Most existing critiques of large-scale interim systems have focused on their inability to generate valid, small-grain information . . . in particular, small-grain information about individual students that can be leveraged for better instructional decision making. Far less attention has been paid to the confidence educators, parents and policy makers can place in the large-grain, normative reportage (scale score trends, national percentile ranks, etc.) that most interim systems also produce. Reports of this kind are an important feature of interim systems because they offer a much more attractive range of program monitoring and public communication options than the reportage provided by most statewide testing systems.

The current study uses descriptive statistical comparisons to illustrate consequential problems that arose in three Illinois school districts when two of those districts depended too heavily on large-grain scoring trends reported by the Northwest Evaluation Association (NWEA) during the five years prior to the pandemic. In all cases, these problems arose from NWEA scoring trends that were substantially more positive than comparable reportage from the National Assessment of Educational Progress (NAEP) and other widely-respected systems.¹

Introduction

“For a while in my life, I was senior research director at Educational Testing Service in Princeton, New Jersey. We had three different teams working on trying to get good diagnostic information from high-stakes accountability tests . . . Nothing came out of that work. It was simply, in most practical situations, impossible to get any insights into what you might do with this information apart from just looking at the total score and seeing ‘this kid’s struggling, this kid’s OK.’

. . . I would love it if we could derive diagnostic information from large-scale standardized tests. Everybody wants things to do double duty. But what we find in assessment is that, generally, when we try to make things do double duty, they don’t do either of them very well.”

Dylan Wiliam, Learning Sciences International Conference on Formative Assessment (April 2022)

A fundamental problem with the testing requirements of No Child Left Behind was that they grossly underestimated what it would take for states and their vendors to develop robust, standards-based testing systems that could help inform instruction. So faced with tight implementation deadlines, most states did something way less than robust. With quiet cooperation from the U.S. Department of Education, they used cut-scores,² content strands³ and an assortment of other pseudo-psychometric reporting tools to dress up existing norm-referenced reporting systems in standards-based clothing. From Day One, both the summative and “formative” reportage that most state systems produced was data rich and information poor.⁴
Over the years, growing dissatisfaction with the value of statewide test reportage has led many school districts to purchase large-scale interim systems directly from commercial vendors. In my own state of Illinois, for example, 70% of districts now spend $50 million annually on commercial interim systems, all paid for from local revenues. The promise of these systems is that big advances in large-scale test design and computer-adaptive technology now make it possible to assess a wide range of small- and large-grain learning information quickly, and report back results almost instantaneously. Some commercial systems also claim that interim testing deepens school and district capacity to support rich, classroom-based assessment that has long been known to increase student learning by between 0.4 and 0.7 standard deviations a year.

Regrettably, there is no independent evidence that commercial interim systems are actually making these kinds of contributions. And sadly, the failure of interim systems to deliver on their promise is something independent assessment professionals have been warning practitioners about for most of the past two decades.

Within a year after NCLB became law, the National Research Council brought leading educators and assessment professionals together from around the country. The purpose of this gathering was to envision practical ways to “bridge the gap between large-scale and classroom assessment.” Program content made it clear that, in NCLB’s new world of high-stakes accountability, the gap that assessment professionals worried about most was the big structural divide between:

- large-scale standardized assessments designed to monitor large-grain learning and growth across large populations; and,
- small-scale, content-specific assessments designed to generate small-grain, diagnostic information that can actively inform classroom teaching and learning.

A workshop led by Lorrie Shepard (University of Colorado—Boulder) paid particular attention to how the psychometric DNA of large-scale assessments limits the kinds of inferences that can be drawn from them.

“While the value of large-scale assessments for [several essential] 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 this 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)” [emphasis added]

Assessment in Support of Instruction and Learning: Bridging the Gap between Large-Scale and Classroom Assessment (2003) p. 9

Two decades later, independent assessment professionals have grown even more wary about the reach and ambition of large-scale interim systems, including so-called “through-year” designs that replace year-end summative tests with statewide, interim-based systems. The flaws of through-year reporting systems may be less obvious than reporting systems used under NCLB. But many through-year designs give every indication that they will be even more data rich and every bit as information poor as their NCLB-era predecessors.
Shortly after pandemic disruptions raised new urgency everywhere about strengthening school and district resources, Scott Marion (National Center for Improvement of Educational Assessment) wrote about these developments in close to plaintive tones.

“Educators and school leaders are trying to figure out how to deal with extraordinary school interruptions this year and an uncertain educational future. They are desperate for tools and resources to help meet these needs. The allure of an assessment that can be administered quickly and produce results to inform teachers, school leaders, and policymakers could be too tempting to resist. Except, it is fiction. It is unfair to dangle false hopes in front of educators who are trying to do what is best for kids.

“I am not ruling out the parsimonious use of some district or state assessments to inform resource allocation and related decisions, but we need to be clear that calling an assessment ‘diagnostic’ or ‘instructionally-useful’ does not make it so. Assessment leaders must be specific about the uses assessments can and cannot support. In fact, this would be a welcome change that should last well beyond the current pandemic.”

“You Say Tomato: Concerns about the Diagnostic Assessment Rhetoric” June 20, 2020, NCIEA Centerline

The Current Study

As the paragraphs above suggest, most existing critiques of large-scale interim systems have focused on their inability to generate valid, small-grain information . . . in particular, small-grain information about individual students that can be leveraged for better instructional decision making. Far less attention has been paid to the confidence end users can place in the large-grain, normative reportage (scale score trends, national percentile ranks, etc.) that most interim systems also produce. Reports of this kind are an important feature of interim systems because they offer a much more attractive range of program monitoring and public communication options than the reportage provided by most statewide testing systems.

The current study uses descriptive statistical comparisons to illustrate consequential problems that arose in three Illinois school districts when two of those districts depended too heavily on large-grain scoring trends reported by the Northwest Evaluation Association during the five years prior to the pandemic. During this period, educators and policy makers in both districts used NWEA scoring trends as the central source of authority for assessing the status and progress of student learning, and for communicating that information to parents and the public at large.

In all the cases described below, problems arose from NWEA scoring trends because they were substantially more positive than comparable reportage from NAEP and other large-scale systems. These anomalies were especially noteworthy because they arose during a period when NWEA technical manuals indicated that the organization was intensifying its efforts to report accurate, nationally representative norms from its large but not-fully-representative national testing sample.

Case #1: Problematic Scale Score Trends over Time for Chicago Public Schools

Scale scores are the foundation for all standardized reporting systems because they translate responses from thousands of content-specific test items into statistically-equivalent learning units that can be compared across tests and over time. Although scoring scales for different testing systems differ numerically, different systems that purport to measure the same kinds of learning should produce roughly equivalent scoring trends for the same test populations.
By way of illustration, Figure 1.1 shows scale score trends over time in 8th grade math achievement as reported by NAEP/TUDA and two different year-end accountability tests used by the State of Illinois between 2009 and 2019 (ISAT and PARCC).\textsuperscript{16} Blue lines show TUDA and State of Illinois test results for Chicago Public Schools. Green lines show NAEP and State of Illinois test results for all Illinois students tested statewide. The dotted gray line in Figure 1.1 shows NAEP scale scores at the 50th percentile of all public school students tested nationwide. Specific numerical scores from each system are not comparable across systems, but scoring trends over time are.

**Figure 1.1**
Roughly Congruent 8th Grade Math Trends Across Tests and Over Time

DATA SOURCES: NAEP Data Explorer [https://www.nationsreportcard.gov/ndecore/xplore/NDE](https://www.nationsreportcard.gov/ndecore/xplore/NDE)

To the average person, it might seem remarkable that trends over time reported from very different standardized tests all reflect roughly the same trends over time. But typically, they do.\textsuperscript{17} In this case, the consistency reported across tests for Chicago, and for the State of Illinois and the nation as a whole is pronounced. What made this consistency possible was that all tests . . . each in their own way . . . measured roughly the same kind of math learning, and sampled comparable test populations in highly representative ways.
For many years, the Northwest Evaluation Association has claimed that the learning assessed by its MAP Growth system is closely aligned with state and national standards and that reportage involving national norms is highly representative of all students tested nationwide. But if that is true, something else must account for why the strength and upward movement of MAP scale scores for Chicago differed so substantially from trends that were reported by NAEP/TUDA and statewide testing for the same students (see Figure 1.2 below). These differences were especially consequential in Chicago because MAP scoring trends from 2013 through 2019 were the central source of authority that district leaders used to assess the status and progress of student learning, and for communicating that information to parents and the public at large. In 2014, for example, MAP data became the principal measure of achievement and growth for Chicago’s School Quality Rating Policy (SQRP). In 2019 they were the primary empirical foundation for the district’s five-year strategic vision for 2019-2024.

**Figure 1.2**
Anomalous 8th Grade Math Trends in Chicago from Spring MAP Testing

DATA SOURCES: NAEP Data Explorer [https://www.nationsreportcard.gov/ndecore/xplore/NDE](https://www.nationsreportcard.gov/ndecore/xplore/NDE)
Chicago Public Schools Assessment Files [https://www.cps.edu/academics/student-assessments/#Assessments38](https://www.cps.edu/academics/student-assessments/#Assessments38)
Case #2: Inflated National Percentile Ranks for Chicago Public Schools

In 2015, NWEA published an updated set of norms for achievement status and growth. A prominent part of the 2015 norming process was to enhance the use of the “model-based post-stratification weighting procedure” which was first employed by NWEA in 2011. A core reason for introducing this procedure in 2011 was to “create estimates that are more completely representative of kindergarten through grade 10 students attending public schools in the U.S. and allow full inferential support.”

The 2015 norms manual elaborated on the post-stratification weighting process and its intended outcomes as follows:

“All test event records used in this study were from operational MAP tests. Only tests that were routinely used in schools during the study’s 3-year timeframe (fall 2011 through spring 2014) were included. Restricting that study to operational tests is, of course, accompanied by the threats that convenience samples pose to studies intended to generalize to a population (Angoff, 19784; Kolen, 2006). To counter these threats and to achieve sample representativeness to the target (U.S. English-speaking school-age) population, a post-stratification procedure was employed. This approach avoids the time and expense needed to recruit special samples to achieve sample representativeness.

NWEA 2015 MAP Norms for Student and School Achievement Status and Growth, page 6

This and other aspects of NWEA’s enhanced norming process led NWEA’s Vice President of Research and Chief Academic Officer to introduce the 2015 manual by confidently assuring NWEA users that the 2015 norming process was thorough and that its numbers and inferences could be trusted.

“More important than convenience, however, is quality—both in terms of accuracy and utility. In addition to the careful creation of this study, we have subjected the findings and report to numerous internal reviews and have engaged external experts to review and critique the report. Its release represents our assurance that the report is ready for use by NWEA partners with confidence and that the numbers and the inferences can be trusted.”

NWEA 2015 MAP Norms for Student and School Achievement Status and Growth, page i

A month prior to the publication of NWEA 2015 norms reports, the National Center for Education Statistics released detailed results from the 2015 NAEP and its urban district counterpart, the Trial Urban District Assessment (TUDA). The release of NAEP and TUDA data made it possible to compare the results of TUDA’s fourth and eighth grade reading and math tests for Chicago with comparable NWEA results from spring 2015. Given the sustained attention that NWEA had paid to developing highly representative national norms in 2011 and 2015, large differences in results reported using 2015 NWEA and NAEP norms were surprising.

Figures 2.1 and 2.2 summarize differences in the percentile ranks that MAP and NAEP norms assigned to average scale scores for fourth and eighth graders in Chicago in 2015. They reveal substantial discrepancies in the values reported by MAP and NAEP for average student achievement compared with all students tested nationwide. The size of these discrepancies was:

• 0.70 to 0.85 grade equivalents (roughly 0.3 to 0.4 standard deviations) at grade four
• 0.90 to 1.20 grade equivalents (roughly 0.4 to 0.6 standard deviations) at grade eight
While many factors can account for smaller discrepancies in normative reportage from different standardized tests, the magnitude of the discrepancies revealed in Figures 2.1 and 2.2 raises validity questions that no district can afford to ignore.

**Figure 2.1**

*2015 MAP & NAEP/TUDA Achievement Comparison for Chicago Public Schools*

![Figure 2.1](image)

**DATA SOURCES:** NAEP Data Explorer [https://www.nationsreportcard.gov/ndecore/xplore/NDE](https://www.nationsreportcard.gov/ndecore/xplore/NDE)
Chicago Public Schools Assessment Files [https://www.cps.edu/academics/student-assessments/#Assessments38](https://www.cps.edu/academics/student-assessments/#Assessments38)

**Figure 2.2**

*2015 MAP & NAEP/TUDA Achievement Comparison for Chicago Public Schools*

<table>
<thead>
<tr>
<th>2015 Norms</th>
<th>National Percentile of Average Scale Score</th>
<th>Grade Equivalent of Average Scale Score</th>
<th>Net Difference between MAP and NAEP/TUDA Grade Equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>NAEP/TUDA</td>
<td>MAP</td>
</tr>
<tr>
<td><strong>GRADE 4</strong></td>
<td>Reading</td>
<td>47</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Math</td>
<td>49</td>
<td>38</td>
</tr>
<tr>
<td><strong>GRADE 8</strong></td>
<td>Reading</td>
<td>61</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Math</td>
<td>58</td>
<td>42</td>
</tr>
</tbody>
</table>

**DATA SOURCES:** NAEP Data Explorer [https://www.nationsreportcard.gov/ndecore/xplore/NDE](https://www.nationsreportcard.gov/ndecore/xplore/NDE)
Chicago Public Schools Assessment Files [https://www.cps.edu/academics/student-assessments/#Assessments38](https://www.cps.edu/academics/student-assessments/#Assessments38)

To explore these differences further, the Center for Urban Education Leadership at the University of Illinois-Chicago (CUEL) ran a parallel comparison of MAP and PARCC results for the Chicago Public Schools in 2016. This comparison drew from a database of normalized, statewide, student-level test results that the CUEL has used to conduct multiple studies of achievement patterns in Chicago and statewide since the early years of NCLB.
Normalized Statewide Scoring Distributions

All large-scale standardized achievement measures in the U.S. are highly correlated with student demographics. Since Illinois has been the most demographically representative state in the nation for at least the last two decades, it is not surprising that nationwide NAEP scoring distributions for All U.S. and the State of Illinois have been virtually indistinguishable. The same has been true for the match between normalized NAEP scoring distributions for Illinois and normalized scoring distributions from annual, statewide accountability testing (see Figure 1.1). Like earlier statewide tests, normalized scoring distributions for PARCC testing that began in 2015 followed this same pattern. This congruence has been helpful for analytic purposes because NAEP/TUDA testing is only conducted every other year and only occurs at grades four and eight.

By applying 2015 State of Illinois PARCC norms to 2016 PARCC scoring distributions for Chicago, it was possible to make direct comparisons between average Chicago achievement based on 2015 NWEA norms and average Chicago achievement based on statewide PARCC norms. The results of this comparison are displayed in Figure 2.3. They show that normative differences in 2016 MAP and PARCC reportage for Chicago were quite similar to the differences in 2015 MAP and NAEP reportage that are summarized in Figures 2.1 and 2.2.

Figure 2.3
Percentage of Chicago Student Scoring At or Above the 50th Percentile Spring 2016 MAP and PARCC Tests
[Based on 2015 National Norms MAP and 2015 State of Illinois Norms for PARCC]

In 2016:
- the percentage of students who scored at or above the 50th percentile based on national NWEA norms was 10 to 30 points higher than the percentage of students who scored at or above the 50th percentile based on statewide PARCC norms
the majority of students tested at every reported grade level scored at or above grade-level averages based on NWEA norms, while the majority of students tested scored below grade-level averages based on statewide PARCC norms.

- differences between NWEA and PARCC norms were considerably larger in grades seven and eight than they were in grade three through six.

Once again, these differences were especially consequential in Chicago because, from 2013 through 2019, NWEA scoring trends were the central source of authority that district leaders used for assessing and communicating the status and progress of student learning at the classroom, school and district level.

Case #3: Conflicting Measures of High School Readiness in Two School Districts in Evanston, Illinois

The City of Evanston is a demographically diverse suburban community that is home to Northwestern University and shares its southern border with the City of Chicago. Like many Illinois suburbs, Evanston has two, independent public school districts that serve students from pre-kindergarten through grade twelve. Evanston-Skokie District 65 serves students from pre-kindergarten through grade eight; Evanston Township High School District 202 serves students from grades nine through twelve.

For many years, the District 65 and District 202 school boards have used a Joint Committee that meets regularly throughout the year to help increase the coherence of policy and practice across districts. During 2013, a major focus of the Joint Committee was the development of joint literacy goals. In 2014, both Boards adopted a common literacy framework “to ensure that all students are proficient readers and are college and career ready by the time they reach 12th grade.”

Five years later, the Joint Committee met to discuss a report (see Appendix A) that had been presented and discussed two days earlier at a regularly scheduled meeting of the District 202 school board. The report compared spring MAP scores for graduating District 65 eighth graders with scores obtained on District 202’s STAR 360 screening test taken during the first month of ninth grade.

Jointly prepared by District 202’s Reading Specialist and Associate Principal for Instruction and Literacy, the report showed starkly different normative results from MAP and STAR testing. Figure 3.1 shows the two major patterns that were detailed in the report:

1. MAP proficiency estimates based students scoring at the 50th percentile or above are consistently higher, and often much higher, than STAR proficiency estimates based on a grade equivalent score of 8.3 (third month of 8th grade)
2. The largest discrepancies between MAP and STAR proficiency estimates are disproportionally reflected in Black and Hispanic test results.
Figure 3.1
Percent of Incoming Ninth Graders Who Test “Proficient” on NWEA MAP and Renaissance Star 360 Reading Tests

![Bar chart showing percent proficient on MAP and STAR tests for Black, Latinx, and White students.]


A lengthy summary of the discussion these data prompted (see Appendix B) was reported a week later in the Evanston RoundTable, an award-winning local newspaper available online and distributed every other week to Evanston households. The small sampling of comments reported below capture the overall spirit of the discussion.

**Committee Member 1:** “The fact of the matter is we’re getting MAP data only, so that’s how we’re placing [students] and then we’re . . . finding after the fact we’re under-identifying kids who need help—in big numbers . . . I’m looking for a sense of urgency to get this fixed . . . and resolve these data problems . . . to advance the goal of making sure each student by grade 12 is proficient.”

**ETHS Principal:** “[teachers] tell us consistently that STAR is for more reliable because it’s what they see on the ground as they teach every day”

**Committee Member 2:** “There needs to be a sense of urgency around this . . . We need to have a better plan and it needs to happen now.

**ETHS Associate Principal for Instruction and Literacy:** “What we really need to consider is do we want to have a measure of reading proficiency and growth for grades 3-12. And if we do, we need to adopt a system that will allow us to do that. If we’re going to use different measures, I think we’re going to continue having these frustrations.”
In a subsequent interview with the Evanston RoundTable, a co-author of the January 16 report offered additional details about how “proficiency” was defined for the MAP and STAR estimates reported in Figure 3.1:

- District 202 had historically assumed that scoring at the 50th percentile on the spring MAP would be at least close to grade-level proficiency.
- District 202 also assumed that, even if MAP norms were somewhat inflated, a 50th percentile score would still be sufficiently rigorous to exceed a STAR grade equivalent of 8.3, the score District 202 used as preliminary screen for supplemental support.

During the month following the Joint Committee meeting, curriculum and assessment personnel from both districts reviewed MAP and STAR scoring patterns from a number of student cohorts. Careful analysis of these patterns revealed that the best predictor for a STAR grade equivalent of 8.3 in the first month of ninth grade was a MAP scale score of 227 in the spring of eighth grade. A spring scale score of 227 in eighth grade is at the 62nd percentile of NWEA’s 2020 national norms. A STAR grade equivalent of 8.3 in the first month of ninth grade is at roughly the 38th percentile of STAR’s national norms.

These details reveal that the already large discrepancies in proficiency estimates that are illustrated in Figure 3.1 did not capture the full magnitude of differences between MAP and STAR norms. Said another way, the competing definitions of MAP and STAR proficiency that are reflected in Figure 3.1 still understated the actual differences between MAP and STAR norms by roughly half a grade equivalent.

Case #4: From National Exemplar to a Little Below Average

Chicago’s Cautionary Tale: Depending too heavily on NWEA metrics that under-reported the precipitous collapse of cohort growth over time from 2013 through 2019

From 2013 through 2019, Chicago was one of the thousands of school districts across the nation that used MAP Growth systems to anchor systemwide assessment and accountability culture in grades kindergarten through eight. And like many other districts, Chicago had a variety of reasons for investing in commercial interims. Beyond general dissatisfaction with the speed and quality of annual statewide test reportage, legislative changes encouraged by Obama-era Race to the Top initiatives led Chicago and many other districts to seek out commercial systems that could support value-added evaluation of schools, principals and classroom teachers. In 2014, Chicago closely aligned its School Quality Rating Policy (SQRP) with MAP Growth reports for grades PK-8. Under this policy, 60% of each school’s annual rating was driven by MAP achievement and growth measures in the spring of each year.

The timing of Chicago’s MAP Growth adoption was especially significant because it came on the heels of more than a decade of nation-leading growth beyond expected gains from end of grade three through the end of grade eight. The most widely publicized documentation of this growth came in a study published in November 2017 by Sean Reardon and Rebecca Hinze-Pifer for cohort learning and growth that occurred from 2009 through 2014.23
Exactly 30 years after then-Secretary of Education William J. Bennett labeled Chicago Public Schools the worst in the nation, new research shows that Windy City schools now lead the country in academic growth.

A new study by Stanford University researchers Sean Reardon and Rebecca Hinze-Pifer tracked reading and math test scores growth among public school students from 2009 to 2014. Across racial groups, the researchers found that Chicago students learned significantly faster from grades 3 to 8 than did students in nearly all other U.S. school districts—gaining about six years’ worth of learning in five years.

The data used in the Reardon/Hinze-Pifer study were part of an ambitious, first-of-its-kind project that is now known as the Stanford Data Education Archive (SEDA). A unique feature of SEDA reportage is that it equates scoring from all fifty state tests and uses national NAEP norms to report cohort achievement and growth in national grade equivalents for over 11,000 school districts nationwide.24

Since the early years of NCLB, the Center for Urban Education Leadership at the University of Illinois—Chicago has maintained a similar but less complex data archive for all 850+ school districts in the State of Illinois. Like SEDA, the UIC archive uses de-identified, student-level files for all students tested statewide. But unlike SEDA, the UIC archive contains complete student-level scoring information from 2001 onward and uses annual state wide norms for all grades tested instead of NAEP’s every-other-year, national norms for grades four and eight. However, because Illinois’ enrollment demographics are highly representative of enrollment demographics nationwide, most State of Illinois norms have fallen within a standard error of national NAEP norms for at least two decades.

Six months prior to the publication of the Reardon/Hinze-Pifer study, researchers from the Center for Urban Education Leadership published an extensive series of studies called Upstate/Downstate that detailed statewide achievement trends over time along with individual profiles for the state’s 55 largest K-12 schools districts25. While these studies did not specifically track cohort growth from grade three through grade eight, they did report findings similar to the Stanford study about achievement and growth in Chicago under NCLB (see Figure 4.1).
Systemwide use of MAP Growth did not begin in Chicago until 2013. But when the Stanford report was published in November 2017, there was a widespread belief that MAP systems were helping to sustain, if not amplify, the nation-leading growth rates that Stanford and UIC had reported for the years leading up to 2013. That perception was strongly reinforced by annual NWEA data that reported year-over-year growth compared with national NWEA growth norms (see Figure 4.2). While these reports did show flattening or slightly declining growth in grades three through eight, annual growth rates in almost all grades remained at or above NWEA growth norms, especially in grades 3, 7 and 8.
In a written communication that the Evanston *RoundTable* published in 2018, senior leaders from NWEA endorsed this association:

“John Cronin, Vice President, Education Research of NWEA, and Yeow Meng Thum, Sr. Research Fellow of NWEA and an author of NWEA’s 2015 Norm study, . . . told the RoundTable, ‘You are likely familiar with the recent study of Chicago student achievement by Stanford professor Sean Reardon, which concluded that Chicago schools have shown a pattern of improvement based on ISAT and NAEP results from 2008 through the 2015 school year. We believe that this corroboration of academic improvement stems from important methodological communality shared by our work. Analysis of growth is actually based on a vertically equated national grade-equivalent metric, not on percentiles per se. In the case of MAP Growth, MAP percentiles are deployed as a normative reporting metric, and not as inputs to the analysis.” [emphasis added]

Evanston *RoundTable*, August 8, 2018: Reported by Larry Gavin, Editor

A fundamental difference in how results are reported by NWEA and the Stanford Education Data Archive is that NWEA reports average growth in *separate yearly increments* while Stanford reports *net growth across five years of schooling* between the end of grade three and end of grade eight. This difference allows Stanford to report growth information in ways that: a) control for differences in baseline achievement status at the end of third grade; and, b) focus directly on the *cumulative impact* on learning that students experience between third and eighth grade. Another important difference is that Stanford uses grade-equivalents to represent scale score shifts over time in norm-referenced ways.

Shortly after the publication of Stanford’s research in 2017, the Center for Urban Education Leadership began to conduct rough, Stanford-style analyses of cohort growth over time using average MAP scale scores for various Chicago sub-groups that completed MAP testing between 2013 and 2018. To control for potential distortions in 2015 NWEA norms, these analyses applied 2011 NWEA percentile ranks to all the data studied.

Figure 4.3 reports aggregate growth data across 130, PK-8 Chicago elementary schools where student enrollments were at least 90% Black and where at least 90% of students came from low-income households. To minimize enrollment distortions, scoring for each of the 130 schools was weighted based on number of students tested. Vertical axes show MAP scale scores. Horizontal axes identify the testing years reported for three actual cohorts and one projected cohort.

In this and other sub-groups studied, two broad trends emerged between earlier and later cohorts in both reading and math. These two trends are illustrated in Figure 4.3 below.

1. Achievement growth at all grade levels rose in earlier years and flattened in later years
2. Cohort growth beyond expected gains consistently declined between earlier and later cohorts because of diminishing value added as cohorts moved from lower to higher grade levels. For example:
   o Earlier reading growth beyond expected gains between fifth grade—2013 and eighth grade—2016 (blue box) was roughly 1.5 grade equivalents; this estimate is based on a three-year shift in average scale scores from 202.8 (25th percentile at end of fifth grade) to 222.1 (49th percentile at end of grade eight)
Later reading growth beyond expected gains between fifth grade—2015 and eighth grade—2018 (green box) was roughly 1.0 grade equivalents; this estimate is based on a three-year shift in average scale scores from 206.8 (34th percentile) at the end of fifth grade to 221.9 (48th percentile) as the end of eighth grade.

**Figure 4.3**
Actual and Projected Cohort Growth in 130, Predominately Black/Predominately Low-Income Schools Based on 2011 Percentile Ranks of Average MAP Scale Scores

DATA SOURCE: Chicago Public Schools Assessment Files [https://www.cps.edu/academics/student-assessments/#Assessments38](https://www.cps.edu/academics/student-assessments/#Assessments38)

As conversations continued with other members of the school research community and senior district leaders, UIC conducted additional studies that drew directly on data from UIC’s statewide data archive. The purpose of these studies was to interrogate the trends illustrated in Figure 4.3 using z-score analyses that compared cohort achievement and growth over time in Chicago with expected gains based on annual statewide scoring norms for all students tested statewide.

An important characteristic of z-score analysis is that it produces greater normative precision than many other normative comparisons. It does that by:

- accounting for shifts in scoring across entire scoring distributions
- accounting for the statistical properties of all scoring distributions by measuring differences in standard deviations from the mean
The broad findings of UIC’s z-score analysis are shown in Figure 4.4. The blue bars in Figure 4.4 represent the very substantial cohort growth beyond expected gains that occurred in Chicago for eight, consecutive, eighth grade graduating classes ending in the Class of 2014. Mirroring prior reporting by Stanford and UIC, almost all growth beyond expected gains during this period was around 0.3 standard deviations or higher. The only exception to this was reading growth in the Classes of 2013 and 2014. In these two cohorts, growth dropped to around 0.2 standard deviations beyond expected gains.

But subsequent to the Class of 2014, longstanding patterns of growth beyond expected gains fell precipitously in both reading and math. And by the time the Class of 2019 graduated, growth beyond expected gains had slipped into negative territory. This pattern mirrored the general findings of earlier MAP-based analyses shown in Figure 4.3, and was consistent with earlier reports of inflated NWEA achievement norms shown in Figures 2.1, 2.2 and 2.3.

**Figure 4.4**
Shifts in Cohort Value Added over Time in Chicago: End of Grade 3 to End of Grade 8
Conclusion

Everybody wants things to do double duty. But what we find in assessment is that, generally, when we try to make things do double duty, they don’t do either of them very well.”

Dylan Wiliam, Learning Sciences International Conference on Formative Assessment (April 2022)

Since the early days of NCLB, the number and percentage of American school districts that have tried to embrace NCME’s 2023 theme of “leveraging measurement for better decisions” has increased dramatically.26 The Northwest Evaluation Association has become the nation’s leading vendor in this area because their consumers have come to believe that NWEA provides them with timely, trustworthy, information that is not readily available from other sources.

For the most part, the language and inner workings of large-scale, standardized assessment are inaccessible to otherwise well-informed people who do not have advanced technical preparation and experience in the field. For this reason, the psychometric industry, like all other professions, is obliged to “first do no harm,” and to honor its fiduciary responsibility to behave like “a person to whom property or power is entrusted for the benefit of another.”27 Many prior studies of small-grain reportage from large-scale interims like MAP,28 and the four cases of large-grain reportage described here, suggest that NWEA . . . now in its 50th year of doing business . . . is falling well short of this standard.

Consistent with this year’s AERA theme of “Interrogating Consequential Educational Research in Pursuit of Truth,” it is also important to take NWEA to task for knowingly misrepresenting interim assessment as a necessary and widely-acknowledged ingredient of comprehensive assessment systems29. For example, in MAP Growth Theory of Action (2022), J.P. Meyer (NWEA’s Senior Director of Psychometric Solutions) and Michael Dahlin (NWEA’s Manager of Research Reporting and Data Analytics) write,

“A comprehensive assessment involves multiple forms of assessment that each emphasize a different purpose and use (Perie, et. al. 2009). Summative, interim, and formative assessments complement each other and provide the appropriate data for decision-making at every level of an education agency. MAP Growth is a component of a comprehensive assessment system.”

MAP Growth Theory of Action (2022)

But Scott Marion, President and Executive Director of the National Center for Improvement of Educational Assessment (NCIEA), was a co-author of the 2009 study that Meyer and Dahlin reference. Three full years before they wrote MAP Growth Theory of Action, Marion wrote,

“Unfortunately, the title of our well-cited paper, Moving towards a comprehensive assessment system: A framework for considering interim assessments, suggested that interim assessments were a required component of comprehensive assessment systems. That implication was not our intent and that’s not what the text of the paper says.”

Marion then raised several reservations about interims that included the following:

“Both vertical and horizontal coherence is necessary for assessment systems to be balanced, but both are difficult to achieve when commercial interim assessments are included as part of the mix of district assessments. If such assessments are based on an explicit model of
learning (and it is not clear that most are), it is incredibly unlikely for the same model of learning to be found in each district where the assessments are being used. Therefore, the interim assessment results could be sending mixed or even incorrect signals about what students have learned relative to how they were expected to develop in a content domain.

Do Interim Assessments Have a Role in Balanced Systems of Assessment (2019)

Four months after Meyer and Dahlin released MAP Growth Theory of Action, Juan D’Brot, a Senior Associate at the NCIEA, wrote an article titled, “Is There a Recipe for Balanced Assessment Systems?.” In it, he wrote,

“I’m not saying that it’s impossible for interim assessments to be useful in informing instruction. I’m just saying we really haven’t seen it happen systematically yet. If you know of any documented examples of interim assessments being used effectively to inform instruction, please reach out and share them with me . . .

“If you were to ask me what the most promising aspect of interim assessments might be, I would advocate for the role of corroboration. The grain size of interim assessment results is too large to be instructionally useful, but it can help confirm or disconfirm whether instructional efforts were on target over a particular set of time.

“However, corroboration will only work if the interim assessment is aligned to the state’s standards, aligned to the state’s performance expectations (as reflected in the standards and the large-scale assessment), and when results can be interpreted by educators and students in light of what curriculum has or has not been covered. Corroboration also requires a very savvy user base. And that requires a significant investment in assessment and data literacy learning opportunities, which are dependent on good assessment literacy programs, resources, and structures to deliver them.

“Interim assessments might be able to support more effective teaching and learning if used carefully and appropriately, which could justify their investment as part of an assessment system. Ultimately, however, interim assessments are NOT required to balance an assessment system.” [emphasis added]

Is There a Recipe for Balanced Assessment Systems? (2022)

Or maybe, as Scott Marion suggested in an April 2021 blog post about interim and through-year assessment,

“It might just be a pile of bricks!”

It Might Just Be a Pile of Bricks (2021)
Endnotes

1. A substantial amount of the work represented in this study was done in close collaboration with Larry Gavin, a long-time editor of the Evanston RoundTable. Without Larry’s continuing support and counsel, it would not have been possible to complete this study in its current form.


4. See, for example, Zavitkovsky, P. (2009) “Something’s Wrong with Illinois Test Results.” Center for Urban Education Leadership, University of Illinois at Chicago https://slidelegend.com/somethings-wrong-with-illinois-test-results-center-for-urban-59b7dee81723ddf2725f1a43.html

5. See hour 2:52:00 of a presentation by State Superintendent Carmen Ayala to the Illinois State Board of Education at their regular monthly meeting on May 18, 2021, available at https://register.gotowebinar.com/recording/724526900778132481

6. For example, online marketing materials for NWEA’s MAP Growth system say, “Easy-to-use, standards-aligned reports put the information teachers need at their fingertips. Reliable insights make it simple for teachers to find common areas of need among their students, identify students who could benefit from intervention, and determine which instructional strategies are generating the most academic growth. Higher-level reports provide administrators with the context to drive improvement across entire schools and educational systems.” https://www.nwea.org/map-growth/


16. Illinois’ shift from ISAT to PARCC testing in 2015 is an interesting example of how very different standardized tests typically yield similar normative results when they assess the same underlying content and the same test populations. There was a widespread perception in 2015 that the ISAT was a relatively easy test, and that PARCC was tougher and more rigorous. And yet, students that scored in the 50th percentile on 2014 ISAT exams typically scored pretty close to the 50th percentile the following year on 2015 PARCC exams.

One likely contributor to the difference in perceptions about the two tests is that the ISAT had a larger number of less challenging items because ISAT cut scores were set between the 20th and 30th percentile of the statewide scoring distribution. This led test publishers to increase the portion of “easier” items in the test battery to increase the reliability of scoring around cut score boundaries. PARCC publishers did the same, but the cut scores they were working with were set at between the 60th and 70th percentile.


17. For example, in the early years of NCLB, Chicago Public Schools transitioned from using the Iowa Test of Basic Skills to the Illinois Standards Achievement Test (ISAT). The “look and feel” of the ISAT was that it was more rigorous test than the ITBS because it had more constructed response items, contained longer, more detailed passage and prompts, etc. But in a comparative analysis of the two tests in 2003, John Easton and his co-authors at the Consortium on Chicago School Research concluded, “In spite of large content and format differences, the ITBS and ISAT behave similarly among CPS students. Their scores are highly correlated and their trends over time are mostly parallel.” How Do They Compare? ITBS and ISAT Reading and Mathematics in the Chicago Public Schools, 1999 to 2002 p. 19.


19. Currently, the “MAP Growth Tour” tab of NWEA’s webpage continues to include a section called “Norming and RIT Scores.” The opening paragraph of this section reads as follows:
“The most accurate norms

“Today’s kids are vastly different than those in school ten years ago, so NWEA™ conducts norming studies frequently—three to five years—ensuring that the comparisons reflect current standards and demographics. And we continually refine our best-in-class research model to make NWEA norms more accurate than any other—and always improving” [emphasis added]

https://www.nwea.org/normative-data-rit-scores/

20. See, for example, Rupp, A. “Understanding Root Causes of Misalignment Between Results from Different Assessments” National Center for Improvement of Educational Assessment, November 2, 2022 https://www.nciea.org/blog/understanding-root-causes-of-misalignment-between-results-from-different-assessments/


22. For more information about the Evanston RoundTable: https://evanstonroundtable.com/about-us/


24. For more information about the Stanford Education Data Archive see https://edopportunity.org/


27. https://www.dictionary.com/browse/fiduciary


29. For example, just a little more than three months after COVID-19 shut down American schools, Steven Underwood, the newly-appointed Director of State Professional Learning and Consulting Services for NWEA, posted a piece on NWEA’s website called, “How to Build a Balanced Assessment System.” In this piece, Underwood repeatedly noted how interim systems like MAP Growth can support deep formative assessment, and help monitor the impact of that assessment on broader curricular outcomes in deeply challenging times. https://www.nwea.org/blog/2020/how-to-build-a-balanced-assessment-system/
Appendix A
Report to ETHS School Board on Discrepancies between Proficiency in Reading as Measured by NWEA MAP and Renaissance STAR 360 Testing

To: Eric Witherspoon, Superintendent
CC: Marcus Campbell, Assistant Superintendent/Principal
     Pete Bavis, Assistant Superintendent for Curriculum and Instruction
From: Scott Bramley, Associate Principal for Instruction and Literacy
     Kiwana Brown, Reading Specialist
Date: January 10, 2019
Re: Update on Reading Program

Introduction

In 2014-2015, District 202 and District 65 adopted the joint literacy goal, which stated that the districts would “ensure that all students are proficient readers and college and career ready by the time they reach 12th grade.” At the time of the goal’s adoption, the Boards referenced that the goal marked the beginning of a K-12 cycle, which means the students who were in kindergarten in 2014-2015 are presently in 4th grade.

The focus for this report is the “proficient readers” part of the goal. In previous meetings, Board members have inquired as to the status of our work toward this goal since its adoption. In addition to the overall status of reading in the district, the report contains an update on the tier 2 reading intervention program. These updates include:

- Incoming freshman trend data
- Reading growth trend data
- Tier 2 reading intervention program enrollment trend data and program revisions
- Preview of current programs supporting reading

Incoming Freshman Data

The grade equivalent (GE) as measured by the STAR Reading assessment is used to measure and report student reading proficiency in District 202. According to STAR, a GE indicates the grade placement of students for whom a particular score is typical. If a student receives a GE of 10.7, this means
that the student scored as well on STAR Reading as did the typical student in the seventh month of grade
10. District 65 shares MAP data with ETHS that is used for placement purposes only. Figures 1-4 show
the percent of incoming freshman demonstrating reading proficiency for each of the past five years as
measured by the STAR GE and the MAP.

![Incoming Grade 9 Reading Proficiency
STAR and MAP Assessments Fall 2014-2018
All Students](image)

**Figure 1.** 2014-2018 Incoming Grade 9 Reading Proficiency (STAR vs. MAP)- All Students

Figure 1 presents the overall GE for each incoming freshman class for the past five years. The trend data
show that approximately 59% of incoming freshmen score at or above the benchmark grade equivalent of
8 years, 3 months. This also means that approximately 41% of the incoming freshmen during this same
due five-year period enter ETHS not reading at grade level. “Since older students are so far behind, the
amount of intervention needed so that they perform ‘on par’ with peers and/or meet grade-level
benchmarks (e.g., state performance standards) will be more extensive, given both the complexity of the
information that older students are expected to know and the longer period of time that some of these
students have struggled with reading” (Vaughn et al., 2008).
Figure 2. 2014-2018 Incoming Grade 9 Reading Proficiency (STAR vs. MAP)- Black/African American Students

Figure 3. 2014-2018 Incoming Grade 9 Reading Proficiency (STAR vs. MAP)- Hispanic/Latinx Students
Figure 4. 2014-2018 Incoming Grade 9 Reading Proficiency (STAR vs. MAP) - White Students

As Figures 2, 3, and 4 show, the GE for each subgroup remains consistent as well, except for black/African American students who show an increase of 16% (22-38%) between the fall 2017 and fall 2018 classes (Figure 2). Despite this increase, there is much work to continue to do in order to support the reading needs of students who identify as black/African American or Latinx.

During the most recent joint meeting of the District 65 and District 202 Boards of Education, members asked if the MAP scores reported by District 65 were predictive. Since the two assessments use different metrics and analytics, there are stark differences in the STAR and MAP scores (a mean difference of 20% for the five years), so the MAP scores are not predictive. Although this analysis shows GE for STAR and percentile rank for MAP, these differences also appear when comparing STAR percentile rank to MAP percentile rank.

ETHS only uses MAP data for placement purposes for incoming freshman students. Starting with the 2014-2015 school year, the district no longer administers the MAP assessment to students enrolled in
our reading intervention programs. Although MAP assesses word recognition, structure & vocabulary, informational texts, craft structure & evaluation, grammar & usage, and writing conventions, the MAP assessment is not normed beyond 8th grade. At this time, we shifted to using the STAR Reading Assessment. STAR assesses the following reading domains: literary text analysis, word skills and knowledge, argument analysis & text evaluation, comprehension strategies & meaning construction, and author's craft. Within STAR, teachers have access to resources that support individualized instruction to target students' reading strengths and areas of growth, including instructional planning reports.
Appendix B


How Many Students Are Not Proficient Readers When They Leave School District 65: 16% or 40%?

1/23/2019

In Jan. 16, a Joint Committee of the District 65 and 202 School Boards discussed the Districts’ Joint Literacy Goal which is to “ensure that all students are proficient readers and college and career ready by the time they reach 12th grade.”

Two days earlier, the District 202 School Board discussed a report that contained charts showing the percent of students who were proficient in reading when they entered Evanston Township High School using 1) the STAR test, which is given by ETHS; and 2) the MAP test, which is given by District 65.

The charts show the tests yield starkly different results.

For example, one chart shows that in the fall of 2018, 60% of all students who entered ETHS as freshmen were proficient in reading using the STAR test, while 84% were proficient in reading using MAP.

Another chart shows that in the fall of 2018, only 38% of Black students who entered ETHS as freshman were proficient in reading using the STAR test, while 66% were proficient in reading on the MAP test. See chart below illustrates the differences.
The report was prepared by Scott Bramley, Associate Principal for Instruction and Literacy, and Kiwana Brown, Reading Specialist, both at ETHS.

Members of the Joint Committee were frustrated that the results were so different, and that the Districts had not come up with a common assessment or a way to statistically align the STAR and the MAP tests. Several people questioned whether the tests and/or definitions of proficiency were aligned. See sidebar below.

The Discussion

Mark Metz, a member of the District 202 School Board, laid out his concerns. He said the School Boards adopted the Joint Literacy Goal in January 2014, and it was “frustrating” that the Districts have not yet developed metrics that were aligned to assess their joint progress toward meeting the goal.

He added that another concern was that the MAP test was not identifying many students who were not proficient in reading and who needed interventions. “We are under-identifying students who are not proficient in reading every year as they enter high school because all the data we [ETHS] have at that point is MAP. The fact of the matter is we’re getting MAP data only, so that’s how we’re placing them and then we’re discovering … we’re finding after the fact we’re under-identifying kids who need help – in big numbers. Then we’re having to go back and retrofit that and get them into special programs. We’re losing time.”

He said the MAP test is identifying about 120 students each year as lacking proficiency in reading, while STAR is identifying about 300 students who are not proficient in reading.

Mr. Metz said it is critical that ETHS be able to identify students who are not reading at grade level. “I’m looking for a sense of urgency to get this fixed … and resolve these data problems … to advance the goal of making sure each student by grade 12 is proficient.”

Dr. Bramley said, in his experience, the STAR results are more accurate than MAP in terms of making placement decisions.

Marcus Campbell, Principal of ETHS, said teachers “tell us consistently that STAR is far more reliable because it’s what they see on the ground as they teach every day.”

Gretchen Livingston, a member of the District 202 Board, said, “There needs to be a sense of urgency around this. … We need to have a better plan and it needs to happen now.” She suggested the Districts could have an interim plan and then develop a long-term plan later.

Pete Bavis, Assistant Superintendent of District 202, said there are three ways to address this:

1. Districts 65 and 202 could adopt a similar assessment system,
2. ETHS is using the SAT suite of assessments for grades 9-12, and District 65 could use the SAT suite of assessments in grades 7 or 8 and then backmap to earlier grades, and

3. The Districts could do a statistical alignment between STAR and MAP, which he said may or may not work.

Dr. Bavis said the first two options are pretty simple solutions. The third requires a lot of time.

District 65 Superintendent Paul Goren said that District 65 is currently giving many tests, and if a new assessment was added, the District would need to eliminate one that is currently being used. He said, “Part of our work over the next year plus is to do a deeper dive into what our assessment system looks like and how we might change it.”

He said based on his work at the Chicago Public Schools, he thought MAP seemed to be a better assessment to help teachers frame instruction, and added that it would be a “medium to a major lift” to add STAR because it was a computer adaptive test and “all of our kids don’t have immediate access to computers.”

Stacy Beardsley, Assistant Superintendent of Curriculum and Instruction for District 65, said a study done by Renaissance, the owner of the STAR test, concluded there was a high correlation between the STAR test and MAP in both reading and math. Assuming the tests are correlated, the two research departments could discuss if there are cut scores on the tests that would provide a consistent benchmark for measuring proficiency. She said, “This is probably the direction that I would want to step first before trying to decide if we do more assessment.”

Dr. Beardsley also said that the District has 15 years of MAP data, and cautioned that when assessment systems are considered, the assessment system needs to provide tools that help teachers make really strong instructional decisions that will benefit students.

“If it doesn’t inform teaching and learning well, then we will have got something that the Board can use, but not our teachers and students,” said Dr. Beardsley.

Jonathan Baum, a member of the District 202 School Board said, “The vast majority of school districts in the nation are K-12. How do they do this?” He asked administrators to research what K-12 districts are doing.

District 202 Board President Pat Savage-Williams said, “We need to move as quickly as possible,” adding that coming up with Option 3 may take another year.

District 65 Board Vice President Anya Tanyavutti said in addition to discussing possible ways to align proficiency levels, the administrators should determine “what are the implications of adopting another assessment tool, and not just what they perceive, but run that by stakeholders, including students.” She said the Districts should move with urgency, but thoughtfully.
Dr. Bramley said, “What we really need to consider is do we want to have a measure of reading proficiency and growth for grades 3-12. And if we do, we need to adopt a system that will allow us to do that. If we’re going to use different measures, I think we’re going to continue having these frustrations.”

Dr. Goren said before the Joint Board meeting on Feb. 25, the Districts will get their research teams together and continue the conversation, and attempt to come up with a short-term solution and a long-term solution.

The STAR and MAP Tests - Definition of Proficiency Levels

At the Joint Committee meeting on Jan. 16, Stacy Beardsley, Assistant Superintendent of District 65, said she was not in a position to comment on the different results shown for STAR and the Measures of Academic Progress (MAP) test because she lacked information concerning the definitions of proficiency used for each test and also did not know what MAP test data was used. She said ETHS uses the highest score an eighth-grader earns on the Fall, Winter or Spring MAP tests in making a placement decision. She said she was not sure if the data presented in the ETHS report used the highest score of the three MAP tests or just the score on the Spring MAP test.

She added there was a significant difference in results between the Fall and Spring MAP tests, and “We see a drop anywhere from 15% to 30% from the Fall scores over meeting college readiness benchmarks.” She said this could be due a lack of motivation of eighth-graders.

But the drop could also be due to a drop in academic achievement.

Scott Bramley, Associate Principal for Instruction an Literacy at ETHS, told the RoundTable:

- In reporting the percentage of ETHS freshmen who are proficient on the STAR test, ETHS used a grade equivalent score of 8.3 as the benchmark to measure proficiency. If a ninth-grader received a grade equivalent score of 8.3, that means that the student scored as well on the STAR Reading test as a “typical” student in the third month of eighth grade.

- In reporting the percentage of students who were proficient on the MAP test, ETHS used the highest score a student had received as an eighth-grader on the Fall, Winter or Spring MAP test. If that score was above the 50th percentile for the years 2014, 2015 and 2016, the student was regarded as proficient. If that score was 220 or above in the years 2017 and 2018, the student was regarded as proficient. An eighth-grade score of 220 in reading on MAP equates to the 50th percentile on the Spring MAP test.

A score at the 50th percentile – which is the median – could be considered a typical score.
While the MAP test is generating significantly higher results than the STAR test, the RoundTable recently reported that students in the Chicago Public Schools were showing significantly higher norm-based results on the MAP test than on several other tests.

For example, 73.5% of CPS eighth-graders ostensibly scored at or above the national average score in reading on the Spring 2017 MAP test. In contrast, only 42% of CPS eighth-graders scored at or above the national average score in reading on the 2017 National Assessment of Education Progress test. The analysis was conducted with Paul Zavitkovsky, a researcher and leadership coach at the Center for Urban Education Leadership Program at the University of Illinois at Chicago.