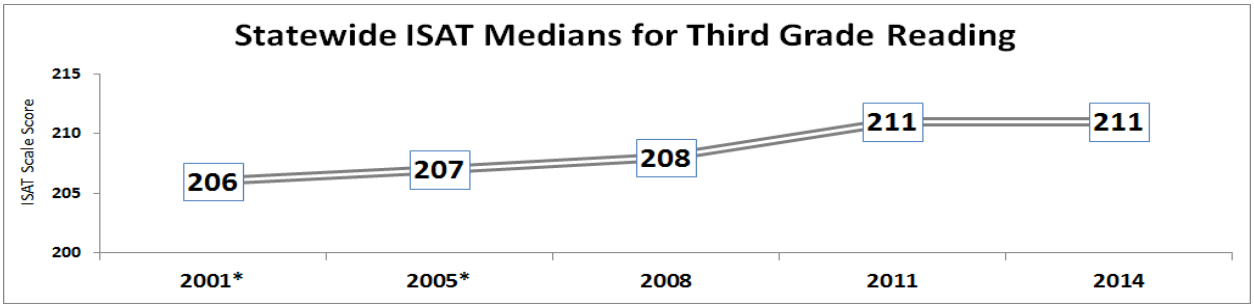
Figure 5.3 continues to look at third grade reading achievement by showing changes in median ISAT scale scores from 2001 through 2014:

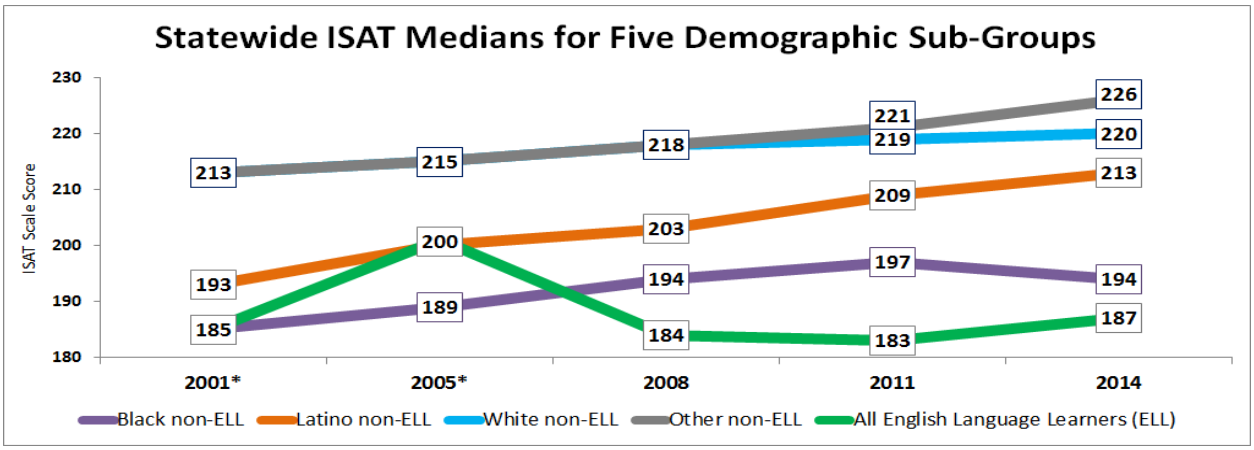
- The upper chart shows changes in median scores for all third graders tested statewide, with actual medians from 2001 and 2005 converted into scale values used between 2006 and 2014
- The lower chart shows changes in median scores for each of the five sub-groups that contributed to overall statewide changes in the upper chart

Figure 5.3

Overall Changes Statewide Don't Match Neatly with Changes in Contributing Sub-Groups

Median ISAT Scale Scores for Illinois Third Graders in 2001, 2005, 2008, 2011 and 2014

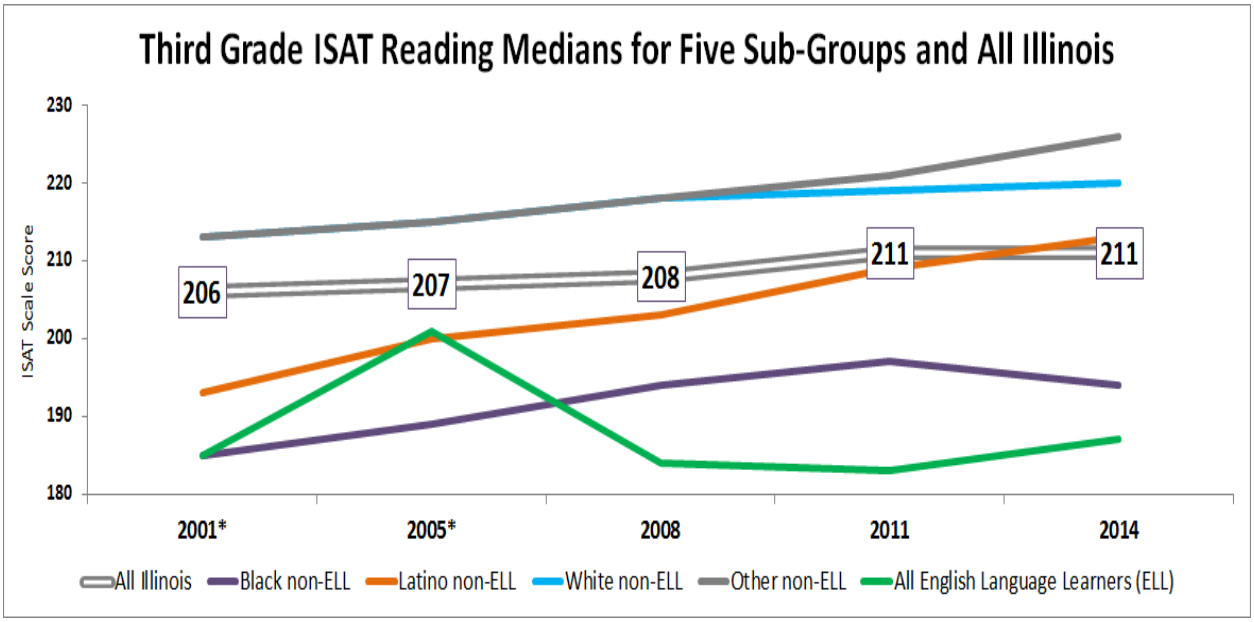


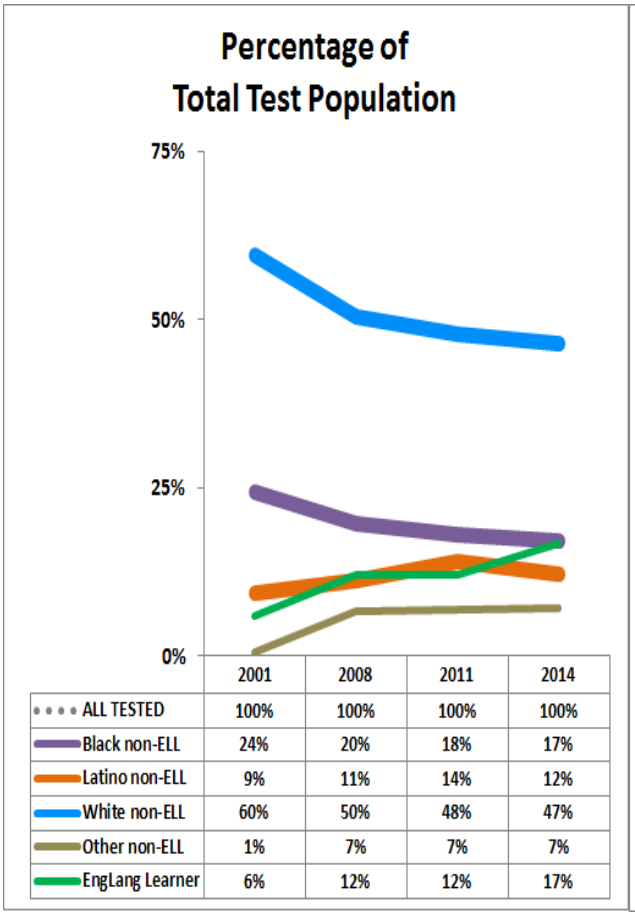


On their face, the two charts shown in Figure 5.3 appear to contradict each other. In the upper chart, statewide reading achievement in third grade flattened between 2011 and 2014. But four of the five sub-groups that contributed to overall scores in 2011 and 2014 showed gains between 2011 and 2014. The only group that didn't was non-ELL Black students, and that group only accounted for 18% of the total statewide test population. The explanation for this apparent conflict lies in a dense mix of underlying changes that occurred within each group, but most especially among non-ELL, White students. These changes are summarized in Figure 5.4.

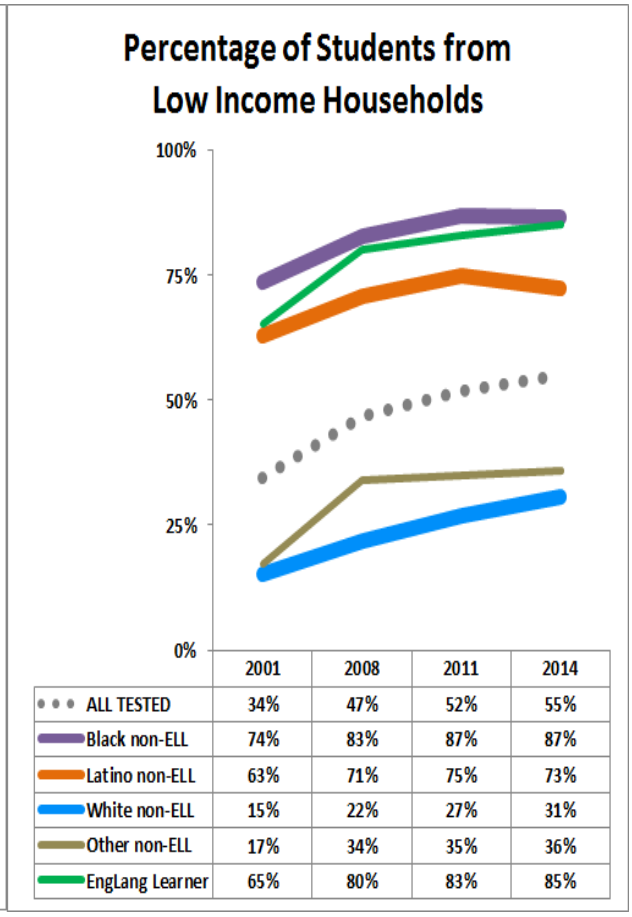
Figure 5.4

Declining Enrollments and Flattening Achievement among Higher Scoring White Students Accounted for Most of the Flattening in Statewide Reading Scores between 2011 and 2014





	2001	2008	2011	2014
ALL TESTED	100%	100%	100%	100%
Black non-ELL	24%	20%	18%	17%
Latino non-ELL	9%	11%	14%	12%
White non-ELL	60%	50%	48%	47%
Other non-ELL	1%	7%	7%	7%
EngLang Learner	6%	12%	12%	17%



	2001	2008	2011	2014
ALL TESTED	34%	47%	52%	55%
Black non-ELL	74%	83%	87%	87%
Latino non-ELL	63%	71%	75%	73%
White non-ELL	15%	22%	27%	31%
Other non-ELL	17%	34%	35%	36%
EngLang Learner	65%	80%	83%	85%

Source: Illinois State Board of Education ftp://ftp.isbe.net/SchoolReportCard/

Figure 5.4 illustrates that changes in statewide scores are produced by a complex mix of shifting demographic characteristics that are all closely associated with achievement. In this case, continuing enrollment declines and flattening achievement among non-ELL White students strongly influenced statewide medians because this group still represented close to 50% of the overall test population in 2011 and 2014. A 1-point gain in this group combined with gains and losses in the other four groups to produce a net effect of no change statewide.

Who's Who in Statewide Demographic Sub-Groups?

An unintended consequence of NCLB accountability requirements was that they reinforced racial and social class stereotypes by conflating things like race, family income and English language proficiency. For example, achievement reports for Latino students did not distinguish between students who were English-proficient and students temporarily identified as English Language Learners (ELL). This artificially depressed overall Latino achievement and under-reported non-ELL achievement, especially at lower grade levels. Similar problems made it impossible to draw meaningful conclusions about achievement among other racial groups because reports failed to account for different concentrations of free/reduced lunch eligibility.

The pie chart below shows how achievement and population size varied among different sub-groups of Latino third graders in 2014.

Achievement and Population Size among Latino Sub-Populations in 2014
Third Graders in Chicago and the Rest of Illinois

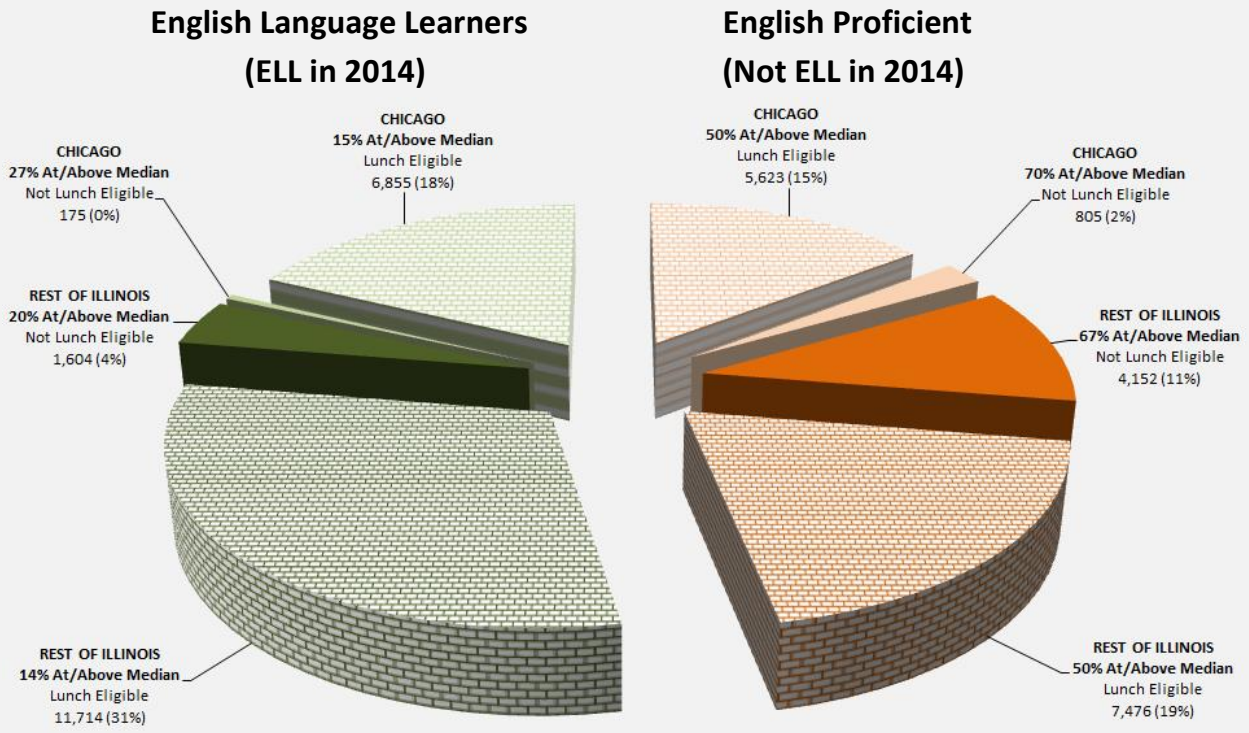


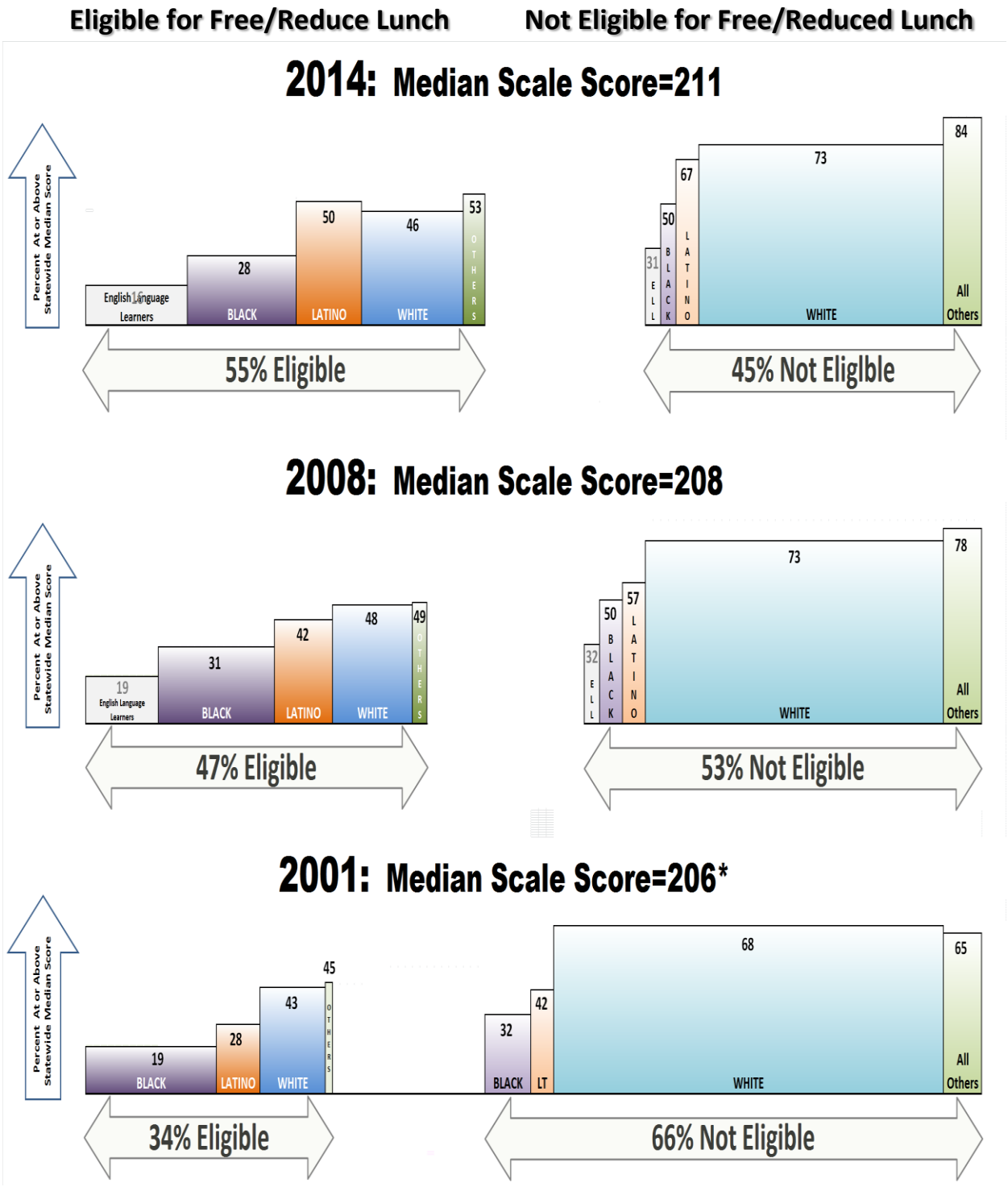
Figure 5.5 illustrates how changing enrollments and changing eligibility for free or reduced lunch in each of five student sub-groups contributed to overall changes in 3rd grade reading achievement between 2001 and 2014 statewide:

- Each bar represents a different demographic group; bars on the left show groups eligible for free or reduced lunch, bars on the right show groups not eligible for free or reduced lunch
- The width of each bar reflects the portion of the total test population that each group accounted for in 2001, 2008 and 2014
- Numbers at the top of each bar show the percentage of students in each group who scored at or above the statewide median score during each year shown

By controlling simultaneously for race, family income status, and proportional contributions to statewide scoring, the charts in Figure 5.5 make it more possible to draw defensible conclusions about changes in third grade reading achievement under NCLB. For example:

- The 5-point gain in median third grade reading scores between 2001 and 2014 occurred during a period when the state's highest achieving populations (students not eligible for free/reduced lunch) shrank dramatically from 66% to 45% of the total test population. Since shifts of this kind normally predict declines in overall achievement, Figure 5.5 offers good evidence that overall instructional effectiveness increased under NCLB.
- Between 2001 and 2008, there were big upward shifts in the percentage of non-ELL students in all sub-groups who scored at or above a rising statewide median for third grade reading. This offers clear evidence that, on average, instructional effectiveness was improving for most students during this period. For example, the percentage of non-ELL Black students who scored at or above the statewide median grew from 19% in 2001 to 31% in 2008 even though the statewide median also increased by the equivalent of two scale points during the same period.
- Between 2008 and 2014, achievement among non-ELL Latinos continued to grow while achievement within non-ELL Black and White populations either flattened or declined. As a result,
    - differences in achievement between non-ELL Latino and non-ELL White students narrowed continuously throughout the NCLB era; by 2014, non-ELL Latino students from low-income households were actually achieving at higher levels than their low income, White counterparts
    - differences in achievement between non-ELL Black and non-ELL White students that narrowed from 2001 to 2008 stayed more or less unchanged between 2008 and 2014

Figure 5.5
Overall Third Grade Reading Scores Continued to Rise Despite
Big Increases in Low-Income Enrollments and Big Declines in White Enrollments



*Scaling on the ISAT changed in 2006. For ease of comparison, actual ISAT medians from 2001 have been converted into 2006-2014 scale equivalents using equipercentile mapping of statewide scores from 2005 and 2006

SECTION 6

Turning Cut Scores into Standards

Under NCLB, the US Department of Education allowed states to create unaligned cut scores that purported to represent mastery of state standards. Cut scores played havoc with the meaning of test results, obscured important achievement trends and undercut public confidence in standardized testing as a whole. They also masked deep similarities in the scoring patterns that were produced by ISAT, NAEP and most standardized tests including new PARCC exams

Unlike earlier standardized tests that reported achievement in comparison with national norms, high-stakes testing under NCLB used cut scores to report achievement and growth over time. A key claim of cut scores was that they represented specific levels of mastery of clearly-articulated learning standards. That was a major departure from earlier reporting that simply compared results against whole-group norms using averages, percentiles and other statistical measures.

Cut scores work best when they predict valued outcomes in the real world. For example, ACT college readiness benchmarks are useful because they predict a 50% probability of getting a "B" or better in freshman level college courses and a 75% probability of obtaining a "C" or better. The Partnership for Assessment of College and Career Readiness (PARCC) has promised that Level 4 cut scores on PARCC exams will be based on a similar calculus.

The problem with cut scores under NCLB is that they created serious technical distortions and opened the door to reporting abuses that simply weren't possible in earlier, norm-referenced reports. So in principle, Illinois cut scores identified the place on statewide test scales where students demonstrated mastery of challenging state standards. And in principle, those cut scores were carefully aligned to track changes in standards-mastery as students moved from one grade level to the next.

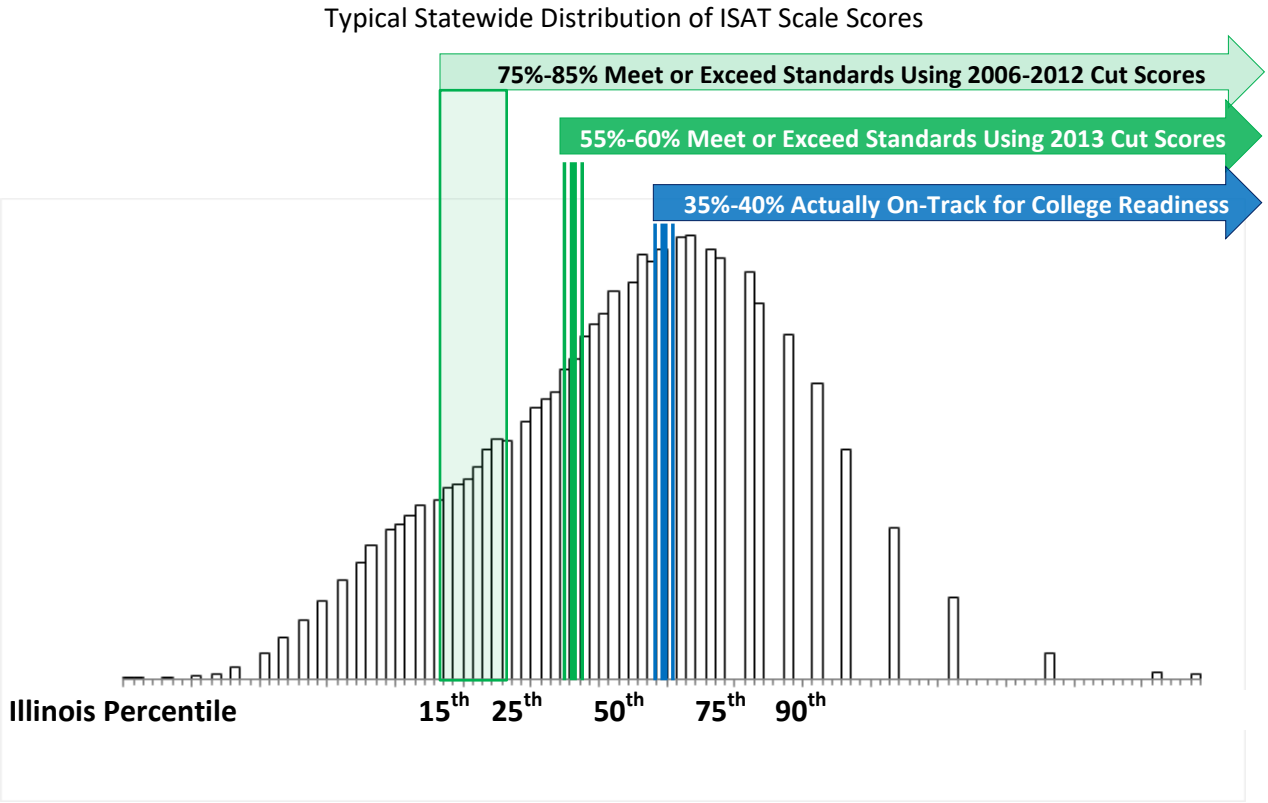
But In practice, Illinois cut scores met neither of these criteria. Instead, levels of skill and knowledge that were needed to meet Illinois cut scores:

- were largely unconnected to the full range of requirements contained in of Illinois State Learning Standards
- varied widely across grade levels and subject areas
- were set one to two years below grade level compared with state and national norms
- fell far below the skills and knowledge required to reach proficiency benchmarks on the National Assessment of Educational Progress (NAEP) or college readiness benchmarks on the ACT

Figure 6.1 shows what typical scoring distributions looked like on the ISAT from 1999 through 2014. It illustrates how different cut scores can report radically different test results from ***exactly the same data***.

Figure 6.1

Different Cut Scores Report Out Different Results from Exactly the Same Scoring Distribution



More Similar than Different

Unaligned cut scores under NCLB made it look like the ISAT assessed radically different forms of academic rigor than the NAEP or other widely used tests like the Measures of Academic Progress (MAP). Figure 6.1 illustrates that, from 2006 through 2012, 75 to 85 percent of elementary and middle school students met or exceeded state standards on the ISAT, but only 35 to 40 percent were on track to meet ACT college readiness benchmarks at the end of eleventh grade. On its face, this difference signaled that age-adjusted items and passages on the ISAT were markedly easier than comparable items and passages on the ACT. Most people drew a similar conclusion about the ISAT and the NAEP. If 80 to 85 percent of students met standards on the ISAT, but only 30 to 35 percent scored proficient or above on the NAEP, it seemed pretty clear that the NAEP was a tougher test.

A less obvious but more accurate explanation is that most tests used under NCLB produced very similar scoring distributions but were ***graded*** in very different ways. Locating cut scores at the lower end of the scoring distribution produced easier, more lenient grades. Locating cut scores higher on the distribution produced harder, more rigorous grades.

In fact, once cut scores are removed from the mix, achievement patterns on the ISAT, NAEP, ACT and most other standardized tests look remarkably similar. The same is true for recently published results from the 2015 PARCC exam. All of these tests predict each other's results with high levels of accuracy.

Figure 6.2 illustrates the close match between:

- percentages of students who scored at or above the statewide median on PARCC's third grade English/Language Arts exam; and,
- percentages of third graders who scored at or above statewide medians on ISAT reading exams from 2001 through 2014.

To improve measurement consistency over time, students temporarily identified as English Language Learners (ELL) have been removed from reported scores.

Figure 6.2

2015 PARCC Scores Closely Mirrored Historical Scoring Trends on the ISAT

Non-ELL Students Scoring at or Above Statewide Medians on 3rd Grade ISAT and PARCC Reading Exams

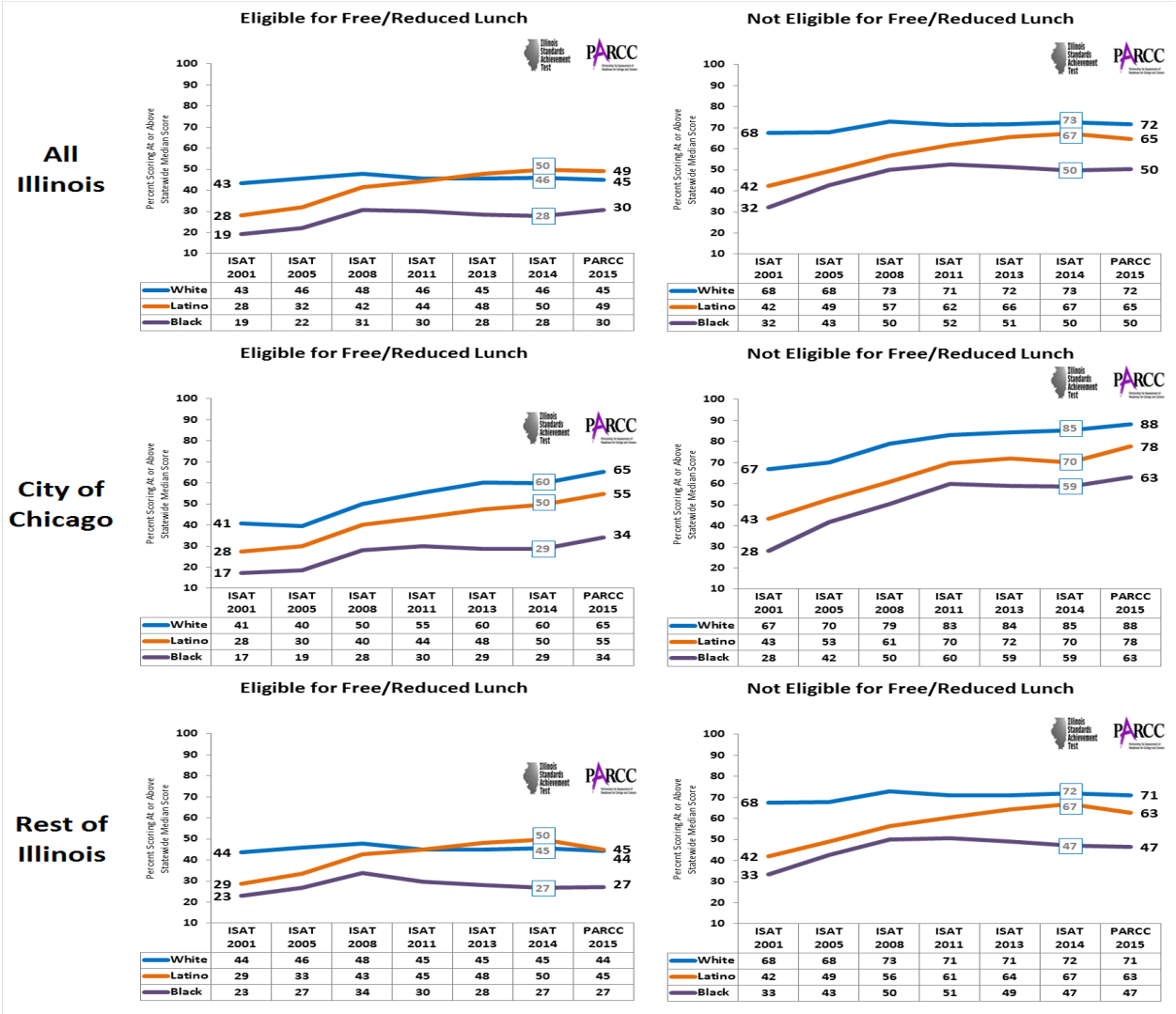


Figure 6.3 shows comparable matches between ISAT and PARCC results at other grades as well.

Figure 6.3

2015 PARCC Results for English/Language Arts Closely Matched ISAT Reading Trends

Non-ELL Students Scoring At or Above Statewide Medians in 3rd, 5th and 8th Grade: 2001-2015

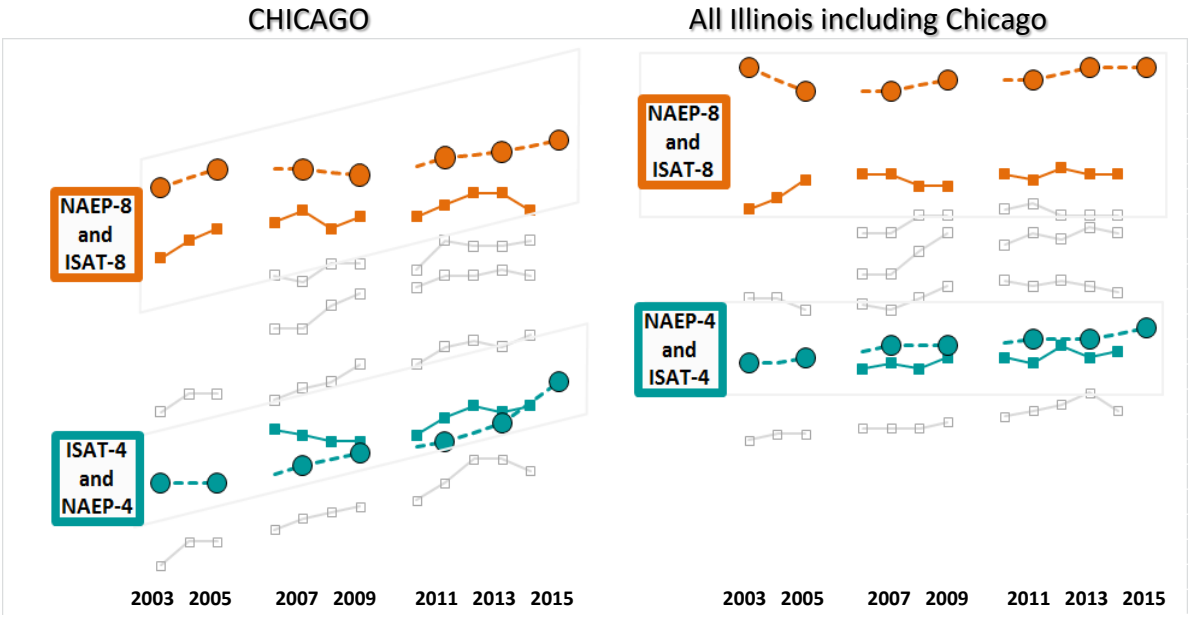
Changes over time in median scores on the ISAT and NAEP have also been very similar throughout the NCLB era. This explains why the top two sections of Figure 6.2 are dead ringers for the NAEP histories that were presented in Section 1.

Figure 6.4 shows how median reading scores on the NAEP and ISAT followed close-to-identical patterns of growth between 2003 and 2015. The blue and orange circles in Figure 6.4 show median scores on the NAEP while blue and orange squares show median scores on the ISAT. Gray squares show ISAT medians for grades 3, 5, 6 and 7 where the NAEP is not administered. For ease of comparison, 3rd, 5th and 8th grade ISAT scores for 2003 through 2005 have been converted to 2006-2014 scale values using equipercentile mapping of statewide scores from 2005 and 2006.

Figure 6.4

NAEP and ISAT Scoring Patterns Mirror Each Other in Chicago and Statewide

Median Scale Scores for NAEP and ISAT Reading



The NAEP and the ISAT Have Different Vertical Scales

It is important to note that NAEP and ISAT scales were ***developed independently and are not numerically comparable***. So the NAEP scores shown in Figure 6.3 are not "higher" than most ISAT scores in fourth grade and in eighth grade. If they were, it would imply that the NAEP was an easier, less demanding test than the ISAT.

The purpose of showing NAEP and ISAT medians using the same vertical scale is simply to highlight how closely their growth patterns match each other over time. This finding supports the view that ISAT results which are not filtered through arbitrary grading practices provide measures of growth that have roughly the same reliability as those produced by the NAEP.

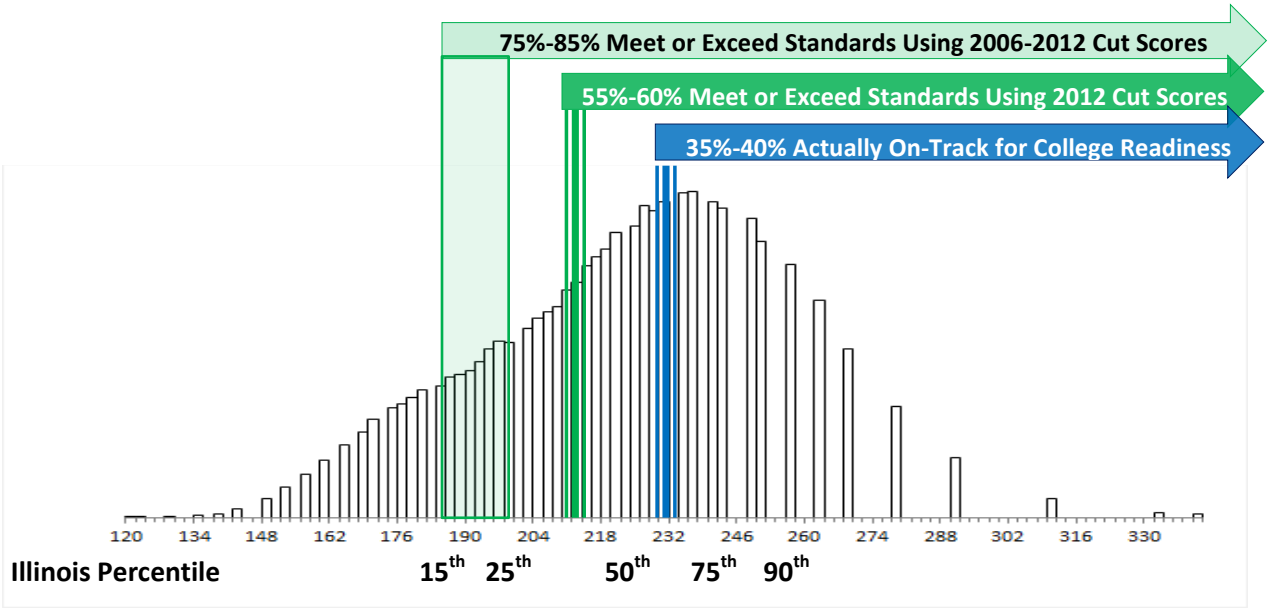
Figure 6.5 shows how percentile match-ups also make it possible to predict ACT scores based on earlier ISAT scores. Percentiles describe the percentage of all students tested who score at or below a particular score. Because, on average, scoring relative to other students is stable over time, scores on later tests can typically be predicted by matching up their percentile ranks with those of earlier scores.

By way of illustration, students who scored at or near 2006 cut scores on the ISAT were at the 15th to 25th percentile of the statewide scoring distribution. Students who score between the 15th and 25th percentiles on the ACT have scale scores of 15 and 16. If ISAT test results predict ACT test results, roughly the same percentage of 8th graders who score at or above the 15th to 25th percentiles on the ISAT will also score at or above 15 to 16 when then take the ACT three years later.

Figure 6.5

Matching Percentiles to Scale Scores Allows Earlier Test Scores to Predict Later Test Scores

Typical Statewide Distribution of ISAT Scale Scores



Typical Statewide Distribution of ACT Composite Scores

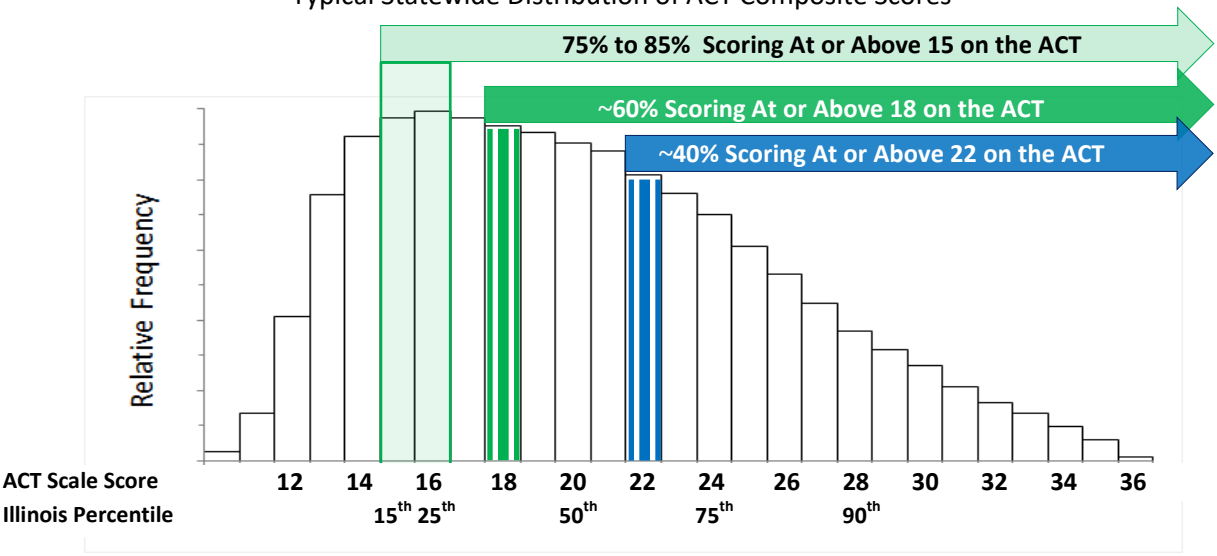


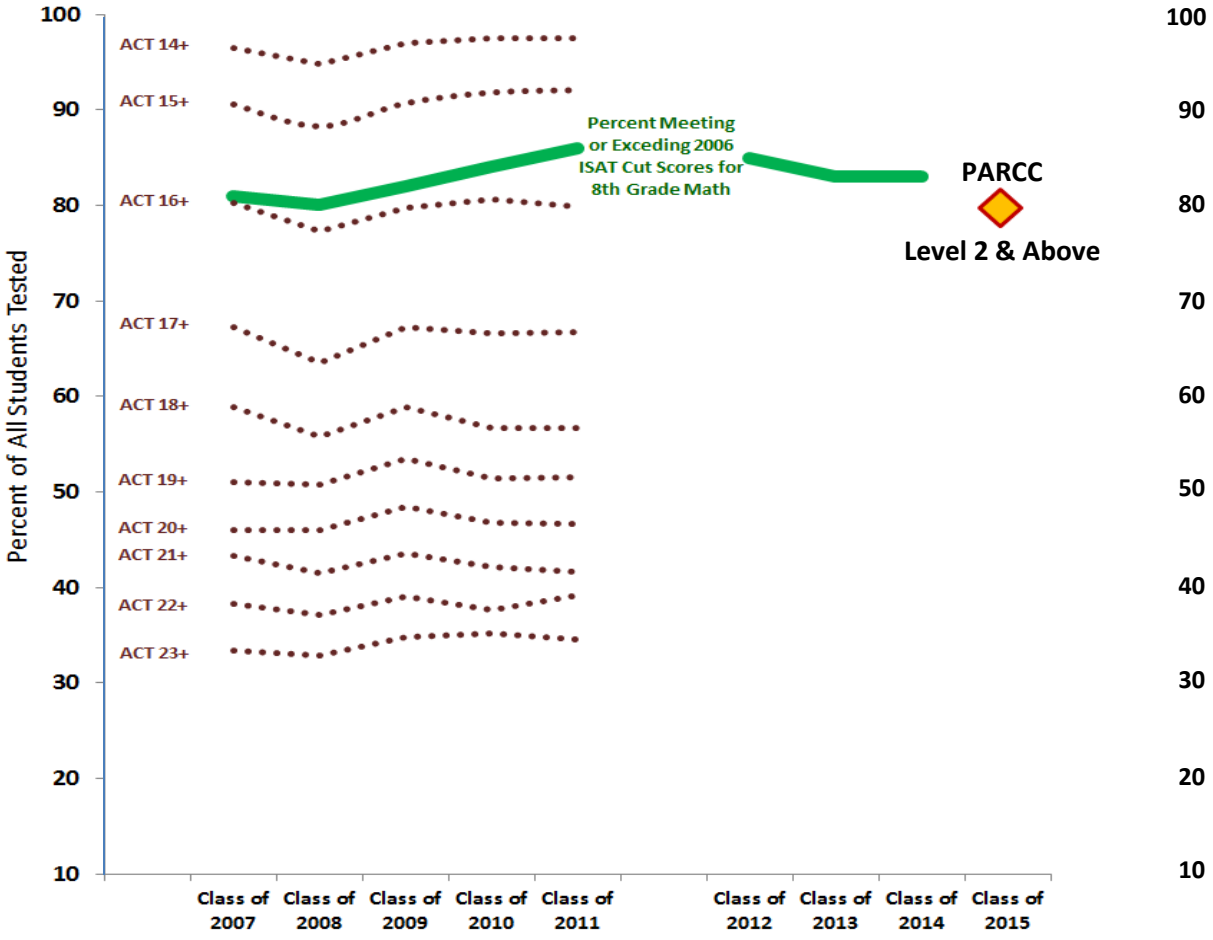
Figure 6.6 illustrates that this is exactly what happened statewide for five consecutive eighth grade graduating classes. The solid green line in Figure 6.6 shows the percentage of students in each graduating class who met or exceeded the 2006 cut score for 8th grade math on the ISAT. Dotted maroon lines show the percentage of students in each graduating class who scored at or above different ACT math scores three years later at the end of 11th grade. These lines show that roughly the same percentage of 8th graders who met or exceeded 2006 cut scores also scored at 15 to 16 or above on the ACT at the end of their junior year of high school.

The gold diamond in Figure 6.6 matches up percentages of 8th graders statewide who met or exceeded 2006 ISAT cut scores with statewide results from the 2015 PARCC math exam. It shows that 2006 ISAT cut scores fell just below the cut score that PARCC uses to separate Level 1 from Level 2 on its five-level proficiency scale. PARCC describes Level 1 as "Did Not Meet Expectations" for college and career readiness. It describes Level 2 as "Partially Met Expectations" for college and career readiness.

Figure 6.6

2006 Cut Scores for 8th Grade ISAT Math Roughly Equated to an 11th Grade ACT Scores of 15 On the 2015 PARCC Exam, They Equated to the Upper Edge PARCC's Lowest Proficiency Level

Connection between Scoring At/Above 2006 Cut Scores in 8th Grade ISAT Math and Later ACT Math Scores at the End of 11th Grade



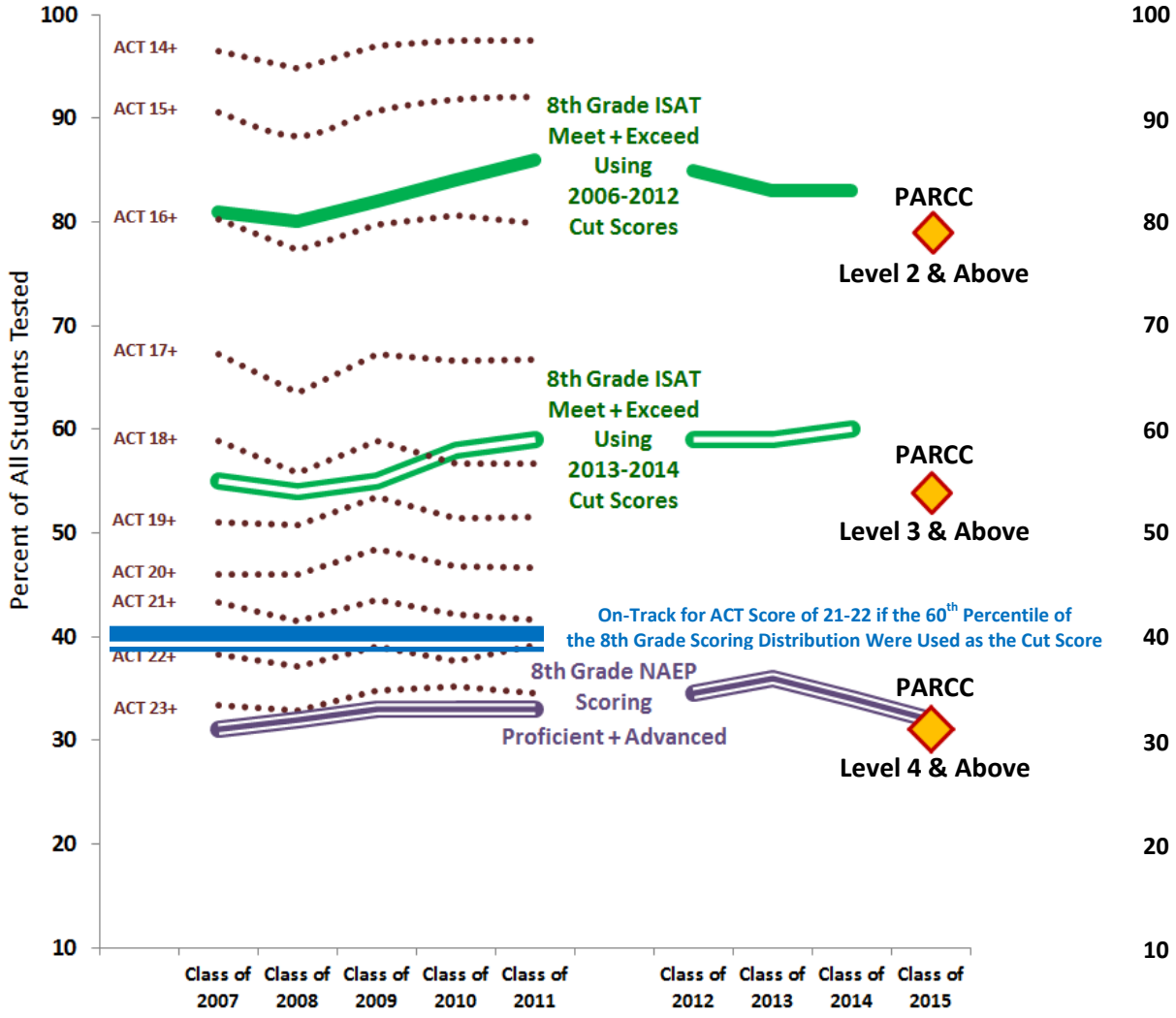
Below, Figure 6.7 expands Figure 6.6 by showing the connection between other 8th grade cut scores and later achievement on the ACT:

- The double green line in Figure 6.7 shows that about the same percentage of 8th graders who scored at or above 2013 cut scores on the ISAT also scored 17-18 or higher on the ACT.
- The purple line in Figure 6.7 shows that about the same percentage of 8th graders who scored proficient or above on the NAEP also scored 23 and above on the ACT
- The blue line in Figure 6.7 shows that a cut score set at the 60th percentile of statewide ISAT distributions would have predicted ACT scores of 21-22 and above with high levels of accuracy

Figure 6.7

8th Grade ISAT Math Scores Were Powerful Predictors of 11th Grade ACT Math Scores
NAEP Proficiency and Level 4 PARCC Proficiency Map Closely to a Score of 23 on the ACT

Statewide Scoring Percentages for Nine, Eighth Grade Graduating Classes: 2007-2015
ACT Scoring Percentages from 2010 through 2014 for 8th Grade Classes of 2007 through 2011



Finally, the orange diamonds in Figure 6.6 indicate that:

- Level 3 and above on PARCC math will likely equate to a score of 18-19 or above on the ACT
- Level 4 and above on PARCC math will likely equate to a score of 23 or above
- Level 4 and above on the 8th grade PARCC math exam roughly equates to proficient and above on the 8th grade NAEP math exam

The Medium is the Message

For most of its sixteen year history, and especially since 2006, low cut scores made the ISAT look like an easy, undemanding test that was based on easy, undemanding standards. In the summer of 2009, for example, a Chicago *Tribune* editorial wrote,

> *". . . we've known for some time now that nobody can put much faith in the ISAT. In 2006, state education officials significantly changed the test. Like magic, the test results took a leap.*
>
> *What really happened: Illinois responded to pressure from the federal No Child Left Behind law by deciding it was simpler to make the tests easier than make the kids smarter.*
>
> *. . . While Chicago students' scores on the dubious Illinois tests have jumped, by another measure -- the National Assessment of Educational Progress -- they have flatlined . . ."* [Chicago *Tribune* July 11, 2009]

The *Tribune* got it partly right when it said that pressure from No Child Left Behind led state officials and their partners in the testing industry to "dumb down the test." But Illinois Learning Standards didn't change in 2006, and test questions didn't change much either. There were plenty of challenging items on the new ISAT. You just didn't have to get any of them right to get a passing grade.

Scores jumped on the ISAT in 2006 because:

- 8th grade math cuts were intentionally lowered by more than a full grade level.
- A botched "equating study" that was supposed to equalize cut scores between old and new ISAT scales ended up inflating ISAT scale scores values by close to a full grade level; the result was that new ISAT cut scores which were supposed to equate with old cut scores were actually lowered by almost a full grade level

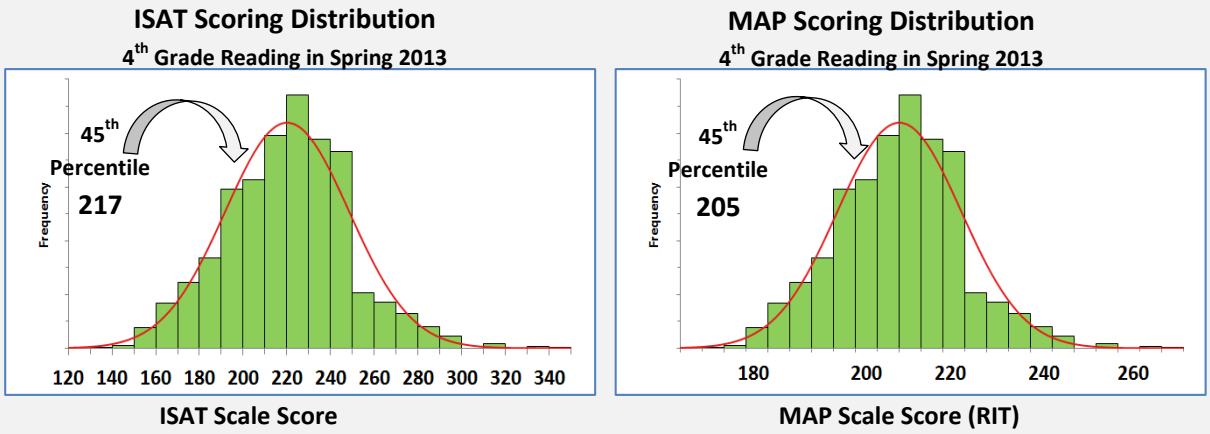
The net effect was that cut scores values plunged from the lower edge of grade level in 2005 to a year or more ***below*** grade level in 2006. The same thing happened in 2013, but this time in the other direction. Cut scores were changed with great fanfare in 2013 to "bring them in line with the Common Core's more rigorous standards" What actually happened was that cut scores moved back to the lower edge of grade level on statewide scoring distributions, just a few percentiles up from their original locations in 1999.

With all its imperfections, the ISAT had pretty much the same ability to assess and predict overall academic achievement as other more reputable tests like the NAEP, ACT, MAP, and more recently PARCC. But the cut scores Illinois used to finesse NCLB accountability requirements created an alternate universe of grading and reporting strategies that had no real connection with the standards they purported to represent. In the end, those strategies distorted what the ISAT actually assessed, undermined public trust and denied useful information to a whole generation of Illinois parents, educators and policy makers.

ISAT versus MAP: Not Much Difference Either

During the later years of NCLB, many districts used local funds to purchase assessments like the Measures of Academic Progress (MAP) to measure student growth and predict later performance on the ISAT and ACT. The process that made these predictions possible is called equipercentile mapping.

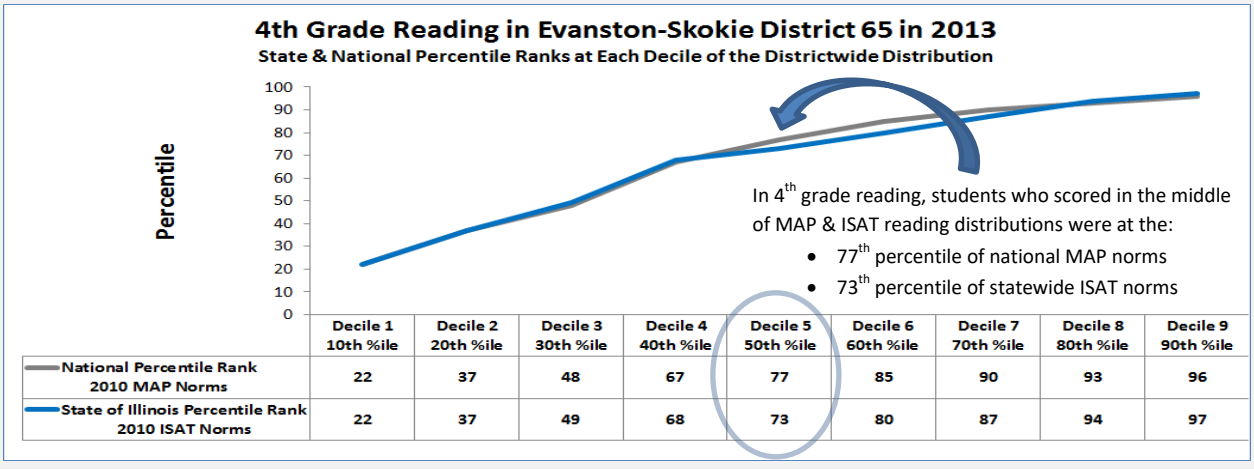
Equipercentile mapping aligns scale score values from two separate tests. It does that by matching up percentile ranks from the scoring distributions of students who took both tests. In 2013, the cut score for meeting standards on the fourth grade ISAT reading exam was raised to 217. The two charts below illustrate how equipercentile mapping was used to find the MAP score for fourth grade reading that had roughly the same "knowledge value" as a 217 on the ISAT reading test.



Illinois' Performance Evaluation Reform Act (PERA) tacitly encouraged districts to purchase tests like the MAP by prohibiting districts from using the ISAT to measure student growth. But once cut scores are removed from the mix, the MAP and ISAT growth measures become virtually indistinguishable.

In 2013, UIC's Center for Urban Education Leadership asked the senior leadership of Evanston-Skokie District 65 for permission to access four years of student-level roster files for all 6,500 students in the district who had taken both MAP and ISAT exams. The district generously provided these data after removing all information regarding the identities of individual students and the schools they attended.

The chart below shows that percentile ranks at every decile of MAP and ISAT scoring distributions were close to identical.



	Decile 1 10th %ile	Decile 2 20th %ile	Decile 3 30th %ile	Decile 4 40th %ile	Decile 5 50th %ile	Decile 6 60th %ile	Decile 7 70th %ile	Decile 8 80th %ile	Decile 9 90th %ile
National Percentile Rank 2010 MAP Norms	22	37	48	67	77	85	90	93	96
State of Illinois Percentile Rank 2010 ISAT Norms	22	37	49	68	73	80	87	94	97

SECTION 7

Inside the Black Box: What Do Standardized Tests Actually Measure?

The rhetoric of standards-based assessment convinced most parents, educators and policy makers that different tests produced different results because they were based on different standards. Higher scores signaled easier standards; lower scores signaled more demanding standards. But assessment professionals have long known that different tests based on different standards often produce close to identical results.

A year after the passage of No Child Left Behind, the Consortium on Chicago School Research compared results from the Iowa Test of Basic Skills (ITBS) with the still-new Illinois Standards Achievement Test. This study highlighted a number of factors that distinguished the "standards-based" ISAT from the older, norm-referenced ITBS. But the study concluded by saying,

> *In spite of large content and format differences, the ITBS and ISAT behave similarly among CPS students. There scores are highly correlated and their trends over time are mostly parallel. In the one case where the trends run counter to each other, the trends are converging and will be parallel in another year or two." [Easton, et.al. (2003) p. 19]*

Tucked into a brief insert at the end of the study, the authors also presented an important caveat. Citing conflicting conclusions from two federal studies about the comparability of different tests, the insert said,

> *The correlation between two tests depends on several factors: the reliability of the tests, the similarity of content and format, and the instructional experiences of the students who take them. When the correlations between tests are high, it is possible to predict performance on one from the other.* ***This does not mean that you can interpret the results of one test in terms of the content of the second****. [Easton et.al. (2003) p.18 emphasis added]*

Anyone who doubts that tests of different content knowledge can produce remarkably similar results need only look at the close correlation among different sub-scores on the high school ACT. Figure 7.1 excerpts a table from ACT's Technical Manual (2007). On the low end, math and reading sub-tests match up 64% of the time. One the high end, English and reading match up 78% of the time.

Figure 7.1

Median Correlations Among Test Scale Scores
for the Six National ACT Administrations in 2005–2006

	English	Mathematics	Reading	Science
English	1.00	.70	.78	.70
Mathematics		1.00	.64	.75
Reading			1.00	.69
Science				1.00

Source: *ACT Technical Manual* (2007), ACT, Inc. p. 61

If matching up two-thirds to three-quarters of the time still seems like a modest correlation, the data shown in Figure 7.2 offer more evidence for how powerful this connection actually is. Figure 7.2 tracks the consistency of test results ***within each subject*** area from the ACT/PLAN exam at the beginning of tenth grade to the full ACT at the end of 11th grade. These data show that scoring over time in the ***same subject area*** is consistent between 67% and 81% of the time . . . not much different from the 64% to 78% consistency that Figure 7.1 shows for results across subjects.

Figure 7.2

Correlation Coefficients
Among ACT Scores and PLAN Scores
(Based on data pooled over high schools, N = 403,381)

PLAN score	ACT score				
	English	Mathematics	Reading	Science	Composite
English	.81	.65	.73	.67	.80
Mathematics	.67	.82	.61	.72	.78
Reading	.68	.56	.71	.61	.72
Science	.62	.65	.60	.67	.71
Composite	.82	.78	.78	.78	.88

Source: *ACT Technical Manual* (2007), ACT, Inc. p. 70

Figure 7.2 shows that consistency across subjects also doesn't deteriorate much over time. The correlation between math and reading is 0.64 on the ACT, but only a slightly lower 0.56 between the 10th grade PLAN and the 11th grade ACT. The correlation between English and reading is 0.73 between the PLAN and ACT but rises only marginally to 0.78 on the ACT itself.

What's Going On?

When very different assessments of very different content produce very similar results, something important is missing from the story. Ironically, the missing piece is a constellation of factors that standardized tests have been roundly criticized for ignoring throughout the NCLB era . . . rigorous content, critical thinking and depth of knowledge.

The data outlined in Section 6, and in many prior studies, show that the ISAT reliably predicted scoring patterns on tests like the NAEP and ACT year after year after year. That's because, regardless of content, items and passages on the ISAT, NAEP, ACT and most other standardized tests are all intentionally designed to produce normal, "bell-curve" distributions. And the gatekeeper for obtaining higher scores on bell-curve distributions has less to do with particular skills and content than with the ***level of difficulty*** that selected skills and content represent.

The most persistent message about testing during the NCLB era has been that scoring on standards-based assessments is based on mastery of specific content. But a decade and a half of standards-based test results tell a very different story. Scoring on the ISAT and most other standardized tests was heavily determined by something else that produced similar test results across tests and across content areas. That something else is what assessment professionals euphemistically describe as "general knowledge."

Consider, for example, this recent finding from a study of middle school factors that predict success in Chicago public high schools:

> *. . . A student's score on either the reading or the math ISAT is a very good indicator of whether he or she will meet the college-readiness benchmark on any of the four subject-area tests on the PLAN. Combining the reading and math ISAT scores together, or combining ISAT scores from multiple years (e.g. students' seventh-grade score with their eighth-grade score), improves the prediction of students' PLAN score on any subject-specific test and on the composite score (the average of all the subject-specific tests). The subject-specific tests are each very predictive of scores in other subjects, almost as predictive as tests within the same subject.* ***This suggests that both the ISAT and the PLAN tests are measuring general knowledge and skills as much as knowledge and skills in any given subject area . .*** . [emphasis added]
>
> Allensworth, Elaine M. et.al. (2014). *Middle Grade Indicators of Success in CPS High Schools*, pp. 93-4

In earlier years, testing professionals had a different name for general knowledge. They used to call it aptitude. Before that, they called it I.Q.

What Does "General Knowledge" Look Like?

Prior to the revolution in neurology and learning science that began in the last quarter of the twentieth century, the consensus view in American social science was that differences in standardized test outcomes reflected inherent and largely immutable differences in student ability. As late as 1994, Richard Hernstein and Charles Murray, both prominent members of the social science community, continued to argue this position in *The Bell Curve*.

While it still seems likely that individual traits and dispositions play a role in standardized test outcomes, the major lines of evidence that were used in *The Bell Curve* have now been widely discredited. In their place, a large and growing body of evidence has developed about the plasticity of human neurology and the substantial impact that teaching and schooling can have on conventional measures of aptitude and intelligence.

No Child Left Behind was predicated on the assumption that 100% of the American students could be taught to "reach . . . proficiency on challenging state academic achievement standards and state academic assessments." But the tools that NCLB used to drive that effort relied almost exclusively on progressive sanctions and high-stakes accountability. Deeper understanding of how tests actually measured academic challenge got short shrift. Just how short that shrift was is the focus of Section 8.

Contrary to stereotype, standardized tests are intentionally imprecise. Like polls before an election, they sample what students know, assign probabilities and margins of error to their findings, and report back their results to interested consumers. Their most important job is to estimate the depth and breadth of students' academic strengths, and to identify where that estimate fits on a standardized continuum of academic capacities. Numerical scales are the yardsticks used to represent that continuum. Scale scores are the "units of knowledge" that make up the yardstick.

As outlined earlier, higher scale scores have at least as much to do with the depth and breadth of student thinking as they do with the volume of discrete skills and concepts that students have mastered. For the most part, students who are able to size up and work through items and passages that reflect higher levels of depth and complexity earn higher scale scores than students who get stumped by those items. None of this keeps standardized tests from providing useful diagnostic information about what students know and are able to do. But the diagnostics they produce say more about higher-order thinking and depth of knowledge than they do about mastery of discrete skills and curricular content.

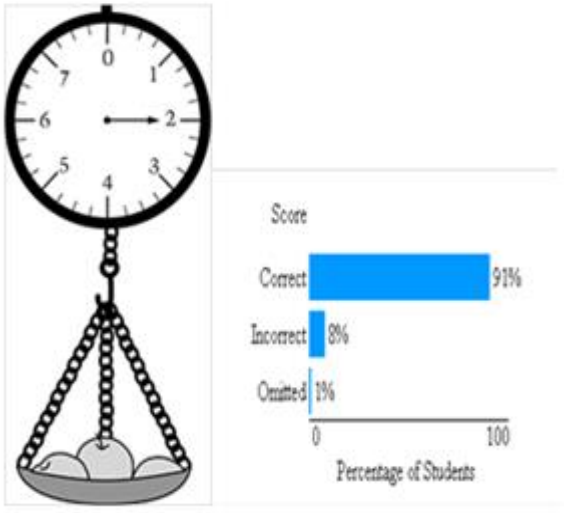
Figure 7.3 shows two fourth grade math items from the 2013 National Assessment of Educational Progress. Item #1 requires students to select the type of measurement that best matches what is going on in the picture. Item #2 requires students to read and interpret a thermometer scale that is calibrated in 2-degree units.

Figure 7.3
Standardized Test Items Use Specific Skills to Sample Broader Ability to Size Up and Work Through Academic Content at Different Levels of Academic Complexity

Math Item from the 2013 NAEP with an "Advanced" Scale Score Rating of 287

ITEM #1

Scale Score	NAEP Proficiency Level
157	BELOW BASIC

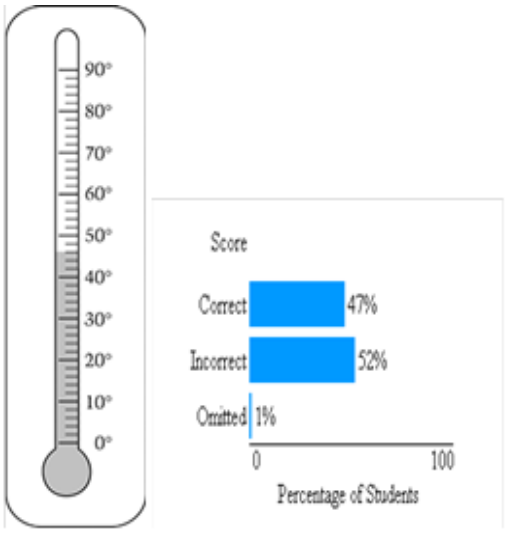


How much do these apples weigh?

- **A. 2 cups**
- **B. 2 feet**
- **C. 2 pounds**
- **D. 2 quarts**

ITEM #2

Scale Score	NAEP Proficiency Level
287	ADVANCED



What temperature does the thermometer show?

- **E. 43°**
- **F. 46°**
- **G. 52°**
- **H. 54°**

Source: NAEP website http://nces.ed.gov/nationsreportcard/itemmaps/

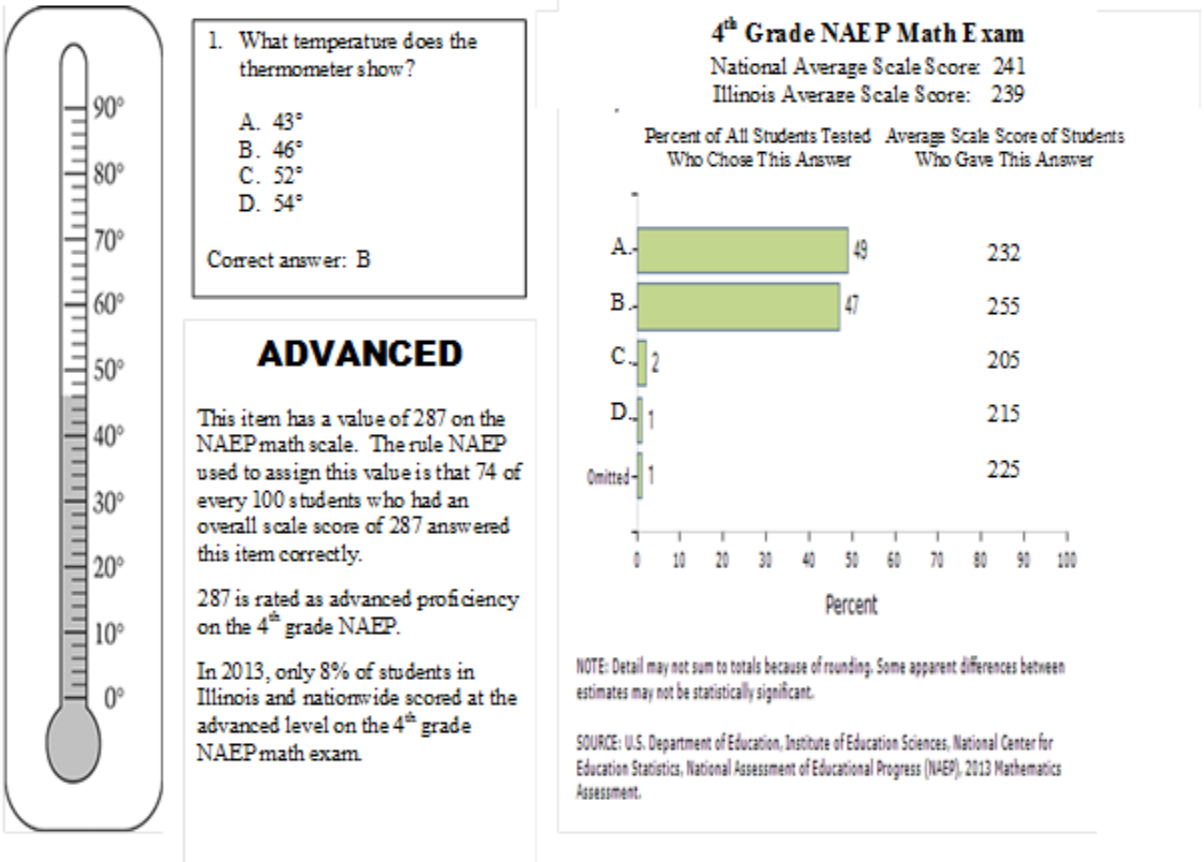
In Figure 7.3, Item #2 is clearly more demanding than Item #1. In 2013, 92 percent of American fourth graders got Item #1 right, while only 47 percent answered Item #2 correctly. How come?

In one respect, both items are criterion-referenced measures that simply require mastery of different discrete skills.

- Item #1 assesses if a student can:
  - read and understand the question
  - recognize the difference between "weigh" and other forms of measure
  - recognize that G. is the only response that measures weight
- Item#2 assesses if a student can:
  - read and understand the question
  - recognize that the gray-filled portion of the picture is a measure of temperature
  - recognize that the numbers on the thermometer represent a graduated scale of temperatures that increases from bottom to top
  - recognize that the numbers in the graduated scale are calibrated in 2-degree increments

But the last bullet in Item #2's list calls for a different kind of thinking than Item #1 does. Figure 7.4 illustrates that most of the students who failed to select the correct answer for Item #2 did so because they did not attend carefully enough to the scale to ***infer*** that each tick represents two degrees instead of one. In fact, almost half of all 4th graders tested made the more literal judgment that 1 tick mark equaled 1 degree and chose 43° as the correct answer.

Figure 7.4
Choosing the Correct Answer for This Item Required
Inferential Reasoning and a Rudimentary Understanding of Ratio and Proportion



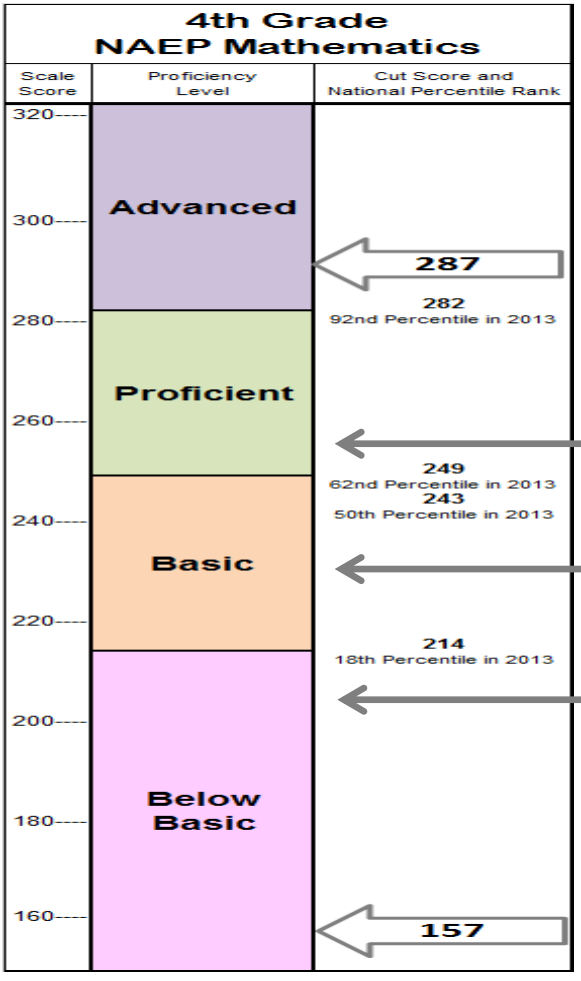
Source: NAEP website http://nces.ed.gov/nationsreportcard/itemmaps/

While a number of discrete mathematical skills were needed to avoid this error (i.e. 50°- 40° =10°; 10 ÷ 5=2°; 3 x 2° = 6°; 40°+ 6° = 46°), getting the correct answer involved a lot more than just executing those discrete skills and carrying them out in the proper sequence. Basic inferential reasoning and rudimentary understand of ratio and proportion were required to size up what the problem required and notice details that made a literal answer like 43° look problematic.

Contrary to stereotype, inferential reasoning and conceptual understanding are central requirements for achieving higher scale scores on virtually all standardized tests. In 2013, for example, students who correctly answered questions like Item #2 three quarters of the time typically scored at or above the 92nd percentile on the 4th grade NAEP math scale.

The role that inferential reasoning and conceptual understanding play in obtaining higher scale scores on tests like the NAEP is further illustrated in Figure 7.5. Fourth graders who answered the item in Figure 7.4 correctly had an average overall scale score of 255 in 2013. This is just above the cut score for "Proficient" on NAEP's 4th grade math scale. By contrast, students who selected the more literal and incorrect answer of 43° had an average overall scale score of 232. This placed them squarely in the middle of "Basic" on NAEP's 4th grade math scale, just seven points higher than the average score of students who simply omitted the problem altogether (see chart data in Figure 7.4).

Figure 7.5
Standardized Test Scales Reflect Mathematical Probabilities Based on Correct-Response Frequencies from Large, Representative Samples of Typical Test Takers



287
The level of difficulty that NAEP assigned to Item #2
Decision Rule: 74% of students who had an overall scale score of 287 answered Item#2 correctly.

255
Average scale score for 4th graders who correctly chose 46° as the answer for Item #2

232
Average scale score for 4th graders who incorrectly chose 43° as the answer for Item #2

205
Average scale score for 4th graders who incorrectly chose 52° as the answer for Item #2

157
The level of difficulty that NAEP assigned to Item #2
Decision Rule: 74% of students who had an overall scale score of 157 answered Item#2 correctly.

Of course, none of the information described above explains why Item #1 has a value of exactly 157 on the NAEP math scale, or why Item #2 is valued exactly 130 points higher at 287. That explanation lies deep inside the black box of statistical relationships that let test makers assign equated knowledge values to items and passages on standardized tests. These values are based on the correct-response frequencies that items produce when they are administered to large, representative samples of typical test takers. The full range of scale scores shown in Figure 7.5 reflects the hierarchy of abstract knowledge values that are produced by this process.

What makes the inner workings of standardized assessment practices so impenetrable for all but a small number of testing experts is that individual elements of the system are mostly defined by the statistical relationship they have with all the other elements of the system. For the two items described in this section, Item #1 maps to a scale score of 157 because fourth-grade students with an overall scale score of 157 had a 74 percent chance of answering this question correctly. Fourth graders with an overall scale score of 287 had a 74 percent chance of answering the Item #2 correctly. And the overall scale that 157 and 287 are a part of is defined by the wider range of probabilities which are produced by the testing system as a whole.

The core message here is that scale scores are statistical abstractions. These abstractions do not assess specific skills and content knowledge. They assess the probability test takers have of being able to respond successfully to different ***types*** of skill, content, and ways of academic thinking. This information makes it possible to rank student proficiencies along a continuum of academic difficulty. In turn, this ranking creates a reliable predictor of future performance within normal margins of error.

The Rap on Standardized Testing

"All They Measure is Rote Basic Skills"

From: FairTest (2012) "What's Wrong with Standardized Tests?"
http://www.fairtest.org/facts/whatwron.htm

Are standardized tests fair and helpful evaluation tools?

Not really. On standardized exams, all test takers answer the same questions under the same conditions usually in multiple-choice format. Such tests reward quick answers to superficial questions. They do not measure the ability to think deeply or creatively in any field.

Do multiple-choice or short-answer tests measure important student achievement?

These kinds of tests are very poor yardsticks of student learning. They are weak measures of the ability to comprehend complex material, write, apply math, understand scientific methods or reasoning, or grasp social science concepts. Nor do they adequately measure thinking skills or assess what people can do on real-world tasks.

* * * * * * *

The late Grant Wiggins was a staunch advocate of progressive curriculum and assessment reform and a vocal critic of rote teaching and learning. *Understanding by Design*, the text he first co-authored with Jay McTighe in 1998, is something like sacred text for educators committed to making deep learning accessible for all students regardless of race, family income or zip code.

In March 2010, Wiggins surprised many of his followers by publishing an article in *Educational Leadership* called, "Why We Should Stop Bashing State Tests." In preparation for that article, Wiggins reviewed a wide range of released items and response frequencies from NCLB-era tests in Florida, Ohio and Massachusetts.

These following excerpts from page 51 describe what Wiggins found.

. . . most of the questions on the math and language arts tests are both appropriate and revealing—especially those that involve inferences about such key concepts as main idea, author purpose, linear relationships, equivalency of fractions and decimals, and so on . . .

A close look at state test results shows me that both test-prep "teaching" and test bashing get it wrong. The test items that our students do most poorly on demand interpretation and transfer, not rote learning and recall . . .

The surprising implication of Wiggins' analysis is that standardized testing doesn't have to be the Darth Vader of progressive school reform. Released test items and full reports of student responses can actually deepen the way we think about teaching and learning in ways that other forms of assessment cannot. They can also give us better insights about how to improve local assessment practices in ways that directly support deeper, more authentic forms of student and adult learning.

EDUCATIONAL LEADERSHIP

March 2010 | Volume **67** | Number **6**
Reading to Learn Pages 48-52
Special Topic / Why We Should Stop Bashing State Tests
Grant Wiggins

An item-by-item look at state test results reveals that students lack higher-level reading and thinking skills.

It is, of course, a common lament: "Oh, those standardized tests! If it weren't for them …" But if you look closely at the released test items and student performance data for states that provide such information, your opinion may change. Mine did. Standardized tests can give us surprisingly valuable and counterintuitive insights into what our students are *not* learning.

The myth is that the tests demand and reward low-level "coverage." The results say otherwise. Consider this item from the 2008 Massachusetts 10th grade English test, which involves the lyrics of a Bob Dylan song dear to me as a child of the '60s and as a musician. The student sees all the lyrics of the song, and then responds to this question: *Based on "The Times They Are A-Changin'," why does the speaker most likely single out "senators, congressmen" and "mothers and fathers"?* Here are the four choices:

A. They understand the problems of society.
B. They represent an outdated set of values.
C. They are the most open to change.
D. They are role models for the speaker.

Well, we "better start swimmin' or [we'll] sink like a stone" in education—because only 58 percent of students chose the correct answer, B. Astonishingly, 19 percent chose A; 12 percent chose C; and 11 percent chose D. In other words, more than 40 percent of 10th graders think the lyrics mean the *opposite* of what they really do. It seems that a huge chunk of our students cannot even make the most basic sense of a biting song lyric.

http://www.ascd.org/publications/educational-leadership/mar10/vol67/num06/Why-We-Should-Stop-Bashing-State-Tests.aspx

SECTION 8

Morphing Standards into Skills

Tests like the ISAT and ACT were well equipped to measure instructional impact on general knowledge, but poorly designed to return standard-specific information to teachers and parents. Test makers finessed this problem by inventing "content strands" and "power standards" that purported to measure mastery of specific standards. They did that knowing full well that standardized test items almost always measure more than one standard at a time, and are less about specific skills than about students' ability to handle different kinds of academic complexity.

Filling the demand for granular, diagnostic information

The tension between what standardized tests actually assessed and what policy makers wanted them to assess has dogged reporting practices throughout the NCLB era. High-stakes accountability created huge demand for granular, diagnostic information that could help schools improve achievement on standardized tests. To meet this demand, the testing industry stretched, and often broke, the boundaries of ethical reportage by inventing reporting gimmicks like "content strands" and "power standards."

Like many state-level testing systems, the ISAT used content strands report standards mastery in broad categories like "main idea," "supporting details," "number sense" and "measurement." Mastery levels for each strand were reported out as the number and percentage of correct answers that students earned in each strand.

Figure 8.1 shows how content strands on the ISAT were publicly reported:

- From left to right, bars on the chart show the percentage of correct answers for all math items tested followed by break-outs of percentages for each of five content strands; green bars show district-level percentages and red bars show statewide percentages
- The table below the chart show the total number of correct answers that were possible in each content strand followed by the average number of correct answers that students earned in each strand at the district and statewide levels
- The descriptions at the bottom summarize the Illinois Learning Standards that were ostensibly measured by each strand; links to specific standards (6A, 6B, 6C, etc.) provided access to more detailed descriptions of each standard.

On their face, content strands appeared to provide detailed, standards-based diagnostics about what students knew and were able to do in each strand. But closer examination reveals a host of troubling problems. The clearest evidence of these was that success rates didn't change much from one content strand to the next. They mostly just reported small variations on an overall score.

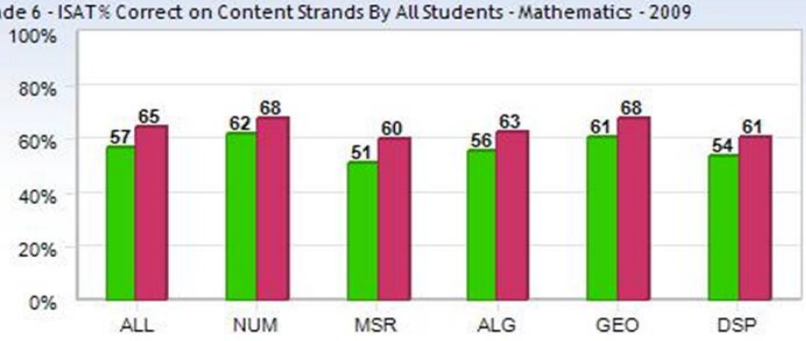
In Figure 8.1, for example, green bars show that there was only an 11 percentage point difference between the highest-scoring strand (NUM) in Chicago and the lowest-scoring strand (MSR). Statewide, that difference dropped to eight points. Normal statistical error rates weren't reported with content strands. If they had been, most differences would have disappeared entirely. Remarkably, this pattern showed up every year, in every subject, and at all grade levels tested from 2001 through 2013. The State of Illinois finally stopped reporting content strands in 2014.

Figure 8.1

ISAT Content Strands Made It Look Like Precise Diagnostic Information Was Being Reported about Mastery of Illinois Learning Standards

CITY OF CHICAGO SD 299

Grade 6 - ISAT% Correct on Content Strands By All Students - Mathematics - 2009



	ALL	NUM	MSR	ALG	GEO	DSP
Points Possible	65	16	10	17	12	10
(N) Correct Dist	37.3	9.9	5.1	9.5	7.4	5.4
(N) Correct State	42	10.9	6	10.8	8.2	6.1

Content Strand	Description	Standards
Number Sense	Demonstrate and apply a knowledge and sense of numbers, including numeration and operations (addition, subtraction, multiplication, division), patterns, ratios and proportions.	6A, 6B, 6C
Measurement	Estimate, make and use measurements of objects, quantities and relationships and determine acceptable levels of accuracy.	7A, 7B, 7C
Algebra	Use algebraic and analytical methods to identify and describe patterns and relations in data, solve problems and predict results.	8A, 8B, 8C, 8D
Geometry	Use geometric methods to analyze, categorize and draw conclusions about points, lines, planes and space.	9A, 9B, 9C
Data, Statistics, Probability	Collect, organize and analyze data using statistical methods; predict results; and interpret uncertainty using concepts of probability.	10A, 10B, 10C

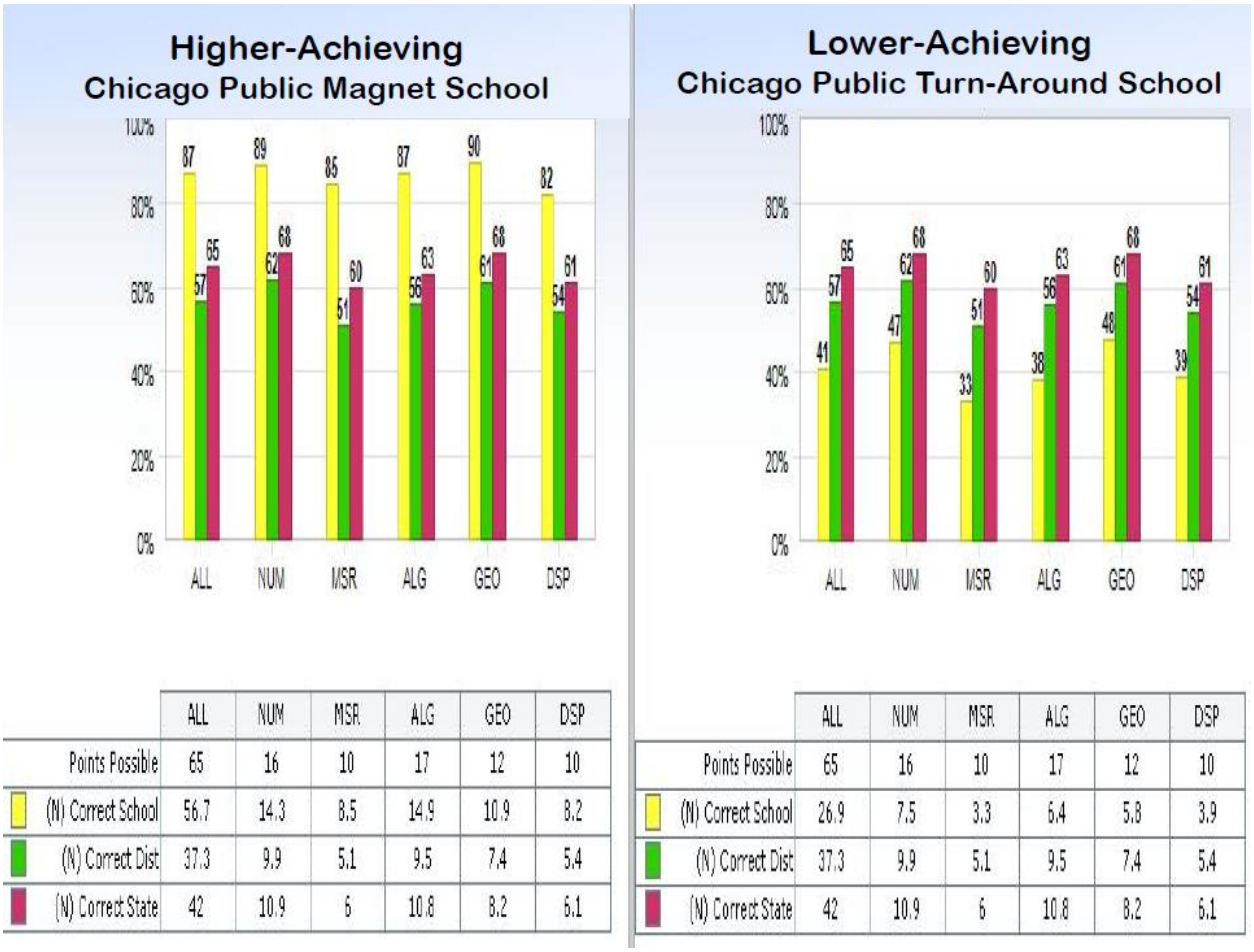
Source: http://iirc.edu **Retrieved May 15, 2012; for updated version see:**
http://iirc.niu.edu/Classic/District.aspx?source=ISAT&source2=ContentStrands&districtID=15016299025&level=D

Content strands didn't just fail to report meaningful differences in teaching and learning across strands. They also rose and fell together in lock-step regardless of the school or district they purported to measure. Statisticians call this "covariation."

Covariation is a problem for any measure that purports to be independent. A key part of what makes a measure "independent" is that it measures totally different things than other independent measures. For content strands to have diagnostic value, they needed to measure skills separately from one another, not rise and fall together in close synchrony across hundreds of elementary and middle schools statewide. But they did, in every subject and at every grade level tested, for thirteen straight years.

The two charts in Figure 8.2 show what this dance of numbers looked like across four very different test populations. The yellow bars in chart on the left show strand information from a high-achieving Chicago public magnet school. Yellow bars in the chart on the right show comparable information from a low-achieving Chicago neighborhood school that had recently been closed and re-opened as a district turnaround school. The green and red bars show the same district and state data that is shown in Figure 8.1

Figure 8.2
ISAT Content Strands Rose and Fell Together in
Exactly the Same Way Regardless of Unit Size or Overall Level of Achievement
6th Grade ISAT Math Achievement in 2009 at the School, District and State Levels



	ALL	NUM	MSR	ALG	GEO	DSP
Points Possible	65	16	10	17	12	10
(N) Correct School	56.7	14.3	8.5	14.9	10.9	8.2
(N) Correct Dist	37.3	9.9	5.1	9.5	7.4	5.4
(N) Correct State	42	10.9	6	10.8	8.2	6.1

	ALL	NUM	MSR	ALG	GEO	DSP
Points Possible	65	16	10	17	12	10
(N) Correct School	26.9	7.5	3.3	6.4	5.8	3.9
(N) Correct Dist	37.3	9.9	5.1	9.5	7.4	5.4
(N) Correct State	42	10.9	6	10.8	8.2	6.1

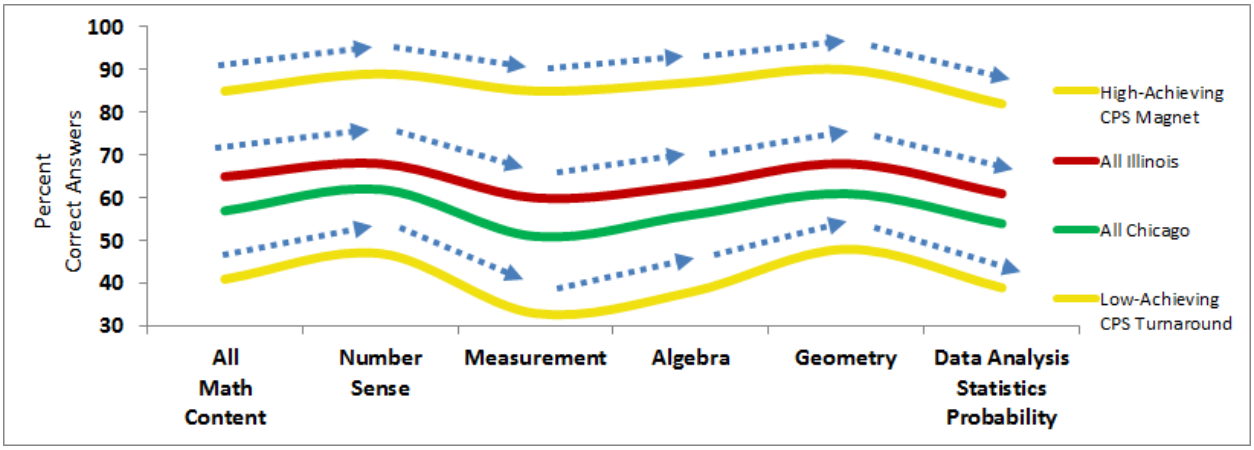
Source: http://iirc.edu **Retrieved May 15, 2012; for updated version see:**
http://iirc.niu.edu/Classic/District.aspx?source=ISAT&source2=ContentStrands&districtID=15016299025&level=D

Move your fingertip from left to right across the tops of content strands for each test population. As you do, notice how the same "m-shaped" wave takes shape for each population. Figure 8.3 shows just how closely these waves match up by removing the bars and showing only the waves. In 2009, this same m-shaped pattern was reflected in test results for every one of 1,000+ elementary and middle schools statewide that gave sixth grade math tests. Different subjects and grades produced different shapes. But lock-step covariation was a stable characteristic of ISAT content strands in every subject and at every grade level tested during every year the ISAT was administered.

Figure 8.3

ISAT Content Strands Rose and Fell Together in Exactly the Same Way Regardless of Unit Size or Overall Level of Achievement

6th Grade ISAT Math Achievement in 2009 at the School, District and State Levels



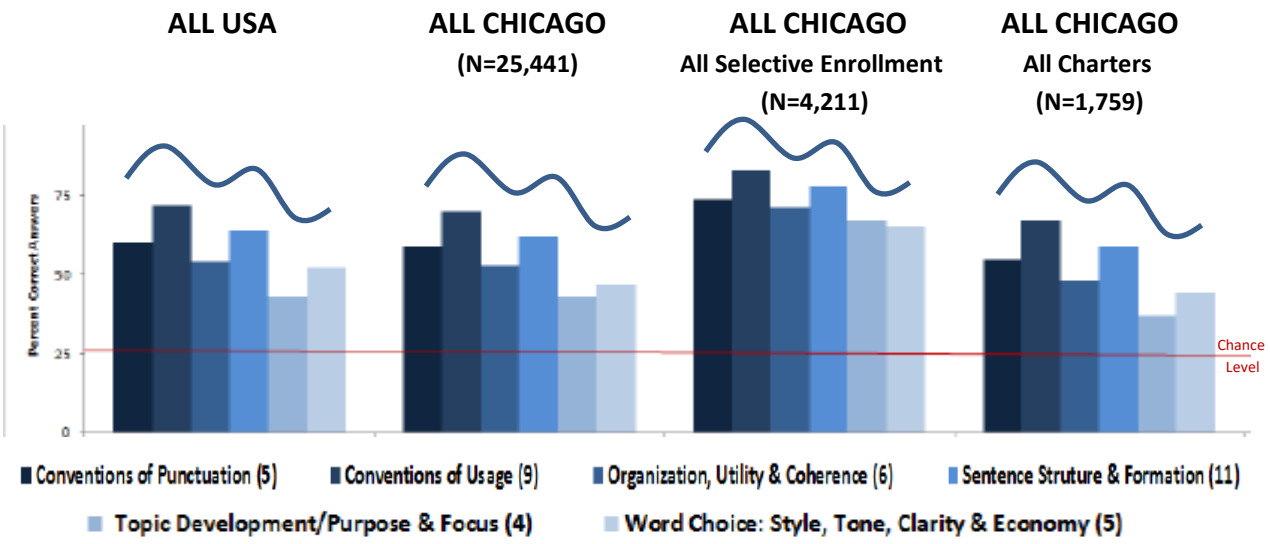
Not Just the ISAT

The ISAT and the State of Illinois were not alone in reporting results that consistently co-varied across content strands. ACT's now-retired EPAS system of high school assessment did, too. The same is true for the ACT's new ASPIRE system. ASPIRE is now being marketed nationally as an alternative to PARCC for tracking progress toward college and career readiness from the primary grades onward.

Figure 8.4 illustrates that EPAS English results co-varied in close-to-identical ways across four different test populations. This is especially noteworthy because the EPAS English test focused mostly on punctuation and writing conventions and was more "skill-specific" than other EPAS sub-tests. The blue wave lines above each bar chart trace the pattern of covariation that occurred on this particular test. In the legend, numbers in parentheses show the number of questions that were associated with each content strand.

Figure 8.4

On the 9th Grade EXPLORE Exam in 2008, English Power Standards Consistently Co-varied



In Chicago and elsewhere, EPAS content strands were often referred to as "power standards." Like ISAT content strands, EPAS power standards moved up and down together in almost perfect symmetry regardless of the size, achievement level or demographic characteristics of the population tested.

Figures 8.5 and 8.6 illustrate how this looked for 10th and 11th graders who took the same PLAN English test in 2008.

Figure 8.5
On the 10th Grade PLAN Exam in 2008, English Power Standards Consistently Co-varied

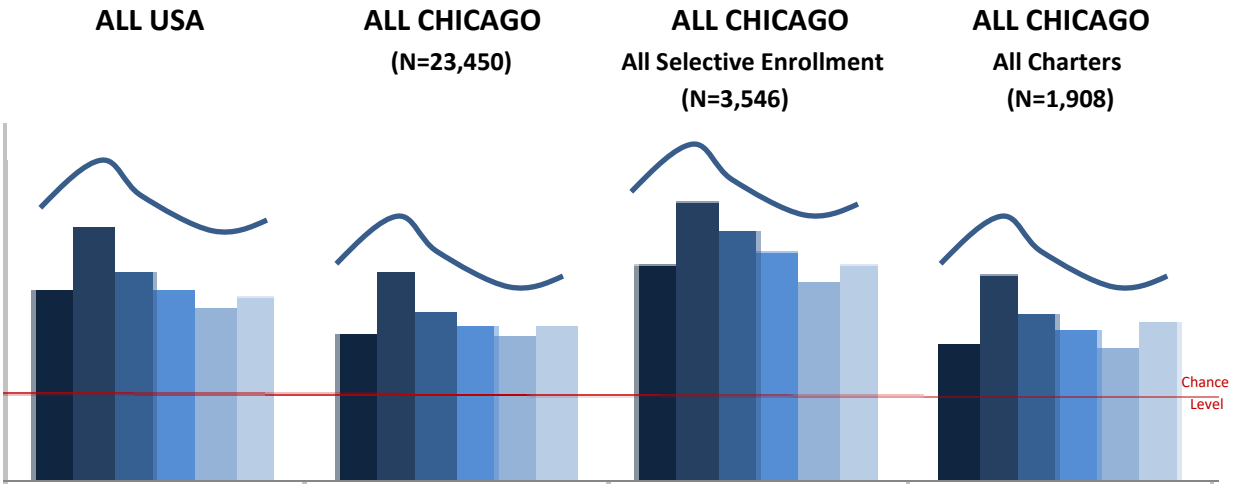
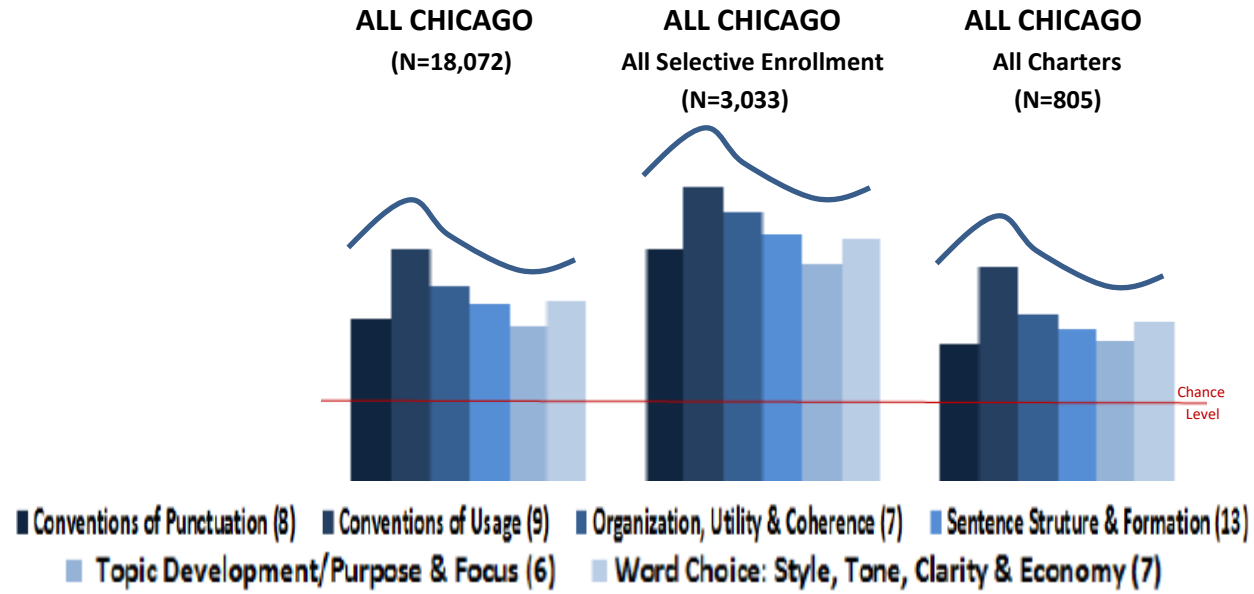


Figure 8.6
In 2008, 11th Grade Scoring Patterns the on PLAN English Exam Co-varied in Almost Exactly the Same Way as 10th Grade Scoring Patterns Did on the Same Test



What's Going On?

Covariation across independent measures is a clear signal that those measures are not, in fact, independent. Covariation signals that other factors that the measures have in common are mostly responsible for the way outcomes change. The evidence summarized below identifies some likely candidates for what those factors are.

Items on standardized tests almost always assess more than one content strand at a time. As described in Section 7, it is exactly this mix of skill, content knowledge and complexity that test makers use to adjust the depth and breadth of "general knowledge" that standardized items are designed to represent. So when groups of experienced teachers are given the task of coding test items by content strands, they almost always assign items to two or more content strands instead of just one.

Figure 8.6 summarizes the results of a workshop exercise that UIC's Urban Education Leadership Program often carries out with school principals and teacher leadership teams. Figure 8.6 comes from a workshop where 37 master teachers from across Chicago analyzed items from an interim assessment of sixth grade math achievement. Workshops that draw on standardized items from other grades and subject areas regularly produce close-to-identical results

Without being shown how the test publisher coded each test item (top row in gray), teachers were asked to complete fourteen test items. They were then asked to mark an "X" next to the content strand(s) that described what the items assessed. Figure 8.7 shows, for example, that most teachers felt Item #1 assessed three content strands (6B-C—computation, operations, estimation, properties; 6D—ratios, proportions, percents; and, 8C-D—writing, interpreting and solving equations). The gray row at the top shows that the test publisher used only one content strand to describe Item #1, and it was different that any of the strands that teachers identified.

Figure 8.7
Most Standardized Test Items Assess More than Ones Content Strand at a Time
Even Though Test Publisher Code Each Item to a Single Strand

Selected *Learning First* Items (Harcourt) Coded by 37 Master Teachers in Chicago's Pathways Leadership Development Program

JOB #2	Make an "X" next to the Illinois Learning Standards that you think are being tested by each question Make an "X" next to more than one standard if you think more than one standard is being tested															
Illinois Learning Goal	Illinois Learning Standard	Test Publisher's Coding of Illinois Learning Standard	6A	6A	6A	6A	6A	6D	6A	6D	6A	6A	6A	6A	6A	6A
		Item Number →	1	2	13	21	28	32	34	36	37	39	40	42	43	45
6	6A	Representations/Ordering		X			X					X		X	X	X
	6 B-C	Computation, Operations, Estimation, Properties	X	X	X	X			X			X		X	X	X
	6D	Ratios, Proportions, Percents	X	X	X	X	X	X		X	X		X	X	X	X
7	7 A-B-C	Units, Tools, Estimation and Applications				X	X	X				X	X	X	X	X
8	8-A	Representations, Patterns and Expressions			X		X	X	X	X	X	X	X	X	X	X
	8-B	Connections Using Tables, Graphs and Symbols			X						X		X		X	
	8 C-D	Writing, Interpreting and Solving Equations	X		X					X						X
9	9A	Properties of Single Figures and Coordinate Geometry														
	9B	Relationships Between/Among Multiple Figures														
10	10	Data Analysis, Statistics, Probability								X	X					

The message of Figure 8.7 is that one likely candidate for why content strand data move up and down together is that items which test publishers identify with just one content strand actually measure multiple strands at the same time. If all or most items actually represent multiple strands, it would be surprising if strands ***didn't*** move up and down in tandem.

It turns out that there is an even more compelling reason for strands to co-vary. As described in Section 7, test publishers intentionally design test items to reflect ***different levels of academic difficulty***. This is an essential part of the equating process that allows test makers to create standardized scales which produce normal, bell-curve distributions.. It also lets test makers assess similar content information at very different levels of academic depth and complexity (e.g. the two NAEP items in Section 7 that both assessed "measurement").

Figure 8.8 illustrates that item difficulty is more or less evenly distributed in each of the seven content strands that the test publisher used to represent Illinois Reading Standards. This even distribution of item difficulty across content strands pretty much guarantees that that achievement will rise and fall in tandem across strands.

Items numbers on the far right of Figure 8.8 are rank-ordered and color-coded by the percentage of correct answers that all students tested got on each item. The easiest third of items tested are shown in green at the top of the table. The middle third are shown in yellow and the hardest third are shown in pink. Item 66 was the easiest item on the test. It shows in green in the top row of the content strand called "1C.1 Literal or Simple Inference." Item 57 sits at the bottom of the Literal or Simple Inference column. It shows in pink because it was the fifth most difficult item on the entire test.

Figure 8.8
Each Content Strand Has a Pretty Even Mix of Easy, Medium and Difficult Items

Items, Content Strands and Levels of Difficulty on an Interim Reading Assessment in Chicago in the Fall of 2009

	Illinois Standard Strands for 8th Grade ETS Benchmark Reading: Fall 2009							Item
	1A.2 Word in Context	1C.1 Literal or Simple Inference	1C.2 Summarizing Main Idea	1C.3 Sequencing and Ordering	1C.4 Evidence-based Conclusions	1C.5 Interpreting Instructions	1C.6 Author's Design/Purpose	
All CPS Easiest 1/3		1C.1-INF.8.14						66
		1C.1-INF.8.14						48
					1C.4-CONC			50
	1A.2-CONT							60
					1C.4-CONC			53
					1C.4-CONC			84
					1C.4-CONC			55
					1C.4-CONC			67
					1C.4-CONC			61
			1C.2-SUMM					63
		1C.1-INF.8.14						88
	1A.2-CONT							82
			1C.2-SUMM					46
			1C.2-SUMM					49
				1C.3-SEQU				72
All CPS Middle 1/3			1C.2-SUMM					89
						1C.5-INST		68
	1A.2-CONT							87
							1C.6-PURP	51
							1C.6-PURP	65
		1C.1-INF.8.14						74
		1C.1-INF.8.14						80
					1C.4-CONC			86
	1A.2-CONT							85
				1C.3-SEQU				59
				1C.3-SEQU				47
	1A.2-CONT							52
							1C.6-PURP	71
		1C.1-INF.8.14						56
					1C.4-CONC			75
All CPS Hardest 1/3							1C.6-PURP	77
					1C.4-CONC			73
			1C.2-SUMM					70
					1C.4-CONC			78
		1C.1-INF.8.14						81
			1C.2-SUMM					79
					1C.4-CONC			58
						1C.5-INST		64
	1A.2-CONT							90
					1C.4-CONC			83
		1C.1-INF.8.14						57
							1C.6-PURP	69
			1C.2-SUMM					76
					1C.4-CONC			62
			1C.2-SUMM					54

Another interesting result of rank-ordering results by item difficulty is that rank-ordering follows pretty much the same pattern regardless of the achievement level of students who take the test. The items that high-achieving students find most difficult are typically the same ones that lower-achieving students struggle with the most. The only real difference is that higher-achieving populations have higher correct-response frequencies on all or most items than low-achieving populations do.

Figure 8.9 illustrates the point by showing correct response frequencies and difficulty levels for five different Chicago populations. From left to right, the 8th grade populations shown are: All Chicago Schools; all schools Chicago's highest-achieving network cluster; a single school in the highest achieving network cluster; all Chicago schools in the district's lowest-achieving network cluster; a single school in the districts lowest-achieving cluster. With minor variations, the same items which were green, yellow and pink for the district as whole were also green, yellow and pink in each of the sub-groups shown.

Figure 8.9
Factors that Make Items More of Less Difficult are
Pretty Much the Same Regardless of Overall Achievement Levels
Correct-response Frequencies by Item for Five Different Student Populations

Item	Answ	All CPS % Correct	High-Achieving % Correct		Low-Achieving % Correct	
			Area	School	Area	School
66	A	85%	89%	100%	71%	75%
48	D	81%	86%	98%	73%	77%
50	B	79%	85%	100%	71%	68%
60	A	73%	77%	87%	70%	68%
53	C	73%	78%	93%	71%	68%
84	D	71%	78%	94%	60%	62%
55	C	71%	76%	93%	61%	68%
67	C	70%	75%	89%	67%	74%
61	B	69%	76%	87%	56%	66%
63	A	69%	73%	96%	58%	60%
88	D	67%	72%	91%	50%	55%
82	C	66%	73%	85%	39%	47%
46	D	66%	69%	72%	62%	60%
49	A	64%	68%	81%	63%	68%
72	B	63%	69%	89%	55%	53%
89	B	63%	70%	87%	55%	55%
68	D	63%	69%	91%	54%	58%
87	C	62%	68%	85%	68%	70%
51	B	60%	61%	72%	60%	64%
65	D	59%	64%	78%	54%	49%
74	C	59%	65%	85%	43%	49%
80	A	57%	65%	91%	38%	40%
86	B	57%	60%	70%	46%	55%
85	A	56%	62%	72%	49%	53%
59	D	55%	61%	70%	44%	36%
47	C	54%	57%	67%	46%	53%
52	D	52%	57%	63%	48%	47%
71	C	52%	55%	80%	45%	51%
56	B	52%	56%	72%	41%	38%
75	A	51%	57%	69%	46%	43%
77	D	50%	55%	85%	38%	40%
73	B	49%	56%	85%	46%	47%
70	B	48%	55%	69%	40%	38%
78	A	48%	54%	74%	34%	34%
81	D	46%	52%	70%	34%	32%
79	D	46%	52%	78%	35%	28%
58	C	45%	52%	74%	35%	45%
64	B	41%	46%	65%	39%	40%
90	C	38%	41%	44%	31%	30%
83	C	38%	40%	54%	28%	32%
57	A	37%	43%	67%	35%	42%
69	A	37%	42%	67%	30%	26%
76	B	34%	36%	56%	37%	45%
62	B	31%	38%	69%	29%	15%
54	A	28%	33%	50%	17%	21%

Like most states, Illinois did not release items and response frequencies for the ISAT itself. For that reason, it was not possible for practitioners or researchers to sort ISAT items in the same way that Figures 8.8 and 8.9 do. But careful reading of the Scaling & Equating section of ISAT Technical Manuals makes it clear that actual ISAT results distributed in more or less the same way as those shown in Figures 8.8 and 8.9. Items sorts ***were*** possible for EPAS tests. Those results closely matched the patterns shown in Figures 8.8 and 8.9.

Morphing Standards into Skills

Packaging ISAT and EPAS test results into content strands delivered two clear messages to teachers and other end users. The first was that both tests assessed discrete bits of knowledge and skill which fit neatly into separate topical categories. The second was that earning higher scores was mostly about mastering greater volumes of content and skill in each topical category.

This section illustrates neither of these things were true. Content strands were mostly arbitrary, after-the-fact labels that made the ISAT, ACT/EPAS and other tests appear to be more "criterion-referenced" than they actually were. This explains why content strands moved up and down in lock-step. It explains why skilled teachers find that standardized test items almost always assess several standards at the same time. And it explains why item difficulty does a better job of describing what tests actually assess than descriptors that focus on specific skills and content information.

Does NWEA MAP Do Any Better?

"Is it about getting data for instruction? Or is it about measuring the results of instruction? In a nutshell, that's what this is all about," said Douglas J. McRae, a retired test designer who helped shape California's assessment system. "You cannot adequately serve both purposes with one test."

Gewertz, Catherine "Test Group Rethinks Questions" *Education Week*, December 5, 2012. Pages 1, 24

During the past decade, many Illinois districts have opted to supplement annual state assessments with interim Measures of Academic Progress (MAP) tests developed by the Northwest Evaluation Association (NWEA). Growth in MAP testing received a significant boost in 2010 with the passage of Illinois' Performance Evaluation Reform Act (PERA). PERA mandated that annual performance evaluations of teachers and principals include measures of academic growth. Because districts were prohibited from using ISAT and Prairie State Achievement results for PERA purposes, many districts turned to the MAP for standardized growth measures that met PERA requirements.

First marketed in 2000, MAP testing is now used in over 7,400 schools and districts around the world. A major part of MAP's appeal is that it reports both nationally-normed achievement and growth data ***and*** highly detailed classroom diagnostics from the same test. Moreover, it reports quickly. All MAP testing is done on-line. Results typically turn around overnight. That contrasts sharply with paper and pencil results which often take three to six months or more to report.

A unique feature which helps the MAP generate ***both*** normed ***and*** skill-specific information from the same test is a sophisticated algorithm that tailors test content to individual students ***while they are being tested***. The algorithm starts off testing at relatively low levels of academic challenge. As testing continues, the challenge level gradually increases until students begin to produce incorrect responses. At that point, difficulty levels are progressively refined until the algorithm determines that a reliable estimate of achievement can be reported. Under normal circumstances, the whole process takes 60 minutes or less.

Ostensibly, individual tailoring and large banks of MAP assessment items allow the MAP to do it all. On the normative side, MAP items are equated along a standardized "RIT" scale (see Figure 3.1) that allows MAP to produce bell-curve scoring distributions. These distributions make it possible to make reliable estimates of achievement and growth against national norms. On the diagnostic side, MAP algorithms and MAP's online format make it possible to create a unique mix of test items for every

student tested. MAP then uses that information to create detailed diagnostic profiles which classroom teachers can access directly online for skills grouping and other instructional purposes.

The Illusion of Precision

A few years prior to the passage of No Child Left Behind, the Consortium on Chicago School Research produced a ground-breaking set of recommendations for assessing academic productivity with standardized test instruments. A fundamental ground rule was that,

> ". . . standardized tests should be directly aligned with standards. Only if . . . assessments are specifically developed to achieve this aim and have been demonstrated to be valid in this regard will teachers, students and parents know whether they are making progress on these important goals.
>
> The content of the standards should dictate the content of the tests. A "back in" solution (choosing among existing tests the one that comes the closest to matching the standards) is inadequate. Under such an approach, test publishers rather than local leaders get to decide the accountability standards for judging schools."
>
> Bryk, et. al. (1998) *Academic Productivity of Chicago Public Elementary Schools*, p. 47

Like all nationally-normed tests, MAP depends on large banks of items and passages to assess student achievement. To be credible, those items and passages have to validly represent essential curricular content and produce reliable estimates of current and future performance. MAP builds validity by making items as reflective as possible of the skills and content that are most often addressed in American textbooks. It builds reliability and predictive power by equating those items so they produce normal, bell-curve distributions of results when administered to large test populations.

MAP's diagnostic reports reflect a major advance over content strands. The most important element of that advance is what MAP calls its "Descartes Learning Continuum." That continuum highlights the role that academic depth and difficulty play in obtaining higher scale scores on MAP's standardized scale.

Figure 8.10 shows a small portion of a MAP "Classroom Breakdown Report" for third grade math. This particular report makes instructional recommendations about "place value, counting and cardinality" skills for students who have scored in the 171 to 180 range on the MAP scale:

- The column on the left identifies skills that warrant periodic reinforcement which are characteristic of the 161 to 170 range
- The column in the middle identifies skills at students' current instructional level that typically need to be developed
- The column on the right identifies skills from the 181-190 range which are likely to be at the outer edge of student understanding. These skills need to be introduced to support later growth

Tracing skills from left to right across each column, teachers can see concrete examples of the close connection between rising scores and incremental upticks in difficulty and complexity. For example, the top row of Figure 8.10 shows the following progression:

- Left column (161-170): Identifies whole numbers under 100 using base-10 blocks
- Middle column (171-180): Identifies whole numbers from 100-999 using base-10 blocks
- Right column (181-190) identifies the numeral and written name for numbers 101-999, e.g. 342 is three hundred forty-two and vice versa

As helpful as Classroom Breakdown Reports may be in the hands of skilled practitioners, an endnote at the bottom of each page reveals important limits to MAP's diagnostic power. The first line in the endnote reads,

> **At the range mid-point, this is the probability students would correctly answer items measuring these concepts and skills.*

This endnote refers to ***probability statements*** that are printed in parentheses in the header of each skills column. In the case of Figure 8.10, these statements mean that students who scored in the 171 to 180 range have:

- a 73% ***likelihood*** of already knowing the skills and concepts in the column on the left
- a 50% ***likelihood*** of already knowing the skills and concepts in the middle column
- a 27% ***likelihood*** of already knowing the skills and concepts in the column on the right

It is unclear how much attention most teachers pay to MAP's fine print about probabilities, or what they make of that language when they do. But what it highlights is the shaky diagnostic tightrope that MAP reports walk between ***actual*** skills mastery and ***probable*** skills mastery.

Figure 8.10

Probabilities Are as Close as MAP Can Get to Reporting Actual Skills Mastery

NWEA Class Breakdown Report for Mastery of Number and Operations Skills in a Third Grade Class

Skills and Concepts to Enhance (73% Probability*) 161 - 170	Skills and Concepts to Develop (50% Probability*) 171 - 180	Skills and Concepts to Introduce (27% Probability*) 181 - 190
Understand Place Value, Counting, and Cardinality	Understand Place Value, Counting, and Cardinality	Understand Place Value, Counting, and Cardinality
• Identifies whole numbers under 100 using base-10 blocks • Identifies the numerical and written name for whole numbers 11 to 20 (e.g., 15 is fifteen, and vice versa) • Counts 1 to 10 objects • Identifies missing numbers in a series through 100 • Recognizes and generates equivalent forms for the same number using physical models for whole numbers 11 to 20 • Orders whole numbers less than 10 • Writes whole numbers in standard and expanded form through the tens	• Identifies whole numbers 100 - 999 using base-10 blocks • Identifies the numerical and written name for whole numbers 21 to 100 (e.g., 62 is sixty-two, and vice versa) • Identifies the numeral and written name for whole numbers 101 to 999 (e.g., 342 is three hundred forty-two, and vice versa) • Identifies missing numbers in a series through 100 • Counts backwards from a given number (given number greater than 10) • Recognizes and generates equivalent forms for the same number using physical models for whole numbers 11 to 20 • Compares sets of objects and identifies which is equal to, more than, or less than the other (1 to 10 objects) • Compares whole numbers through 999 • Counts objects that are grouped into tens and ones • Identifies the place value and value of each digit in whole numbers through the tens place	• Identifies the numeral and written name for whole numbers 101 to 999 (e.g., 342 is three hundred forty-two, and vice versa) • Identifies the numeral and written name for whole numbers to 1000 to 9999 (e.g., 3456 is three thousand, four hundred fifty-six, and vice versa) • Identifies the numeral and written name for whole numbers 10,000 to 100,000 • Compares whole numbers through 999 • Rounds 2- and 3- digit whole numbers to the nearest ten • Rounds 3-digit whole numbers to the nearest hundred • Counts objects that are grouped into tens and ones • Identifies whole numbers under 100 given place value terms (e.g., 3 tens and 4 ones = 34) • Identifies the place value and value of each digit in whole numbers through the tens place • Identifies the place value and value of each digit in whole numbers through the hundreds place • Identifies the place value and value of each digit in whole numbers through the thousands • Identifies the place value and value of each digit in whole numbers through the hundred thousands • Compares and orders decimals to the hundredths place (same number of digits after decimal)

Explanatory Notes
* At the range mid-point, this is the probability students would correctly answer items measuring these concepts and skills. Both data from test items and review by NWEA curriculum specialists are used to place Learning Continuum statements into appropriate RIT ranges. Blank cells indicate data are limited or unavailable for this range or document version.

Like all standardized tests that produce normal, bell-curve scoring distributions, MAP tests use equated items to make reliable estimates of students' overall achievement level. Each item's level of difficulty determines how likely it is that students at different achievement levels will answer the item correctly. So even though the number of skills that are tested directly in an online testing session is just a small fraction of all MAP items, Classroom Breakdown Reports can show probabilities for ***every skill in the item bank*** based on students' overall scale score.

Footnotes and technicalities aside, the broad message that Classroom Breakdown Reports communicate is that they offer busy teachers detailed prescriptions for what needs to be taught in order to raise student achievement. Their clear message is that the recipe for improving academic achievement is:

- periodic reinforcement of the skills shown in the left column
- focused instruction of the skills shown in the middle column
- gradual introduction of the skills shown in the right column

Widespread use of the MAP throughout Illinois and across the nation suggests that many teachers and school leaders value this support and may not be all that worried about differences between actual mastery and probable skills mastery. It is also possible that direct linkages between teacher/principal evaluations and growth in MAP achievement give everyone involved an added incentive to follow MAP's recipe.

The most powerful critique of MAP's approach to instructional support is that it does not provide end-users with the ***actual items and passages*** that are used to assess what students know and are able to do. Instead, like content strands on steroids, MAP reports code test items and translate them into specific skills and categories that then appear in the Classroom Breakdown Report.

The net effect of this filtering process is that it reduces reading and math curriculum to lengthy lists of discretely teachable skills and makes end-users entirely dependent on MAP for accurate descriptions of what was actually tested. This becomes especially problematic at higher scoring ranges where, by definition, students have to size up and work through more than one skill or concept at a time in order to obtain higher scores.

In schools and districts where MAP testing is conducted three times a year, and teacher/principal evaluations are tied to MAP results, the line between classroom instruction and perpetual test preparation can easily begin to blur. Under those conditions, incentives can be strong to turn Classroom Breakdown Reports into the *de facto* math and reading curriculum of the school. These kinds of conditions are the ones that have led increasing numbers of educators and parents to raise deep concerns about over-testing, and to question the value of standardized testing as a whole.

In the end, the most practical indictment to date of MAP's approach is that it has not yet produced meaningful gains in student achievement. Independent research on MAP and other similar assessment systems offers no compelling evidence that they are helping teachers move the needle on achievement (see insert below). Nevertheless, schools and districts across Illinois continue to spend millions of dollars and thousands of instructional hours each year in hopes they will.

You go to war with the army you have.

Lots of Dollars and Many Instructional Hours

But No Independent Evidence that MAP or Other Interim Assessments Help Improve Achievement

EDUCATION WEEK

http://www.edweek.org/ew/articles/2013/10/02/06testing_ep.h33.html?qs=interim+assessments

Demand for Testing Products, Services on the Rise

By Sean Cavanagh Published Online: October 1, 2013

"In an analysis released last year and completed for the Software and Information Industry Association, a major trade group, consultants John Richards and Leslie Stebbins surveyed vendors selling products to schools, then extrapolated those findings to a broader set of companies based on the composition of the market.

They estimated that the current market for technology-based testing and assessment products and services in fiscal 2011 was **$1.6 billion**. Preliminary results that are still being analyzed show the market grew by at least 20 percent for fiscal 2012 . . .

* * * * *

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement (2012)

http://www.air.org/resource/impact-measures-academic-progress-map-program-student-reading-achievement

. . . MAP teachers were not more likely than control group teachers to have applied differentiated instructional practices in their classes. Overall, the MAP program did not have a statistically significant impact on students' reading achievement in either [of the grades studied]" p xii

"The study investigated one primary and two secondary confirmatory research questions:

1. Did the MAP program (that is, training plus formative testing feedback) affect the reading achievement of grade 4 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?
2. Were MAP resources (training, consultation, web-based materials) delivered by NWEA and received and used by teachers as planned?
3. Did MAP teachers apply differentiated instructional practices in their classes to a greater extent than their control counterparts?

"The report also addressed one exploratory question:

4. Did the MAP program affect the reading achievement of grade 5 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?

"The results of the study indicate that the MAP program was implemented with moderate fidelity but that MAP teachers were not more likely than control group teachers to have applied differentiated instructional practices in their classes. Overall, the MAP program did not have a statistically significant impact on students' reading achievement in either grade 4 or grade 5."

EDUCATION WEEK

Interim Assessments Yield Disappointing Results in Indiana Study

http://blogs.edweek.org/edweek/inside-school-research/2014/04/large_study_suggests_that_inte.html

Holly Yettick April 5, 2014

"As the roll out of the assessments for the Common Core State Standards approaches, school districts have been spending millions of dollars per year on diagnostic exams in the hopes that these interim results will help improve scores on high-stakes state exams given toward the end of the school year.

"But research presented this week at the annual meeting of the American Educational Research Association in Philadelphia suggests that, though diagnostic assessments may lead to some increases in 3rd-8th grade math scores, they have no effect in reading and a small negative effect in the lower grades in both subjects.

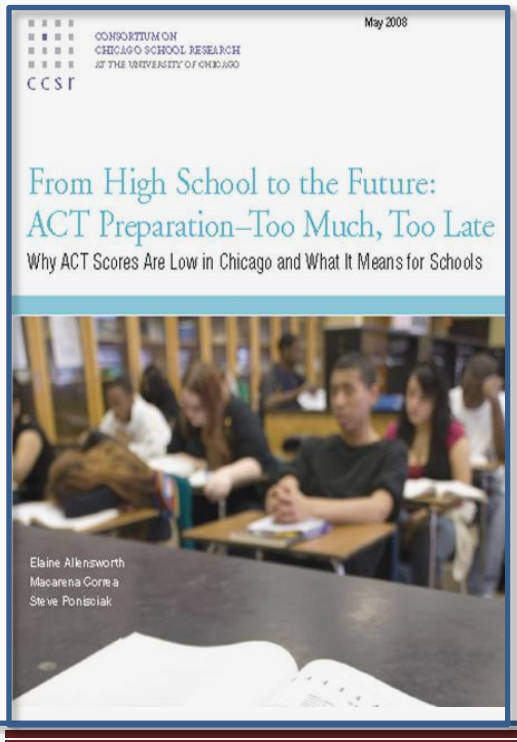
"The paper, written by researchers at Michigan State University and the American Institutes for Research and funded by the federal Institute of Education Sciences, summarized the results of two different experiments that took place in Indiana. In 2009-10, 20,000 students in 50 schools participated in the research. Half were randomly assigned to take the popular mCLASS and/or Acuity interim assessments throughout the school year. Treatment group students in grades K-2 took mClass. Treatment group students in grades 3-8 took Acuity. The remaining students were the control group, so they went about business as usual. In 2010-11, a separate but similar experiment included a total of 30,000 students in 50 schools."

* * * * *

THE UNIVERSITY OF CHICAGO

UCHICAGO CCSR

https://consortium.uchicago.edu/publications/high-school-future-act-preparation-too-much-too-late



Executive Summary p. 2

"There is no evidence that scores benefit from learning testing strategies or from practicing on test questions outside of taking a full, timed practice test.

"In fact, improvements from the PLAN to the ACT are smaller the more time Teachers spend on test preparation in their classes and the more they use test preparation materials . . .

". . . Teachers need better strategies for preparing their students for this challenging high-stakes test. Using class time to practice the test is not producing higher scores."

PART 3

STANDARDIZED TESTING AT THE CROSSROADS

"In August of 2008, the chairman of the Federal Reserve Bank called an emergency meeting with then-President George W. Bush to inform him that the entire financial system was melting down. Bush, shocked, responded by asking, 'How did we get here?' . . . Part of the answer is that the system was designed to fail. Naturally, the banks did not ***want*** to fail. They did not ***want*** the economy to fall apart. But these results were nevertheless natural outgrowths of the choices they made about measuring and rewarding performance. Investment banks failed to hold their employees accountable for key decisions that were well within their control**"**

Harris, Douglas (2011) *Value-Added Measures in Education* p. 14, 16

PART 2 of this study describes a constellation of grading and reporting practices under NCLB that systematically misrepresented what standardized tests actually assessed:

- Section 5 shows that arbitrary cut scores introduced deep distortions into publically reported results and cynically misrepresented what it meant to "meet" Illinois Learning Standards
- Section 6 illustrates that apparent differences in results produced by ISAT, MAP, NAEP and PARCC exams mostly disappear when tests are graded in the same way
- Section 7 shows that the reason most standardized tests produce similar results is that they are designed to measure "general knowledge" more than specific skills and content knowledge
- Section 8 illustrates that NCLB reporting practices actively reinforced rote teaching and learning by morphing measures of general knowledge into content strands, power standards and lengthy lists of discrete skills

By just about any measure, the problems described in PART 2 reflect stunning violations of public trust for which no one has yet been held accountable. To date, however, national policy recommendations about the future of standards-based assessment have given no special priority to the ethics and accuracy of public reportage.

In 2013, for example, the Stanford Center for Opportunity Policy in Education published *Criteria for High Quality Assessment.* This report outlined eighteen "indicators of quality for next-generation assessments." Two of those indicators address the way results are communicated to parents, educators and other end-users:

- *Rich feedback on student learning and performance*
- *Tasks that reflect and can guide valuable instructional activities*

In 2014, the Council of Chief State School Officers (CCSSO) leaned heavily on Stanford Center recommendations in a report called, "Criteria for Procuring and Evaluating High-Quality Assessments." It translated Stanford indicators into twenty-four criteria, two of which focus on test reportage:

- *Score reports illustrate a student's progress on the continuum toward college and career readiness, grade by grade, and course by course. Reports stress the most important content, skills, and processes and show how the assessment focuses on them, to show whether or not students are on track to readiness.*
- *Reports are instructionally valuable, are easy to understand by all audiences, and are delivered in time to provide useful, actionable data to students, parents, and teachers*

Early in 2016, the National Center for the Improvement of Educational Assessment (NCIEA), published its *Guide to Evaluating Assessments Using CCSSO Criteria for High Quality Assessments: Focus on Test Content.* This report operationalized criteria for evaluating ***test content only*** and deferred work on other CCSSO criteria . . . including communication of results . . . until later in 2016.

Shortly after the NCIEA issued its guide, the Thomas Fordham Institute and the Human Resources Research Organization (HumRRO) published parallel evaluations of PARCC, Smarter Balanced, ACT-ASPIRE and the Massachusetts Comprehensive Assessment System (MCAS). Both Fordham and HumRRO used NCIEA methodology to conduct their work. For this reason, both reports were silent on how well PARCC, Smarter Balanced, ASPIRE and MCAS communicated results to end-users.

Meanwhile, Rome has been burning:

- Initial enthusiasm for Common Core standards and assessments by national teachers' unions has eroded as states and districts moved to link new standards and assessments to new forms of high-stakes accountability; in Chicago, the teachers' union now actively opposes both Common Core State Standards and PARCC assessments
- Nine hours of new PARCC testing in 2015 was enough to turn general annoyance over "too much testing" into organized political resistance. 11% of Chicago students and over 4% of all students statewide "opted out" of PARCC testing in the spring of 2015; legislation granting parents the right to opt-out of standardized testing is now pending in the General Assembly
- Of the original 23 states and the District of Columbia that planned to participate in PARCC testing, only eleven states and the District of Columbia ended up administering PARCC tests in the spring of 2015. Currently, only six states plus the District of Columbia are scheduled to conduct PARCC testing in the spring of 2016.

Reframing the Problem

The nation's most vocal advocates for Common Core State Standards and second-generation, standards-based assessments have done an excellent job of communicating that their prescription for better schooling is to get everyone to eat more broccoli. Their message has been:

- NCLB gave us lax standards and easy tests
- Lax standards and easy tests need to be replaced by tougher standards and harder tests which more accurately reflect the demands of a competitive, 21st century world

This excerpt from Fordham's recent assessment study captures the sentiment:

> *For too many years, state assessments have generally focused on low-level skills and have given parents and the public false signals about students' readiness for postsecondary education and the work force. They often weren't very helpful to educators or policymakers either. States' adoption of college and career readiness standards has been a bold step in the right direction. Using high-quality assessments of these standards will require courage: these tests are tougher, sometimes cost more, and require more testing time than the previous generation of state tests Will states be willing to make the tradeoffs?. [p 24]*

PART 3 illustrates why hard test/easy test is a misleading mantra that mostly misses the point. The evidence says that most NCLB-era tests were able to measure academic progress toward college readiness with high reliability. What they lacked was the ability to report back deep, rich information about how students are thinking and where they're getting stuck.

PARCC now has this ability, but gave few hints of it in its first round of test reports. What parents and educators saw instead were new numbers, new proficiency ratings and lower success rates. This made PARCC look like it wasn't much more than a longer/harder/costlier version of the old ISAT.

- Section 9 describes why hard test/easy test unfairly scapegoats older state tests and misses what is most important about PARCC's potential contributions
- Section 10 recalls the original promise of standards-based assessment and describes changes in second-year PARCC reportage that could begin to repair the damage that NCLB left in its wake

SECTION 9

Moving beyond Hard versus Easy

The claim that PARCC tests are a different species that cannot be compared with NCLB-era tests is only partly true. Statewide, 2015 PARCC results closely matched earlier ISAT results once both tests were graded in the same way. The biggest new asset that PARCC brings to the table is ***rich detail*** *about how students are thinking and where they're getting stuck. But the first round of PARCC reports made scant mention of this asset, and failed to package it in ways that educators and parents could easily use.*

The National Drumbeat for Increased Rigor and Depth of Knowledge

In recent years, most of the national conversation about standards and accountability has focused on the wisdom of replacing first-generation, state and local standards with a more coherent and rigorous set of Common Core State Standards. Common Core advocates have argued that conventional, mile-wide, inch-deep curricular content leaves most students poorly prepared for the intellectual demands of a post-industrial era. The Common Core created guidelines for math and English language curricula that call for deeper understanding of underlying concepts and greater ability to engage in complex, real-world problem solving

A persistent theme of the Common Core has been that first-generation, state standards from the NCLB era were too diffuse, too topical and too simplistic to prepare students for the world they are about to enter. In 2010, for example, the Fordham Institute conducted an exhaustive, state-by-state comparison of Common Core and individual state standards. The study's conclusion, illustrated in the maps below, was that the quality of most state standards fell far short of those in the Common Core.

FIGURE9 9.1
Fordham Institute Comparison of Common Core and Earlier State Standards

English/Language Arts Standards **Mathematics Standards**

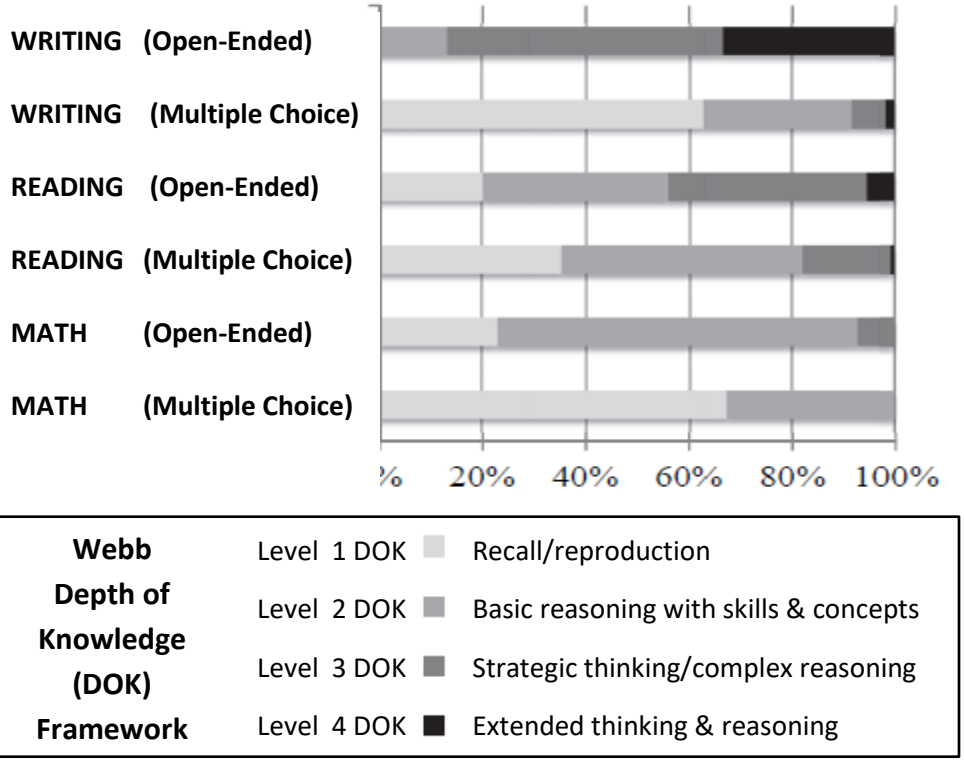
Source: Carmicheal et.al. *The State of State Standards—and the Common Core in 2010* pp. 6-7

Illinois standards for English/Language Arts and Mathematics both received a grade of "D" in the Fordham study and were rated as "among the worst in the country."

State-level assessment instruments have been scrutinized in similar ways and with very similar results. In 2012, Rand Education carried out a study for the William and Flora Hewlett Foundation that assessed the cognitive demand of items on 17 state-level assessments using Norman Webb's Depth of Knowledge (DOK) framework. Rand researchers found that none of the questions on multiple choice **math** tests, and less than 20% of questions on multiple choice **reading** tests, assessed DOK at Level 3 or higher.

Figure 9.2

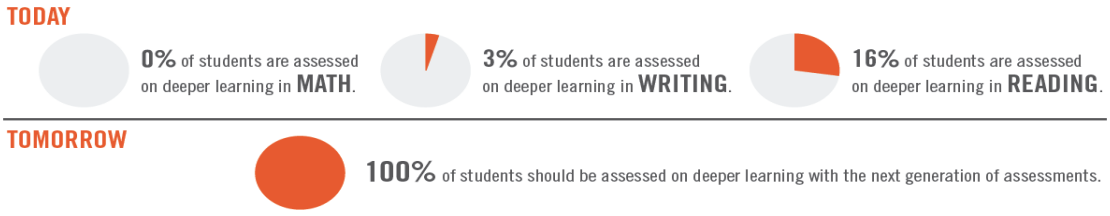
Rand Education Study of Webb Depth-of-Knowledge Levels on 17 State Assessments



Source: Kun, Yuan and Vi-Nhuan Le (2012) "Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests," RAND Education, p. xiii http://www.rand.org/pubs/working_papers/WR967.html

In 2014, Advance Illinois raised the profile of the Rand study in Illinois with a publication called *Making Assessments Work.* This study underscored and amplified Rand's findings by saying,

> *Current state tests rarely measure students' depth of knowledge. Questions typically focus on basic comprehension and information recall rather than conceptual understanding and analysis across disciplines. The new assessments aim to change that by gauging student's higher-order thinking.*



Source: Advance Illinois, (2014) *Making Assessments Work: Supporting Teaching and Learning in the Common Core Era*, page 6 http://www.advanceillinois.org/publications/making-assessments-work/

Items from the ISAT were not included in Rand's analysis because only a handful of sample ISAT items were ever publically released. But close examination of the validation sections in ISAT Technical Manuals shows that DOK distributions on the ISAT were consistent with Rand's overall conclusion. The ISAT, too, was heavily weighted with DOK 1 and DOK 2 items.

Figure 9.3

DOK Levels on the ISAT Were Consistent with Those Reported by the Rand Study

Number and Percentage of Items by DOK Level in ISAT Validation Studies

ISAT Gr 3-8		DOK 1	DOK 2	DOK 3	DOK 4
Reading	Number	18	154	63	0
	Percentage	8%	66%	27%	0%
Math	Number	104	191	11	0
	Percentage	34%	62%	4%	0%

Source: Illinois State Board of Education, *Illinois Standards Achievement Test 2013 Technical Manual* pp. 99-100, 184 http://www.isbe.net/assessment/pdfs/isat_tech_2013.pdf

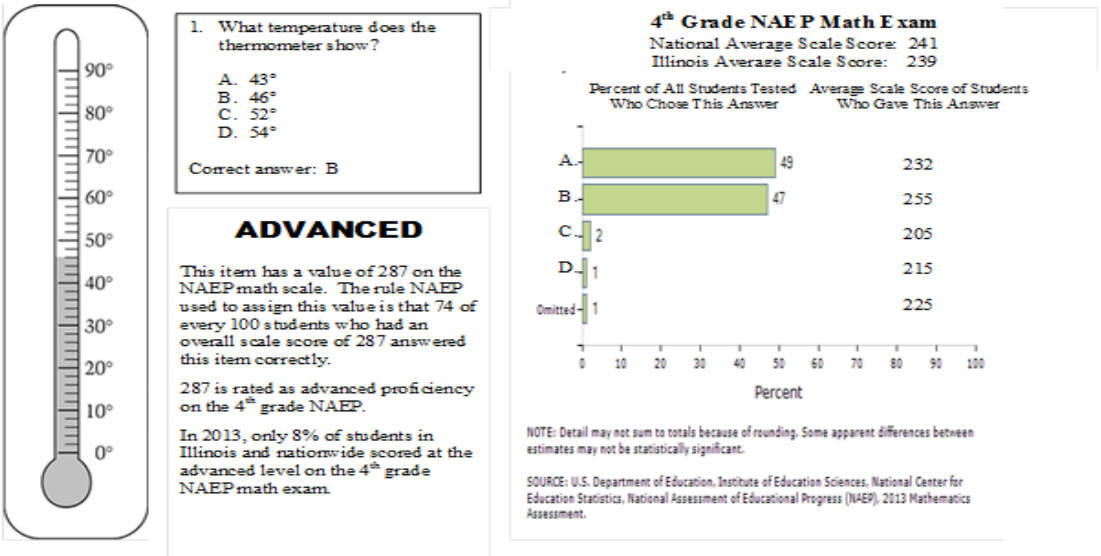
The clear message from Fordham, Rand and others has been that standards and tests under NCLB just weren't rigorous enough to assess students' progress toward college readiness. New tests were needed to produce honest assessments that didn't "top-out" out far below levels required to predict college success.

More Ways to Get Predictive Power than DOK Alone

On its face, Rand's analysis is pretty compelling. But it turns out that DOK is only one of several tools that test makers use to assess higher-ordering thinking and predict academic progress toward college readiness. Figure 9.4 (shown earlier as Figure 7.4) illustrates the point.

Figure 9.4

Choosing the Correct Answer for This Item Required Inferential Thinking and a Rudimentary Understanding of Ratio and Proportion



Source: NAEP website http://nces.ed.gov/nationsreportcard/itemmaps/

The Rand study's shorthand for properties that define different DOK levels was:

- **DOK 1**: Recall of a fact, term, concept, or procedure.
- **DOK 2**: Use information, conceptual knowledge, and procedures in two or more steps.
- **DOK 3**: Requires reasoning, developing a plan or sequence of steps; has some complexity and more than one possible answer
- **DOK 4**: Requires an investigation, time to think and process multiple conditions of the problem, and non-routine manipulations.

Source: Kun, Yuan and Vi-Nhuan Le (2012) "Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests," pp. 14-15 http://www.rand.org/pubs/working_papers/WR967.html

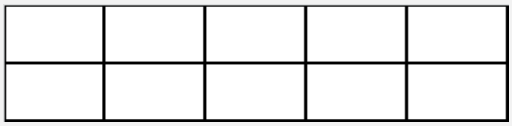
Using Rand's criteria, Figure 9.4 fits squarely in the DOK 2 category. It requires students to use information, conceptual knowledge and procedures in two or more steps. Nevertheless, fourth graders who consistently answered items like Figure 9.4 correctly scored at the 92nd percentile or above and are rated by NAEP as having "Advanced" proficiency.

Figure 9.5 shows a more complex item from the 4th grade NAEP. It requires a constructed response and meets all the criteria for DOK 3.

Figure 9.5

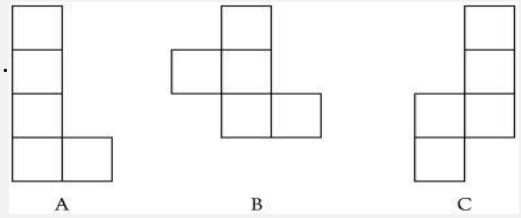
This Item Has Multiple Steps, Calls for More than One Answer and Requires Strategic Thinking

Question refers to a number tiles on a paper strip. Please remove the 10 number tiles and the paper strip from your packet and put them on your desk.



1. Turn the tiles facedown so that the blank side is showing.

It is possible to arrange 5 tiles so that at least one side of each tile completely shares one side of another tile. Here are 3 different ways to do this:



Two figures are not considered different if one figure can be turned or flipped to match the other.

The figures below are not examples of proper arrangements or new arrangements.

Using 5 of your tiles, show 3 other different ways to arrange the tiles. Trace the tiles to show each figure. **Show the lines separating the individual squares.**

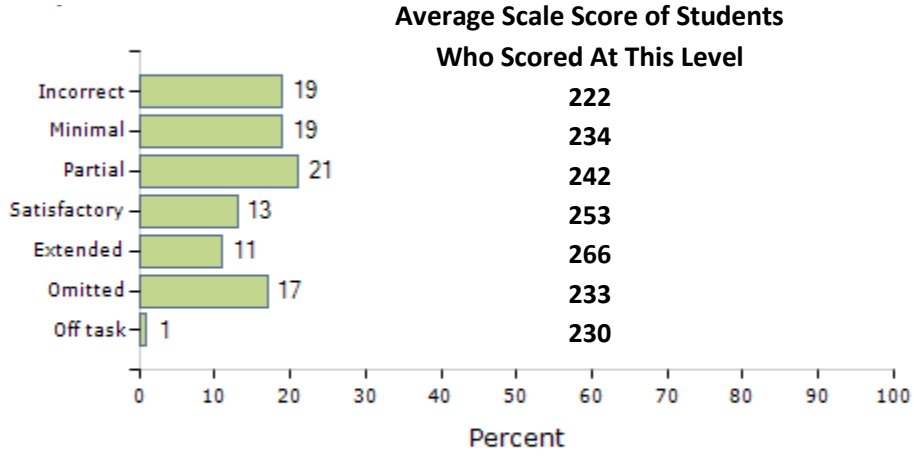
Source: NAEP website http://nces.ed.gov/nationsreportcard/itemmaps/

Procedurally, the thinking required for a satisfactory response to Figure 9.5 is more complex than that required by Figure 9.4. But the scale score value (287) that NAEP assigned to Figure 9.4 is ***identical to the scale score value it assigned to Figure 9.5***. And the average scale (253) of students who constructed satisfactory responses to Figure 9.5 was actually two points ***lower*** than the average scale score (255) of students who correctly answered Figure 9.4.

Figure 9.6

Items with Higher DOK Ratings Aren't Necessarily More Difficult and Don't Necessarily Assess Progress toward College Readiness Better than Items with Lower DOK Ratings

Average Scale Scores and Response Frequencies of Constructed Responses to Figure 9.5 on the 4th Grade NAEP



NOTE: Detail may not sum to totals because of rounding. Some apparent differences between estimates may not be statistically significant.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment. http://nces.ed.gov/nationsreportcard/itemmaps/

The point of comparing the DOK 2 item in Figure 9.4 with the DOK 3 item in Figure 9.5 is that DOK is only one of several tools that test makers use to represent and assess academic depth. The ISAT may have been low on items with high DOK. But it still measured and predicted progress toward college readiness about as well as NAEP, MAP, ACT and PARCC did.

Rigor-marole

When testing advocates and school officials say that more rigorous tests are needed to measure and predict progress toward college readiness, their argument seems strong. Over half of students who "met standards" on the ISAT failed to reach college readiness benchmarks on the ACT. Respected research organizations confirm that tests like the ISAT were low on DOK. And students and teachers report that most items on the ISAT "felt" easy, while most items on the PARCC and ACT "feel" hard.

Looks like a duck, walks like a duck, quacks like a duck . . .

Enter the black box of standardized testing. For over a century, statisticians have been developing assessment tools that let them make reliable predictions with remarkably little information. Election polls predict results within a few percentage points based on phone interviews with less than a

thousand likely voters. Short personality inventories like Myers-Briggs and DISC startle most adults with the depth and intimacy of the reports they produce.

The apparent magic of all this comes from sampling techniques and banks of carefully-vetted items that reliably ***represent*** broad constellations of preferences, dispositions and traits. Standardized achievement tests work the same way. So long as the purpose of standardized testing is just estimation and prediction, that work can be done with a handful of carefully road-tested items.

The great irony of NCLB-era testing is that low-cut scores were the only thing that kept tests like the ISAT from reporting progress toward college readiness (see Figure 9.8). More ironic still, high-stakes accountability forced test makers to load up the ISAT up with less demanding items so that low cut scores would be more reliable. This didn't prevent the ISAT from measuring progress toward college readiness. But it made the ISAT look and feel like an easier test than it actually was.

Figure 9.7 shows a hypothetical example of correct and incorrect answers on an 8th grade ISAT reading test by student who is on-track for an ACT score a 21 or 22. Blue boxes represent correct answers; red boxes represent incorrect answers. Figure 9.7 illustrates that:

- Tests like the ISAT had a relatively small number of items (50 to 75) and could not measure achievement at all levels with equal amounts of depth and accuracy
- Including larger numbers of items with the same level of difficulty as the cut score was important because it increased reliability at the point where high-stakes judgments were made
- It was still possible to estimate achievement at higher achievement levels, but smaller numbers of higher-difficulty items made those estimates somewhat less reliable than estimates at lower achievement levels.
- To be on-track for college readiness in Figure 9.7, the student had to answer 39 items correctly, but just a small portion of those were higher-difficulty items.

Figure 9.7

The ISAT that High Stakes Accountability & Adequate Yearly Progress Led Illinois to Build

Low Cut Scores for "Meeting Standards" Forced Test Designers to Over-Represent Low-Difficulty Items. Doing that Increased the Reliability of Scores that Carried the Greatest Consequences for Schools and Districts

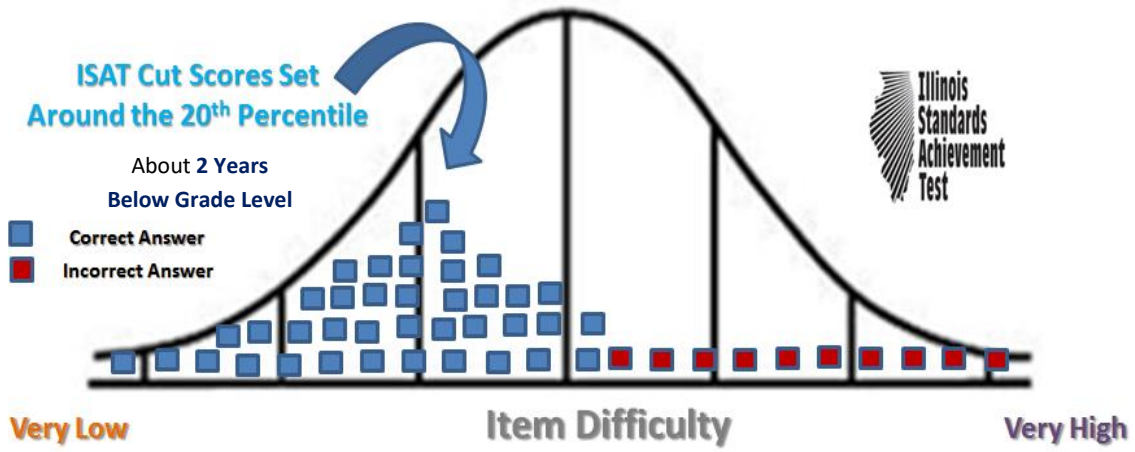


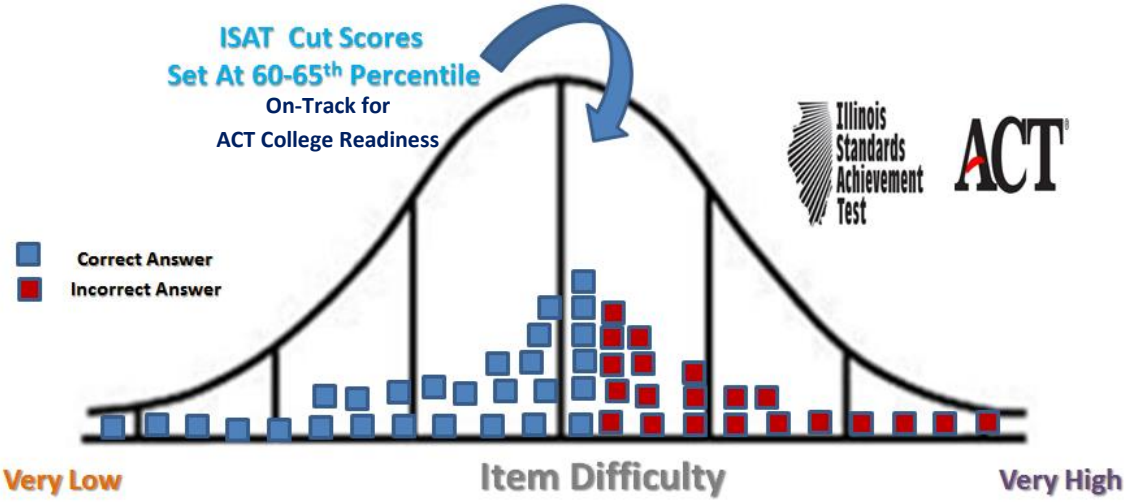
Figure 9.8 offers a rough approximation of how item difficulty is distributed on the ACT. It shows how the eighth grader from Figure 9.7 who scored at the 60-65th percentile on the ISAT would be likely to answer questions three years later on the ACT.

- Although the total number of test questions in both hypothetical examples is the same, the test taker in Figure 9.8 scores at more or less the same level but with fewer correct answers (28 versus 39)
- Because the ACT is more interested in raising score-reliability at higher levels of item-difficulty, a larger proportion of questions on the ACT than the ISAT represent higher levels of difficulty

Figure 9.8

The ISAT We Would Have Had if Cut Scores Had Been Pegged to College Readiness Benchmarks

A More Balanced Distribution of Low-, Medium- and Higher-Difficulty Items*



*Adapted from Table 2.5, "Difficulty Distributions and Mean Discrimination Indices for ACT Test Items 2011-2012" in Technical Manual, The ACT (2014) page 14 http://www.act.org

Under NCLB, there were strong reasons not to set ISAT cut scores at levels that would have invited stiff federal sanctions. But there was no real justification for being dishonest with students, parents, educators and the public at large about the actual level of achievement those cut scores represented. By the same token, there can be little doubt that grading and reporting strategies during the NCLB era were grossly misleading. But it is equally misleading to claim that standardized tests under NCLB were simply too easy to measure or predict progress toward college readiness.

The Proof is in the Pudding

The dominant narrative about PARCC and other post-NCLB assessments is pretty much the same no matter who you hear it from. This is new a new breed of tougher, more rigorous exams that can't be meaningfully compared with earlier assessments. Radical differences in test content and test design create new baselines for assessing progress in future years. In short, we're starting over from scratch.

In December 2015, the Illinois State Board of Education repeated this message in a series of communications that accompanied the release of new PARCC results. Many of those documents

emphasized again that PARCC is something completely new and cannot be compared with prior tests. In the sample shown below, ***blue italics*** have been added to highlight key parts of the text.

Illinois State Board of Education

100 North First Street • Springfield, Illinois 62777-0001
www.isbe.net

James T. Meeks
Chairman

Tony Smith, Ph.D.
State Superintendent of Education

Three things you need to know about

2015 PARCC Assessment Results

2015 PARCC Assessment Results
December 2015, ISBE Division of Public Information

1 Individual student score reports will be available to parents starting Dec. 11 for the first Partnership for Assessment of Readiness for College and Careers (PARCC) exam administration. These scores will look different than scores from previous state tests and may appear lower than what educators and parents typically expect for some students. This difference does not mean our students know less or are less capable. ***The PARCC exam scores will reflect higher expectations for what students should know and be able to do to stay on track for college and careers***.

2 ***PARCC exam results cannot be compared to test scores from the state's previous assessments*****.** The PARCC exam is a different test that uses extended tasks and technology-enhanced items ***to more accurately measure students' critical thinking, problem solving, and writing skills***, which are all necessary for students to succeed in higher education and/or their career field after high school.

3 ***This first year of PARCC testing will serve as a baseline for future test scores and will determine the performance targets for years to come.*** In future years, scores will be available to educators and parents earlier. This will equip teachers with information they need to differentiate instruction and support for every student.

Source: www.isbe.net

But once the results of PARCC and earlier tests are graded in the same way, most differences disappear. Figure 9.9 offers another illustration of this hard-to-believe result using 15 years of eighth grade reading scores from five Illinois school districts.

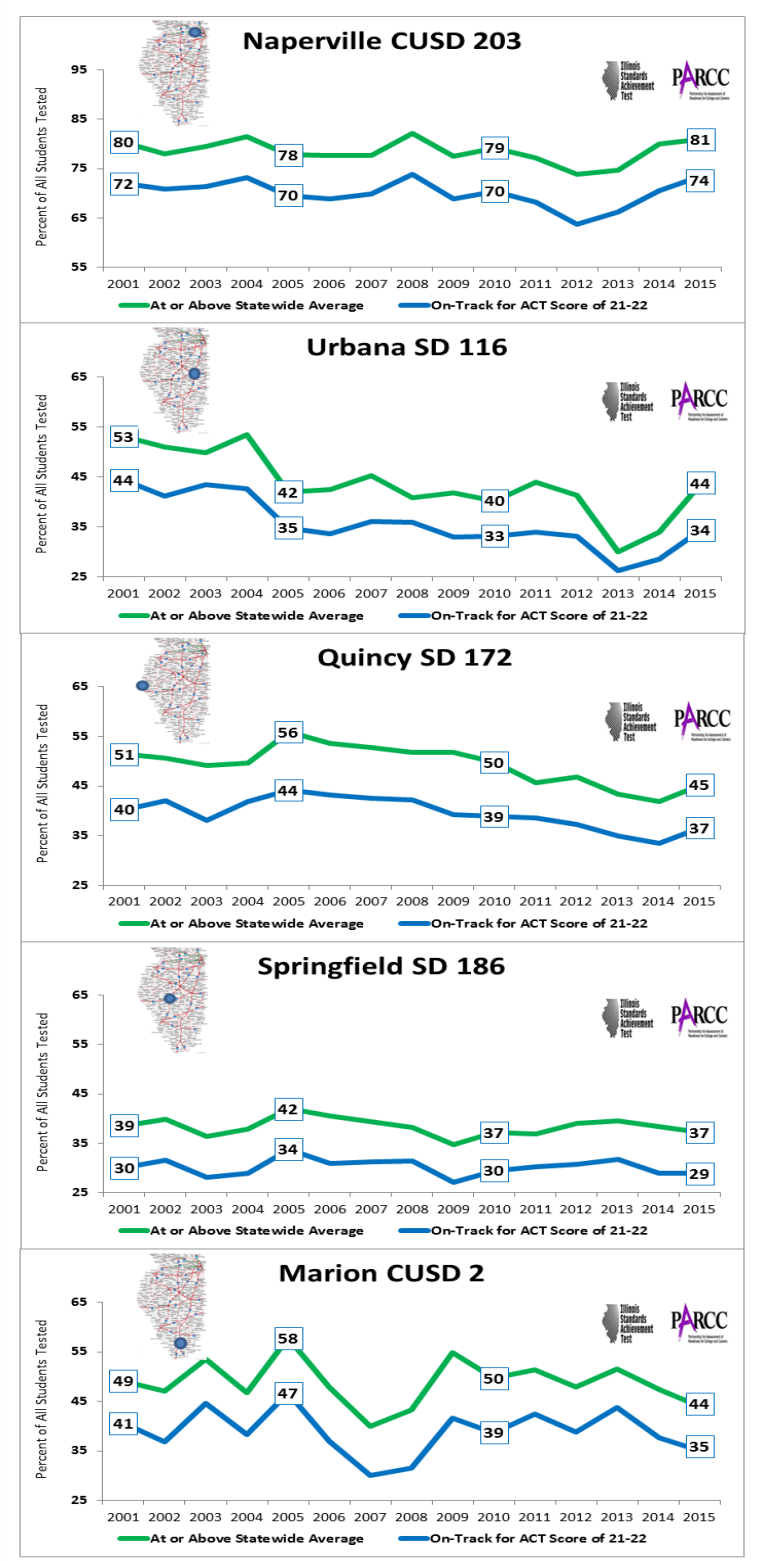
Like earlier illustrations in Sections 4, 5 and 6, the ISAT and PARCC results shown in Figure 9.9 reflect statewide norms instead of unaligned cut scores. The green lines in Figure 9.9 show the percent of students in each district who scored at or above statewide averages between 2001 and 2015. Blue lines show the percent of students in each district who scored:

- at or above Level 4 on the 2015 PARCC exam
- at or above the 60th percentile of statewide ISAT scoring distributions between 2001 and 2014; during the NCLB era, 60th percentile on the ISAT was a reliable 8th grade predictor of 21-22 on the 11th grade ACT

Figure 9.9

Using Statewide Norms, PARCC Results Become Simple Extensions of Long-Term ISAT Trends

8th Grade Reading: Percent At/Above State Averages & Percent On-Track for ACT College Readiness



Predictive versus Descriptive Power

The evidence is clear. PARCC's ability to measure and predict college readiness isn't markedly better than ISAT's was. From the beginning, most NCLB-era tests were quite capable of measuring and predicting progress toward college readiness . . . if policy makers had chosen to use them that way.

The problem with the ISAT and most other NCLB-era tests wasn't their inability to measure and predict progress toward college readiness. Their problem was that they were not designed to report standards-based, diagnostic information. But instead of acknowledging that limitation, test publishers and state officials papered it over with pseudo-diagnostic gimmicks like content strands and power standards (see Section 8). These gimmicks misled educators and parents, and grossly distorted what tests actually assessed.

The real promise of PARCC is that it is specifically designed to gather rich information about what students know and where they are getting stuck. It does this by using additional test time to have students describe their thinking about a wide range of multiple choice and extended-response tasks.

In particular, extra test time lets PARCC do two key things that most NCLB-era tests could not do:

- First, it gives PARCC the opportunity to tap more deeply into students' ability to assemble evidence and build compelling explanations to support their answers. Reported back in user-friendly ways, this information carries huge potential for supporting classroom conversations at all grade levels, teacher-to-teacher conversations in grade/departmental teams, and parent-student conversations around the kitchen table.
- Second, extra time lets PARCC ask a full range of questions across all levels of difficulty. This contrasts sharply with tests like the ISAT which traded off balanced representation at all difficulty levels to increase the reliability of scores that were close to high stakes cut score boundaries (see pp. 76-77)

Differences in the density of information that ISAT and PARCC collected are easy to see in the scoring distributions of both tests. The upper chart in Figure 9.10 shows the distribution for all 149,152 students who took the 8th grade ISAT reading test in 2014. The 53 vertical bars show the number of students who scored at each of the reported scale scores. The 261 spaces between bars indicate scale scores for which no information was reported.

The lower chart in Figure 9.10 shows the scoring distribution for all 143,545 students who took the 8th grade PARCC English/Language Arts test in 2015. This distribution shows close to a four-fold increase in the breadth and density of the information produced by PARCC (201 PARCC score points versus 53 ISAT score points). The increase is especially pronounced in the upper half of the scoring distribution.

Color bands on both charts show the percentile ranges that match PARCC's five proficiency levels:

- Color bands on the PARCC distribution show actual scale score ranges for each proficiency level
- Color bands on the ISAT distribution show scale score ranges that most closely approximate the difficulty level of comparable ranges on the PARCC

The table below illustrates how PARCC's denser information set is also more evenly distributed across proficiency levels than ISAT information was.

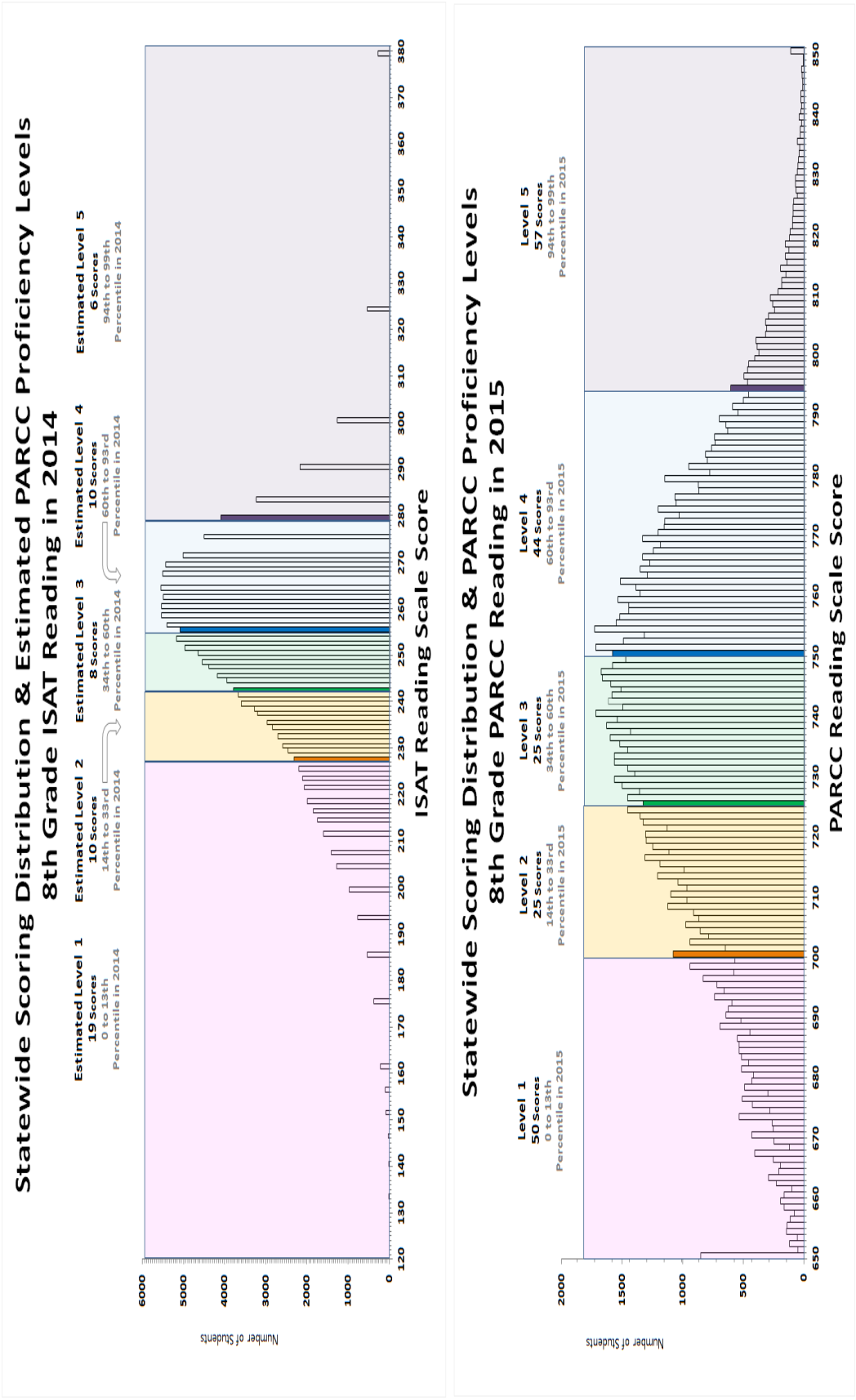
Number of Scores Reported/% of Total Possible Scores

	Level 1	Level 2	Level 3	Level 4	Level 5	Total Possible
ISAT 2014	19/7%	10/4%	8/3%	10/4%	6/2%	261
PARCC 2015	50/25%	25/12%	25/12%	44/22%	57/28%	201

Figure 9.10

PARCC Generated Close to

Four Times More Descriptive Power in 2015 than the ISAT Did in 2014

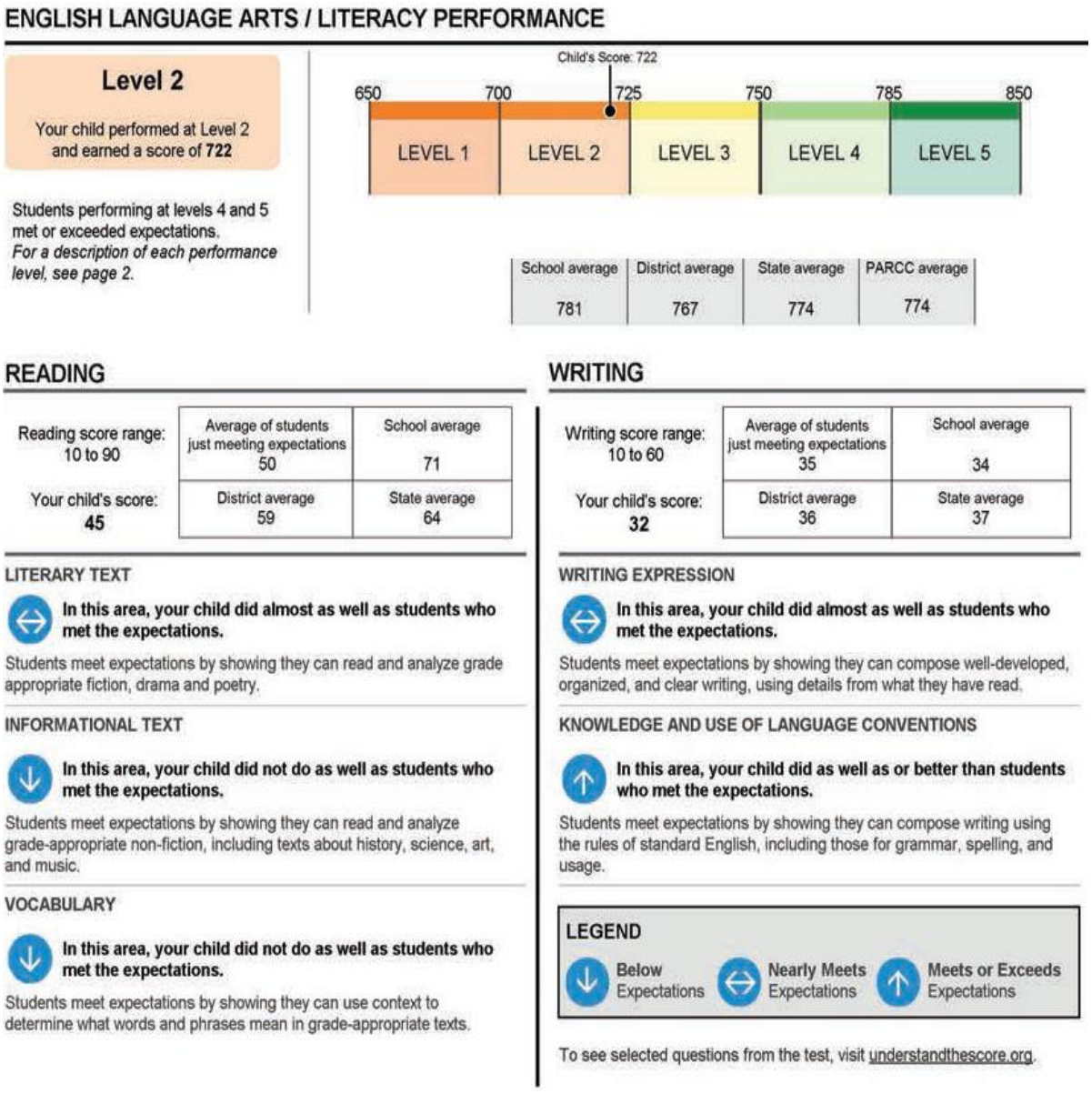


Early indications from PARCC were that they would be releasing representative samples of items, passages, constructed responses and scoring frequencies at the completion of each testing cycle. Yet, inexplicably, the first round of PARCC reports made scant mention of these important new assets. Instead, most of PARCC reportage focused on normative comparisons with local, state and national results . . . exactly the kind of reportage that standards-based assessment has been promising to replace for more than two decades. Released items were relegated to a large, poorly-indexed depot in PARCC's Partnership Resource Center at https://prc.parcconline.org/assessments.

The normative emphasis of PARCC reporting was especially pronounced in its standard score report for students and parents. A sample of that report is shown below.

Figure 9.11

From Page 1 of PARCC's Individual Student Report



ENGLISH LANGUAGE ARTS / LITERACY PERFORMANCE

Level 2

Your child performed at Level 2 and earned a score of **722**

Students performing at levels 4 and 5 met or exceeded expectations.
For a description of each performance level, see page 2.

Child's Score: 722

650 | 700 | 725 | 750 | 785 | 850

LEVEL 1 | LEVEL 2 | LEVEL 3 | LEVEL 4 | LEVEL 5

School average	District average	State average	PARCC average
781	767	774	774

READING

Reading score range: 10 to 90	Average of students just meeting expectations: 50	School average: 71
Your child's score: **45**	District average: 59	State average: 64

LITERARY TEXT

In this area, your child did almost as well as students who met the expectations.

Students meet expectations by showing they can read and analyze grade appropriate fiction, drama and poetry.

INFORMATIONAL TEXT

In this area, your child did not do as well as students who met the expectations.

Students meet expectations by showing they can read and analyze grade-appropriate non-fiction, including texts about history, science, art, and music.

VOCABULARY

In this area, your child did not do as well as students who met the expectations.

Students meet expectations by showing they can use context to determine what words and phrases mean in grade-appropriate texts.

WRITING

Writing score range: 10 to 60	Average of students just meeting expectations: 35	School average: 34
Your child's score: **32**	District average: 36	State average: 37

WRITING EXPRESSION

In this area, your child did almost as well as students who met the expectations.

Students meet expectations by showing they can compose well-developed, organized, and clear writing, using details from what they have read.

KNOWLEDGE AND USE OF LANGUAGE CONVENTIONS

In this area, your child did as well as or better than students who met the expectations.

Students meet expectations by showing they can compose writing using the rules of standard English, including those for grammar, spelling, and usage.

LEGEND

Below Expectations | Nearly Meets Expectations | Meets or Exceeds Expectations

To see selected questions from the test, visit understandthescore.org.

Pretty much all of the information that PARCC provided on the front page defined individual achievement in comparison with other students and groups. This how-are-you-doing-compared-with-others message is spelled out even more explicitly on the back:

> ***How can I use the reading and writing scores?***
>
> *The best way to make sense of these scores is to* ***compare them to the average*** *for students who met the expectations and the average for students in your child's school, district and state* [emphasis added]

Figure 9.12

From Page 2 of PARCC's Individual Student Report

What are the PARCC tests? The tests measure how well students have learned grade-level material in English language arts/literacy and mathematics. Students who meet or exceed expectations are on track for the next grade level and, ultimately, for college and careers. The tests include questions that measure your child's fundamental skills and knowledge, and require students to think critically, solve problems and support or explain their answers. The test is one of several ways to help parents and teachers understand how well children are learning.

What do the performance levels mean? The performance levels listed below describe how well students met the academic expectations for their grade level.

- Level 1: Did not yet meet expectations
- Level 2: Partially met expectations
- Level 3: Approached expectations
- Level 4: Met expectations
- Level 5: Exceeded expectations

How do the test scores this year compare to those in past years? The knowledge and skills tested this year are different - and in some cases more rigorous - than in the past. If your child's score is different than you expected, meet with your child's teacher to understand what that means and how you can help your child improve his or her performance.

How will my child's school use the test results? Results from the test give your child's teacher information about his/her academic performance. The results also give your school and school district important information to make improvements to the education program and to teaching.

How can I use the reading and writing scores? The best way to make sense of these scores is to compare them to the average for students who met the expectations and the average for students in your child's school, district, and state. Also, look at the information below the scores. How is your child doing in each area of reading and writing? Ask your child's teacher how you can give your child more opportunities to be challenged and how you can support his/her academic needs.

Probable range. The probable range in the score on this test is plus or minus 7 points. This is the amount of change that would be expected in your child's score if he/she were to take the test many times. Small differences in scores should not be over interpreted.

Source: http://www.isbe.net/hot-topics.htm?col2=open#toolkit

How do all these numbers and comparisons connect to specific learning standards? The report mentions "expectations" a total of 17 times on the front page and 9 times on the back. And parents are properly advised to consult their child's teacher about what numbers mean in relation to those expectations. What is still not clear to most teachers is what they should say when parents ask.

When NCLB mandated large-scale, standards-based assessments in 2002, it did so based on the premise that the country needed a better way to measure learning than simply comparing students with each other. For the most part, first generation state tests could not rise to that challenge. Between 2010 and 2013, the US Department of Education (DOE) responded to that problem by investing close to $400 million in a second generation of large-scale tests that could deliver meaningful, standards-based information to students, parents, educators and the public at large.

A key lesson from fourteen years of testing under NCLB was that the "army we had" back in 2002 was ill-prepared to deliver on the promise of standards based assessment. The army we have today is much better prepared. Whether our current generation of generals will deploy this new army in ways that deliver on the original promise of standards-based assessment is still an open question.

Rigor-marole

EDUCATION WEEK

States Are Setting Higher Proficiency Bars on Tests, Study Finds

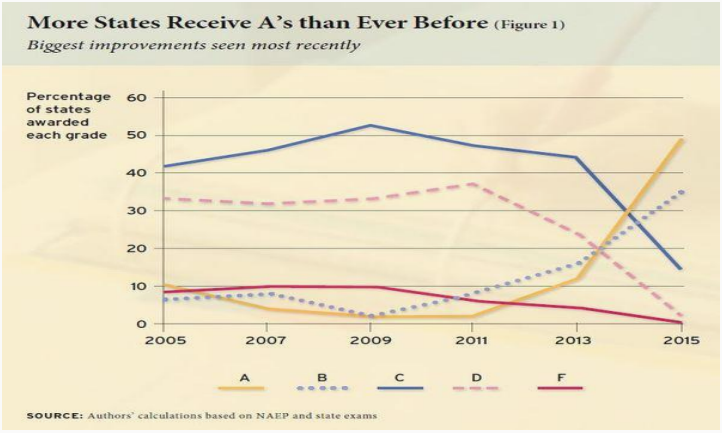
By Catherine Gewertz on January 27, 2016 6:45 AM

The study, published in the journal Education Next, finds that since 2011, 45 states have raised the levels at which students are considered "proficient" on state tests. Thirty-six of the 45 did so within just the last two years. . . .

Higher Grades

Twenty-four states earned A's overall for closely reflecting NAEP's definition of proficiency in 2015. In a 2011 version of the Ed Next study, only three states earned A's. In the 2005 version, only six states did. Eighteen states' ratings jumped by two letter grades or more since 2013.

"In short," writes researcher Paul E. Peterson, with co-authors Samuel Barrows and Thomas Gift, "standards have suddenly skyrocketed." . . .



Source: http://blogs.edweek.org/edweek/curriculum/2016/01/states_setting_higher_proficiency_bar_on_tests.html

* * * * * * *

AIR
American Institutes for Research

National Benchmarks for State Achievement Standards
February 2016

This report uses national benchmarking as a common metric to examine state achievement standards and compare how high these standards are compared to the National Assessment of Educational Progress (NAEP) achievement levels. It also compares how much students are expected to learn in some states with how much they are expected to learn in other states. The study uses NAEP grades 4 and 8 reading and mathematics as benchmarks for individual state achievement standards. The study also benchmarks the achievement standards of Smarter Balanced Assessment Consortium (referred to in this study as Smarter Balanced), Partnership for Assessment of Readiness for College and Careers (PARCC), and ACT Aspire. Benchmarking Smarter Balanced, PARCC, and ACT Aspire provides a common metric (i.e., the NAEP scale) that can be used to compare the stringency of their achievement standards. [page 1]

http://www.air.org/resource/national-benchmarks-state-achievement-standards

SECTION 10

Getting Serious about Using Assessment to Support Teaching for Understanding

Throughout the long history of educational assessment in the United States, it has been seen by policymakers as a means of enforcing accountability for the performance of teachers and schools. For a relatively low outlay, assessments could expose academic weaknesses and make it possible to pressure schools and teachers to improve. But, as long as that remains their primary purpose, assessments will never fully realize their potential to guide and inform teaching and learning. Accountability is not the problem. The problem is that other purposes of assessment, such as providing instructionally relevant feedback to teachers and students, get lost when the sole goal of states is to use them to obtain an estimate of how much students have learned in the course of a year.

Gordon Commission on the Future of Assessment in Education (2011)
A Statement Concerning Public Policy, p.5

Your System . . . Any System . . . Is Perfectly Designed to Produce the Results You're Getting

Everybody agrees. "Many 17-year-olds do not possess the higher order intellectual skills we should expect of them."

The thing is, this quote doesn't come from recent work on the Common Core. It comes from *A Nation at Risk*, the document that launched the standards and accountability movement over 30 years ago. Thirty years earlier, the same critique launched the curricular reforms of the 1960's and before that, spawned the Progressive Era of John Dewey.

We've been agreeing about this problem for quite a while. But we're still not very good at teaching all or most of our students "the higher-order intellectual skills we should expect of them."

Close to a century of failed efforts to teach higher-order intellectual skills to more than a small handful of "advanced" or "gifted" students points to one of two conclusions:

- Either most American students are simply not capable of higher-order intellectual operations
- Or the culture of American teaching has not yet learned how to teach higher-order intellectual operations routinely and at scale to all or most students

For most the last century, social science told us that the problem lay mostly with students. Inherent differences in individual aptitude . . . reflected in normal, bell-curve distributions . . . set tight limits on what schools could reasonably be expected to accomplish.

These days, we know better. We know from studies of instructional effect size that, on average, high-quality classroom assessment practices alone can effectively cancel out the negative impact of low socio-economic status on standardized achievement. We know from 35 years of school improvement literature that a small fraction of schools already post achievement results that consistently exceed what their demographics would predict. And we know from long-term, international comparisons that most American classrooms still do not yet use instructional practices for teaching deep understanding that are commonplace in all of the world's highest-achieving countries.

In 2010, Grant Wiggins captured the hard truth of what we know in the first of two short paragraphs in "Why We Should Stop Bashing State Tests":

> *Here is our problem in a nutshell. Students are taught formulas that they learn and spit back unthinkingly— regardless of subject matter—all in the name of "meeting standards." Yet, as so many assessment results reveal, a large portion of U.S. students are so literal minded that they are incapable of solving fairly simple questions requiring interpretation and transfer—which is surely the point of the state standards.*

But a crucial omission in the next paragraph could easily lead some readers to misinterpret Wiggins to mean that the source of the problem is simple-minded teachers and administrators.

> *Is that the fault of the testing system? Or have our teachers and school administrators badly misunderstood what kind of curriculum and instruction a standards-based education demands? No research supports the oft-heard claim that the tests "demand" superficial coverage. (On the contrary, we see the most slavish test prep in inadequate schools, not the best schools.) An education focused on student understanding—a prioritized curriculum focused on transfer—would not yield such depressing results.*

Wiggins' omission is that the shortcomings he described have deep organizational and cultural roots, and that these roots have been systematically nurtured by high-stakes reporting practices under NCLB. Recent comparisons of teaching and learning systems around the world make it clearer what these roots actually look like.

International Comparisons

James Stigler, James Hiebert and their international research team (1999; 2009) have spent over 15 years reviewing thousands of hours of videotaped instruction from around the world as part of the Third International Math and Science Study (TIMSS). Some of their most important findings include the following:

- Teaching is more of a ***cultural activity*** than an individual one:
  - Underlying patterns in the way people teach are very different from one country to the next but are remarkably similar within countries
  - This contradicts the widespread sense among teachers worldwide that their core instructional strategies are highly individualized
- Viewed internationally, features of instructional practice that often receive the most attention have no consistent relationship with standardized measures of achievement:
  - These features include things like lecture and recitation versus independent learning, whole-group versus small group instruction, and use of real-world situations versus conventional curricular content
  - All of these practices vary as much among high-achieving countries as they do between higher and lower achieving countries
- There is one underlying feature of teaching that does consistently appear in all higher-achieving countries:
  - Teachers in all high-achieving countries regularly engage students in active struggle with core concepts and procedures that ***have not yet been explicitly taught***
  - This contrasted sharply with more overtly didactic forms of instruction that Stigler and Hiebert found in American classrooms
  - In ***every one*** of the American classrooms they studied, teachers spent large amounts of time reviewing material and practicing procedures without expecting students to grasp the underlying concepts on which skills and procedures were based

The Culture of American Teaching and the Grammar of American Schooling

Culture is what happens when groups of people are socialized into particular ways of thinking and acting. The particular way of thinking and acting that characterizes learning cultures in most American schools starts with the belief that learning is a step-by-step process that moves in a steady sequence from simple skills to more complex skills. All or most complex learning depends on mastery of simpler, more basic skills. Without that mastery, more complex forms of mastery are impossible to attain because the foundation has not yet been laid to support them.

This way of thinking about the structure of knowledge and the process of learning describes exactly what Stigler and Hiebert saw in their video analysis of dozens of American classrooms. Educational historians (Tyack, 1974; Cuban, 1993; Lagemann, 2002; Payne, 2008) tell us we have been enacting and reinforcing this way of thinking for close to a century through a variety of organizational structures, program designs, curriculum materials, assessment designs, and instructional strategies. In 1993, David Tyack and William Tobin gave it a name. They called it the "grammar of American schooling".

No Child Left Behind offers a powerful object-lesson in just how profoundly Tyack and Tobin's "grammar" continues drive policy and practice at all levels of schooling. A core promise of NCLB was that it would support standards-based instruction by introducing greater validity, reliability and objectivity into the assessment process. What we got instead were reporting practices that consistently distorted what tests assessed but conformed nicely with the grammar of conventional practice:

- First, local and state-level educators operationalized state standards by breaking them down into massive constellations of skill-specific performance indicators
- Then, most states used very low cut scores to define "standards mastery." This forced test publishers to overload state tests with low-difficulty items in order to improve the reliability of results around high-stakes cut-score boundaries (see Section 9)
- Then, test publishers closed the loop by inventing "diagnostic" reports like content strands, power standards and lengthy skills lists. This psychometric sleight of hand knowingly oversimplified what tests actually assessed (see Section 8). But it produced diagnostics that matched up well with the expectations of conventional instructional practice
- Finally, school and district personnel used "the army they had" to meet state and federal accountability requirements. This army took the form standardized, interim assessments that promised to pinpoint curricular priorities that would improve performance on high-stakes tests. While there is no independent evidence that standardized interim assessments make any contribution at all to improved achievement (see Section 8), they have now become a fixture in most schools statewide

Your system . . . any system . . . is perfectly designed to produce the results you're getting.

There is a Better Way

At the dawn of the NCLB era, Paul Black and Dylan Wiliam generated big excitement in the assessment community with their now-classic meta-analysis of classroom assessment practices called, "Inside the Black Box." Published in the *Kappan* in 1998, "Inside the Black Box" summarized two decades of evidence that all pointed to a single, spectacular conclusion. Engaging students with frequent, high-quality feedback about how they are learning and where they are getting stuck improves standardized achievement by an average of 1 to 1 ½ grade levels (an effect size of 0.4 to 0.7 standard deviations).

In 2001, the University of Chicago's Consortium on School Research echoed Black and Wiliams' findings in a ground-breaking study called, *Authentic Intellectual Work and Standardized Tests: Conflict or Co-existence?* Over a three year period, the study assessed the connection between standardized achievement growth and the intellectual demand of classroom assignments.

A key feature of this study was that it mostly took place in schools with high percentages of Black and Latino students from low-income households. On average, 89% of the students at these schools were eligible for free or reduced lunch, 53% were African American and 39% were Latino.

The central finding of *Authentic Intellectual Work* was that students at all achievement levels who were regularly exposed to intellectually challenging assignments had substantially more growth in standardized achievement the students who were not. This finding flew in the face of conventional

notions that only higher-achieving students can handle and benefit from intellectually challenging tasks. Much like Black and Wiliam's findings, average effect sizes associated with intellectually challenging tasks were between 0.4 and 0.6 . . . about a full grade level higher than the norm associated with more typical classroom tasks.

In 2008, John Hattie published *Visible Learning*, the most extensive meta-analysis to date of factors that impact student achievement. Hattie showed that the effect sizes reported by Black and Wiliam and the Chicago Consortium were about the same as the impact that socio-economic status (SES) has on achievement (0.57 standard deviations; roughly one grade level). Hattie also showed that the impact on achievement which comes from engaging ***teacher***s with high-quality feedback about their day-to-day practice was even higher (0.9 standard deviations, or close to two full grade levels)

Few people will be surprised that frequent, high-quality feedback has a big impact on student and adult learning. But many will be surprised to hear that the size of that impact is, on average, big enough to cancel out the negative effects of low SES. That is a hopeful piece of information for an education system that has struggled for decades to increase instructional effectiveness with students from low-income households.

But the opposite is also true. Positive effect size is a measure of impact over-and-above what is typical. When frequent, high-quality feedback about learning and practice gets positive effect sizes of between one and two grade levels, it means very little of that feedback is currently occurring in typical schools and classrooms.

Confronting the Elephant in the Room: The Persistent Poverty of Local Assessment

The most vocal critics of standardized testing often describe how richer, more authentic forms of classroom assessment would make most present-day standardized testing unnecessary. And they are right. The problem is that repeated efforts to scale up richer, more authentic forms of local assessment have never gained traction in more than 20%-25% of American classrooms. After close to a century of effort, 75%-80% of American teachers continue to use grading and assessment practices that are largely unchanged from those used in the early years of the 20th century.

It was exactly this weakness that gave standards and accountability reformers the warrant to initiate massive out-sourcing of classroom assessment under NCLB. Reformers argued that schools and districts needed help from testing experts and commercial testing organizations to assess mastery of new standards throughout the school year. Why? Local educators didn't have the expertise to do the job properly on their own.

It wasn't just standards and accountability reformers who made this claim. Independent studies have been documenting the poverty of classroom assessment for years. In *How to Assess Higher Order Thinking Skills in Your Classroom* (2010), for example, Susan Brookhart writes:

> *Studies analyzing classroom tests, over many decades, have found that most teacher-made tests require only recall of information (Marso & Pigge, 1993). However, when teachers are surveyed about how often they think they assess application, reasoning, and higher-order thinking, both elementary (McMillan, Myron, & Workman, 2002) and secondary (McMillan, 2001) teachers claim they assess these cognitive levels quite a bit . . .*
>
> *The reason that recall-level test questions are so prevalent is that they are the easiest kind to write. They are also the easiest kind of question to ask off the top of your head in class . . . This situation is true for even the best teachers . . .*

> *Teachers who put together tests quickly, or who use published tests without reviewing them to see what thinking skills are required, are likely to end up asking fewer higher-order-thinking questions than they intended. Contrary to some teachers' beliefs, the same thing also happens with performance assessments.* *[pp. 1-2]*

Similarly, Grant Wiggins, wrote,

> *. . . despite the constant criticisms leveled at state tests,* ***local assessment is arguably the far weaker link in the whole chain of would-be reform.*** *Many of us have seen firsthand how invalid and low-level many local tests are. And studies have shown for years that in terms of Bloom's taxonomy, most teacher questions only hit the first two levels (knowledge and comprehension) instead of the higher levels (application, analysis, synthesis, and evaluation). In one high-income suburban New Jersey district that some colleagues and I studied, we found no test question that required any higher-level thinking in all their marking-period tests. Even more surprising, there was no difference across honors and regular-track versions of the same course* ***[emphasis added****].*
>
> *Wiggins, Grant (2010) "Why We Should Stop Bashing State Tests" Educational Leadership 67:6 p. 51*

A lesson from NCLB, if we choose to learn it, is that standardized tests cannot bail us out of this problem. Properly reported, standardized test information can help support better thinking about instructional practice. And it can help to sharpen the focus local assessment. What it cannot do is replace local assessment, or carry out analysis that only teachers can do.

Back to the Future

In the 1985 movie classic, *Back to the Future,* Michael J. Fox's teen age character, Marty McFly, re-set his family history by traveling back to 1955 and helping his then, teenage father act more heroically at a key moment in his parents' relationship. State and local leaders can't turn back the clock the way Marty McFly did. But big improvements in large-scale test design, and big increases in the flexibility of federal law, offer something almost as good. They offer an opportunity to use hard-won lessons from NCLB to re-set the conversation about assessment.

In 2003, the National Research Council (NRC) published a report called, *Assessment in Support of Instruction and Learning: Bridging the Gap between Large-Scale and Classroom Assessment.* This report:

- highlighted big differences in what large-scale and local assessment are designed and able to do
- explored how more intentional integration of large-scale and local assessment could help create stronger supports for teacher and student learning

The NRC report outlined three core uses for large-scale tests that, under NCLB, were often reduced to long lists of skills:

> The first is ***program diagnosis***. Assessments that make it possible to compare the performance of a large number of students can be used to identify patterns of strengths and weaknesses that are in turn critical for identifying any needed improvements in curriculum or instruction.
>
> Assessments developed for large-scale use, to provide evidence about district- or statewide performance, can also ***exemplify . . .*** the educational goals described in standards and curriculum documents. In other words, assessment tasks and examples of student work make concrete just what students will actually know or be able to do if they meet defined standards.
>
> Large-scale assessments are also useful for one-time ***certification*** or screening; for example, to identify students who are not ready for grade-level work in reading and who need follow-up targeted assessment to determine their specific needs for remediation.

Prophetically, the report also noted the following:

> While the value of large-scale assessments for these purposes is clear, it is equally clear that they are ***not useful for many other important educational purposes particularly that of providing detailed understanding of individual students' performance***. Professional standards are firm on the point that it is not a test itself that can be established as valid, but particular inferences that may be made from the test data (see *National Science Education Standards* (NSES) Standard 13.2, NRC, 1996). . . .
>
> . . . "The best way to help policy makers understand the limitations of an external, once-per-year test for instruction is to recognize that good teachers should already know so much about their students that they could fill out the test booklet for them." ***[emphasis added] pp. 11-12***

Figure 10.1

Large Scale and Local Assessments Have Different Structural Characteristics that Make Them Good for Some Purposes but Ineffective and Inappropriate for Others

BOX 2-2 Unique Characteristics of Large-Scale and Classroom Assessments

Large-Scale Assessments	Classroom Assessments
Provide comparative data, both normative and standards based, that allow policy makers, teachers, parents, and students to make judgments about the adequacy of performance and the specific curricular and instructional areas where improvement is needed.	Must be ongoing and integrated seamlessly into instruction so that teachers and students are receiving frequent but unobtrusive feedback about their progress.
Provide quality feedback to teachers about patterns of errors that could be the target for instructional interventions in the future.	Assess some desired proficiencies in each knowledge domain that cannot be effectively assessed on a large-scale assessment, such as a student-designed experiment or a piece of creative writing revised over time.
Must be cost-effective and feasible; in particular, the benefit to students from information gain must be worth the instructional time lost to testing and test preparation.	Provide quality ongoing feedback to teachers about patterns of errors that could indicate the need for modification of instructional strategies.
	Help teachers to identify and reconstruct students' misconceptions.
	Provide quality feedback to students about their performance and specific guidance about how to improve (most useful when students are given descriptive, criterion-based feedback rather than merely providing number or letter grades):
Results must be reported to stakeholders so as to enable meaningful use of assessment data and forestall misinterpretations.	• Help students to identify and reconstruct their misconceptions.
	Help students to assess their current levels of understanding in relation to well-articulated learning goals and what they, as students, clearly understand to constitute quality work:
	• Involve peer- and self-assessments as well as teacher judgments. • Place more emphasis on allowing students to participate in developing and analyzing the results of the assessments rather than viewing assessments as something that is done to them by teachers.

SOURCE: Adapted from NRC (1993, 1996, 1998, 2000, 2001a, 2001b, 2001c, and 2002), National Council of Teachers of Mathematics (1995), Commission on Instructionally Supportive Assessment (2001), and Popham (in press).

[pp. 9-10]

Reciprocal Accountability

Test reportage under NCLB created an alternate universe of diagnostic information that tests themselves were incapable of producing. These reports misrepresented what standards actually called for and under-reported what tests actually assessed. Worse yet, they tacitly endorsed a set of instructional practices that encouraged rote learning. In short, officially-sanctioned information systems bear much of the responsibility, but little or no accountability, for the failings of NCLB.

As a society, we have been saying for the better part of a century that we want our schools to pay more attention to higher-order thinking and authentic intellectual work. But all the evidence points to a culture of American teaching that is not yet prepared to do that work at scale. In *School Reform from the Inside Out* (2004), Richard Elmore offered a promising entry point for problems of this kind. He called it the "principle of reciprocal accountability:"

> *[You] should be expected to perform at the limits of [your] capacity, but [you] should not be expected to do those things for which [you] do not have the capacity unless [I] accept the joint responsibility with [you] to create that capacity. . . . My authority to command or induce you to do something you are not currently doing depends, in large part, on your capacity to actually do it. You may be motivated to do it. You may agree with me that it should be done. Or you may be willing to do it because just because I have a legitimate grant of authority to require you to do it. But if you can't do it because you do not have the capacity to do it, then my authority is diminished because I have induced or required you to do something you cannot do. I can flog you harder, I can penalize you, I can threaten you, but I cannot make you do something you do not know how to do.*
>
> *Elmore, Richard F. (2004) School Reform from the Inside Out p. 244-45*

It is entirely reasonable to expect teachers to change practices that systematically deny deep understanding to all but a small minority of students. But the failure of reform efforts over many decades to resolve this problem makes it clear that most teachers do not have the capacity to confront this challenge on their own. This is not an indictment of teachers' character. It is an indictment of educational policies that:

- underestimate the depth and complexity of the challenge; and,
- do not create accountabilities for institutional support that match those imposed on school-based practitioners

Systematic misrepresentation of large-scale test results under NCLB is only the most recent case in point.

Between 2010 and 2013, the US Department of Education (DOE) invested $186 million in PARCC to produce new assessment capacity that is vastly superior to most NCLB-era tests. In the words of the 2003 National Research Council report, PARCC can now produce information that ***exemplifies*** what standards call for with "assessment tasks and examples of student work [that] make concrete just what students will actually know or be able to do if they meet defined standards."

This information is useful on its face. But its more important contribution is the models it can provide to support local grade and departmental teams in the hard work of building greater depth into ***classroom*** assignments, assessments and student work. This is where the real work of reforming the grammar of American schooling gets done . . . or not . . . depending on the reciprocal accountability that schools, districts and states are willing to accept for its success.

PARCC represents an important first step toward the more integrated system of large scale and local assessment that was envisioned by the NRC in 2003. Why PARCC missed the opportunity to showcase its most important asset during its first round of test reports is anybody's guess. What is clear is that most of the DOE's investment in PARCC will be squandered if this asset remains buried

in a poorly-indexed warehouse of online pdf files. The original promise of standards-based assessment was to provide educators and parents with meaningful, standards-based information about what students are learning and where they are getting stuck. Making actual items and student responses easily accessible in user-friendly formats during the second round of PARCC reports will go a long way toward delivering on that promise.

PARCC
Partnership for Assessment of Readiness for College and Careers

October 23, 2015

The states that make up the PARCC consortium took the exceptional step of releasing test items from the PARCC assessments to give teachers a powerful tool to inform classroom instruction. The release of the items gives parents insight into the kinds of questions students are seeing on their tests, so assessments aren't a mystery. The test items were built with robust mathematics problems and authentic reading passages selected and reviewed extensively by dozens of educators from PARCC states.

PARCC states see these released items as valuable instructional tools that will give teachers better insight into how students may demonstrate mastery of the standards and how they might be helped on their pathways to academic success—whether in earlier graders or, for older students, college and careers.

The released test questions represent roughly one full test per grade level in each subject area. In addition to the questions, the learning standards associated with each test item are indicated and scoring rubrics are included that show what is required to score at each performance level. Examples of scored student responses are also available for teachers and students to see actual work and the corresponding points earned on the student example.

The PARCC tests were built by educators. They were built on higher standards - meant to challenge students to demonstrate skills that are needed to succeed in everyday life, not to memorize facts. Providing the test questions shows parents, teachers and students the skills that are being measured; problem solving, critical thinking, comprehension and analysis. The examples help students and parents see not only what is being asked, but how the answers are being measured to better understand what is being expected of students.

Together, these materials give educators considerable insight into how the PARCC test measures student understanding of the standards and will help educators plan instruction in their classroom.

This is the first release of items. PARCC is committed to releasing at least as many items in each subsequent year, which will demonstrate the diversity and breadth of items.

See more at: https://prc.parcconline.org/library/using-released-items-instructional-tools

CONCLUSION

Four decades of school effectiveness research, and continuing evidence from dozens of individual schools across Illinois, leave no real doubt that schools are fully capable of:

- offering rigorous and engaging instruction to all students
- reducing gaps in school effectiveness that leave average achievement among Black students far behind that of their Latino and White peers
- reversing declining achievement among low income White students and flattening achievement among White students from middle and upper income households

The policy question is no longer whether these problems can be solved. The policy question is how to solve them at scale. Reasonable people will continue to disagree about the best ways to do that. But all of those ways will require good information.

Between 2005 and 2015, the State of Illinois invested over $315 million in standardized testing mandated by NCLB. At the district level, millions of local dollars and thousands of instructional hours have also been invested in "interim" assessments which tested pretty much the same thing that mandated tests did but reported results differently.

What kind of return has this investment produced?

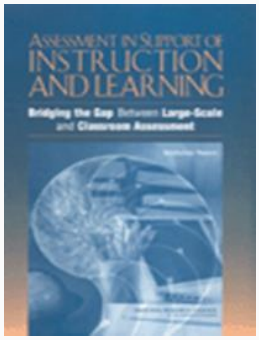
In 2011, the University of Chicago's Consortium on Chicago School Research reviewed two decades of student-level data to assess the progress of school reform efforts in Chicago. A key finding of that study was that most publicly reported data were "simply not useful" for gauging actual progress in student achievement.

Chicago has not only been at the forefront of school reform policies but also has been ahead of most of the rest of the country in collecting data and tracking student and school performance. Yet, even with a heavy emphasis on data use and accountability indicators, the publicly reported statistics that are used by CPS and other school districts to gauge progress are simply not useful for measuring trends over time. . . . As there is a greater push at both the state and federal levels to use data to judge student and school progress, we must ensure that the statistics that are used are comparable over time. Otherwise, future decisions about school reform will be based on flawed statistics and a poor understanding of where progress has been made.

Trends in Chicago's Schools across Three Eras of Reform, Luppescu et. al. (2011) p. 8

If anything, the findings of this study deliver an even harsher indictment than the Consortium study did. All of the trends reported in PART 1 have been developing for years. But officially-endorsed reporting practices described in PART 2 knowingly distorted state and local data and left parents, educators and policy makers unaware of what was actually happening. Small wonder that most people missed major shifts in local and regional achievement, missed chronic stagnation in middle school achievement, and couldn't fully recognize sustained improvements in Latino achievement statewide. Small wonder that many practitioners came to question the value of data-driven decision making. And small wonder that growing numbers of Illinoisans are now asking hard questions about the usefulness of standardized testing as a whole.

A new generation of statewide assessments offers state and district leaders a unique opportunity to recommit to the original promise of standards-based test reportage . . . honest, high-quality information that is meaningful and useful for parents, educators, and the public at large. It is way past time to make good on that promise.



Assessment in Support of Instruction and Learning

Bridging the Gap Between Large-Scale And Classroom Assessment (2003)

National Research Council 2003

http://www.nap.edu/catalog/10802/assessment-in-support-of-instruction-and-learning-bridging-the-gap

"The gap between classroom and large-scale assessments has caught the attention of several National Research Council (NRC) committees, and one result has been a clear consensus that instruction and learning are best supported in educational systems when large-scale and classroom assessments are aligned with each other and with standards, curriculum, instruction, and professional development' [p. 2]

LARGE-SCALE ASSESSMENTS

While large-scale assessments can be controversial, and are easily misused, they are an important way of obtaining certain kinds of extremely valuable information about students. Large-scale assessments, those that are designed to provide evidence about large numbers of students, are the primary means by which accountability evidence is obtained in the United States. Indeed, there is little dispute that accountability—the provisions made for those who use, fund, and oversee public education to review and evaluate its effectiveness—is a crucial element in the continued success of public education.

As Lorrie Shepard of the School of Education, University of Colorado, Boulder, outlined at the workshop, there are three particular uses for which largescale tests are essential. The first is *program diagnosis*. Assessments that make it possible to compare the performance of a large number of students can be used to identify patterns of strengths and weaknesses that are in turn critical for identifying any needed improvements in curriculum or instruction. Assessments developed for large-scale use, to provide evidence about district- or statewide performance, can also *exemplify*, as Shepard termed it, the educational goals described in standards and curriculum documents. In other words, assessment tasks and examples of student work make concrete just what students will actually know or be able to do if they meet defined standards. Large-scale assessments are also useful for one-time *certification* or screening; for example, to identify students who are not ready for grade-level work in reading and who need follow-up targeted assessment to determine their specific needs for remediation.

Shepard also noted that large-scale assessments often provide teachers an opportunity for effective professional development. Development of tests, scoring, curriculum development, and standards-based professional development are all occasions when efforts to improve classroom assessment strategies can be woven into the program. Shepard argues that more could be gained through these opportunities if teachers had improved access to materials that model teaching for understanding, such as extended instructional activities, formative assessment tasks, and scoring rubrics with summative assessments built in to them.

While the value of large-scale assessments for these purposes is clear, it is equally clear that they are not useful for many other important educational purposes, particularly that of providing detailed understanding of individual students' performance. Professional standards are firm on the point that it is not a test itself that can be established as valid, but particular inferences that may be made from the test data (see *National Science Education Standards* (NSES) Standard 13.2, NRC, 1996). . . .

As Shepard stated, "The best way to help policy makers understand the limitations of an external, once-per-year test for instruction is to recognize that good teachers should already know so much about their students that they could fill out the test booklet for them." Shepard listed some of the contrasts, shown in Box 2-3, between large-scale and classroom assessments that make clear why different instruments are usually needed for different purposes

[pp. 11-12]

BIBLIOGRAPHY

ACT. (2007). *ACT Technical Manual*. Iowa City: ACT, Inc.

ACT. (2014). *Technical Manual, The ACT*. Retrievable at: http://www.act.org

ACT. (2008). *The Forgotten Middle*. Iowa City: ACT, Inc.

Advance Illinois. (2014). *Making Assessments Work: Supporting Teaching and Learning in the Common Core Era.* Retrievable at: http://www.advanceillinois.org/publications/making-assessments-work/

Allensworth, E.M., Correa, M., and Ponisciak, S. (2008). *From High School to the Future: ACT Preparation--Too Much, Too Late.* Chicago: University of Chicago Consortium on School Research.

Allensworth, E.M., Gwynne, J.A., Moore, P. and de la Torre, M. (2014). *Looking Forward to High School and College: Middle Grade Indicators of Readiness in Chicago Public School.* Chicago: University of Chicago Consortium on School Research.

Annie E. Casey Foundation. (2012). *Double Jeopardy: How Third Grade Reading Skills and Poverty Influence High School Graduation.* Retrievable at: http://www.aecf.org/resources/double-jeopardy/

Black, P. and Wiliam, D. "Inside the Black Box," *Phi Delta Kappan*, September 2010, Vol. 92 (1) pp 81-90.

Brookhart, S. (2010). *How to Assess Higher Order Thinking Skills in Your Classroom.* Alexandria: Association for Supervision and Curriculum Development.

Bryk, A.S., Thum, Y.M., Easton, J.Q., Luppescu, S. (1998). *Academic Productivity of Chicago Public Elementary Schools.* Chicago: University of Chicago Consortium on School Research.

Carmichael, S.B., Martino, G., Porter-Magee, K., Wilson, W.S. (2010). *The State of State Standards—and the Common Core in 2010.* Washington, D.C.: Thomas B. Fordham Institute

Cavanagh, Sean. "Demand for Testing Products, Services on the Rise." *Education Week,* Vol. 33 (6), October 1, 2013, pp. 16-17,

Commission on Standards for School Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics.* Reston: National Council of Teachers of Mathematics.

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement*. (NCEE 2013–4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Council of Chief State School Officers. (2014). "Criteria for Procuring and Evaluating High-Quality Assessments." Retrievable at: http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

Cronin, J., Dahlin, M., Adkins, D., Kingsbury, G.G. (2007). *The Proficiency Illusion.* Washington, D.C.: Thomas B. Fordham Institute.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., Ho, A., Linn, R.L., Pearson, P.D., Popham, J., Resnick, L., Schoenfeld, A.H.,

Shavelson, R., Shepard, L.A., Shulman, L., Steele, C.M. (2013). *Criteria for High-Quality Assessment.* Palo Alto: Stanford Center for Opportunity Policy in Education, Stanford University.

Easton, J.Q., Correa, M., Luppescu, S., Park, H-S,, Ponisciak, S., Rosenkranz, T., and Sporte, S. (2003). *How Do They Compare? ITBS and ISAT Reading and Mathematics in the Chicago Public Schools, 1999 to 2002.* Chicago: University of Chicago Consortium on School Research.

Elmore, R.F. (2004). *School Reform from the Inside Out.* Cambridge: Harvard University Press.

FairTest. **(2012)**. **"**What's **Wrong with Standardized Tests?" Retrievable at:** http://www.fairtest.org/facts/whatwron.htm

Fraser, Steven (Ed.). (1995). *The Bell Curve Wars: Race, Intelligence, and the Future of America*. New York: Basic Books.

Gewertz, Catherine. "Test Group Rethinks Questions." *Education Week*, Vol. 32 (13), December 5, 2012, pp. 1, 24.

Gordon Commission on the Future of Assessment in Education. (2013). *A Statement Concerning Public Policy.* Princeton: The Gordon Commission. Retrievable at: http://www.gordoncommission.org/rsc/pdfs/gordon_commission_public_policy_report.pdf

Gould, Steven J. (1996). *The Mismeasure of Man.* New York: W.W. Norton Co..

Guskey, T. (2015). *On Your Mark.* Bloomington: Solution Tree Press.

Harris, Douglas. (2011). *Value-Added Measures in Education*. Cambridge: Harvard Education Press

Hattie, John. (2009). *Visible Learning.* New York: Routledge

Hernstein, R. and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: Free Press.

Ho, Andrew D. "The Problem with 'Proficiency:' Limitations of Statistics and Policy under No Child Left Behind," *Educational Researcher*, August-September, 2008, Vol. 37 (6), pp. 351-360.

Illinois State Board of Education Assessment Division. (2012). *Illinois Standards Achievement Test 2001 Technical Manual.* Springfield: Illinois State Board of Education.

Illinois State Board of Education Assessment Division. (2006). *Illinois Standards Achievement Test 2006 Technical Manual.* Springfield: Illinois State Board of Education.

Illinois State Board of Education Assessment Division. (2012). *Illinois Standards Achievement Test 2012 Technical Manual.* Springfield: Illinois State Board of Education

Illinois State Board of Education, *Illinois Standards Achievement Test 2013 Technical Manual.* http://www.isbe.net/assessment/pdfs/isat_tech_2013.pdf

Jacob, B.A. and Rockoff, J.E. (2011). *Organizing Schools to Improve School Achievement.* Washington, DC: The Hamilton Project, Brookings Institution. Retrievable at: http://www.edweek.org/media/gradeconfiguration-13structure.pdf

Kun, Yuan and Vi-Nhuan Le. (2012). "Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests." RAND Education. Retrievable at: https://www.rand.org/content/dam/rand/pubs/working_papers/2012/RAND_WR967.pdf

Lagemann, E.C. (2002). *The Elusive Science*. Chicago: University of Chicago Press.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S. and Busick, M.D. (2012). *Translating the Statistical Representation of Effects of Education Interventions into More Readily Interpretable Forms* NC SER 2013-3000, Institute for Education Sciences, US Dept. of Education.

Luppescu, S., Allensworth, E., Moore, P., de la Torre, M., Murphy, J. and Jagesic, S. (2011). *Trends in Chicago's Schools across Three Eras of Reform.* Chicago: University of Chicago Consortium on School Research.

National Center for the Improvement of Educational Assessment. (2015). *Guide to Evaluating Assessments Using the CCSSO Criteria for High Quality Assessments: Focus on Test Content.* Retrievable at: http://www.nciea.org/publication_PDFs/Guide%20to%20Evaluating%20CCSSO%20Criteria%20Test%20Content%20020316.pdf

National Research Council. (2003). *Assessment in Support of Instruction and Learning: Bridging the Gap between Large-Scale and Classroom Assessment.* Retrievable at: https://www.nap.edu/catalog/10802/assessment-in-support-of-instruction-and-learning-bridging-the-gap

Newmann, F.M., Bryk, A.S., Nagaoka, J. *Authentic Intellectual Work and Standardized Tests: Conflict or Co-existence?* Chicago: University of Chicago Consortium on School Research.

Northwest Evaluation Association. (2011). *NWEA 2011 RIT Scale Norms.* Retrieved from *http://www.murray.cps.edu/pdf2013-14/NWEA_2011_NormsReportRead_Math.pdf*

Payne, C. (2008). *So Much Reform, So Little Change*. Cambridge: Harvard Education Press.

Peterson, P.E., Barrows, S. and Gift, T. "After Common Core, States Set Rigorous Standards," *Education Next,* Vol. 16 (3), Summer 2016. Retrievable at: http://educationnext.org/files/ednext_XVI_3_peterson_standards.pdf

Phillips, G. (2016). *National Benchmarks for State Achievement Standards*. American Institutes for Research. Retrievable at: *http://www.air.org/resource/national-benchmarks-state-achievement-standards*

Regional Education Laboratory Central. "What does the research say about sixth-grade placement? Should they be in an elementary school or a middle school?" Retrievable at: https://www.relcentral.org/what-does-the-research-say-about-sixth-grade-placement-should-they-be-in-an-elementary-school-or-a-middle-school/

Schultz, S.R., Michaels, H.R., Dvorak, R.N., Wile, C.R.H. (2016). *Evaluating the Content and Quality of Next Generation High School Assessments: Final Report.* Alexandria: Human Resources Research Organization (HumRRO).

Schwerdt, G., & West, M. R. (2011). *The impact of alternative grade configurations on student outcomes through middle and high school.* Cambridge, MA: Institute for Economic Research, Harvard University and Harvard Graduate School of Education. Retrievable at:

http://www.hamiltonproject.org/assets/legacy/files/downloads_and_links/092011_organize_jacob_rockoff_paper.pdf

Stigler, J. and Hiebert, J. "Closing the Teaching Gap," *Phi Delta Kappan,* Vol. 91 (3), November, 2009, pp. 32-37.

Stigler, J. and Hiebert, J. (2009). *The Teaching Gap, Reissue Edition*. New York: Free Press.

Tyack, D. (1974). *The One Best System*. Cambridge: Harvard Education Press.

Tyack, D. and Cuban, L. (1995). *Tinkering Toward Utopia*. Cambridge: Harvard Education Press.

Tyack, D. and Tobin, W. "The Grammar of Schooling," *American Educational Research Journal*, Vol. 31 (3), September 1994, pp. 453-479.

Wiggins, Grant. "Why We Should Stop Bashing State Tests," *Educational Leadership,* Vol. 67 (6), March 2010, pp. 48-52.

Wiggins, G. and McTighe, J. (2005). *Understanding by Design, 2nd Edition.* Alexandria: Association for Supervision and Curriculum Development.

Wiliam, D. and Leahy, S. (2015). *Embedding Formative Assessment*. West Palm Beach: Learning Sciences International.

Yettick, Holly. "Interim Assessments Yield Disappointing Results in Indiana Study," *Education Week*, April 5, 2014. Retrievable at: http://blogs.edweek.org/edweek/inside-school-research/2014/04/large_study_suggests_that_inte.html

Zavitkovsky, Paul E. (2009). *Something's Wrong with Illinois Test Results*. Chicago: Urban Education Leadership Program, University of Illinois at Chicago.