Center for Education Policy Research
HARVARD UNIVERSITY

STRATEGIC DATA PROJECT

SDP FELLOWSHIP
CAPSTONE REPORT 2016

Think Local, Act National: Local Analysis of the Nationally-Normed MAP

Eva Yiu, Howard County Public Schools
Ebony Langford-Brown, Howard County Public Schools
KC Holder, Northern Michigan University Charter Schools
Tom Hay, Uplift Education
Barb Soisson, West Linn-Wilsonville School District

Executive Summary

Howard County Public School System (HCPSS), Northern Michigan University Charter Schools Office (NMUSCO), Uplift Education (Uplift), and West Linn-Wilsonville School District (WLWSD) all use the Northwest Evaluation Association (NWEA) Measures of Academic Project (MAP) test as a source of formative data. Given the diversity of the districts and schools that use NWEA MAP, these agencies wanted to know: How can nationally-normed NWEA MAP data meet local prediction and benchmarking needs?

Each site approached the question with a different angle. In order to establish benchmarks, the Strategic Data Project (SDP) Fellows conducted the following analytics in their respective organizations:

- NMUSCO compared student demographics and baseline test scores to those of the NWEA-created virtual comparison group.
- Uplift used a linear regression model with MAP scores and Uplift assessment data to predict Texas STAAR performance.
- WLWSD regressed MAP scores on Smarter Balanced math scores and determined their correlation.
- HCPSS regressed PARCC scale scores on MAP scores to determine how well MAP performance benchmarks align with the PARCC indicators of college-readiness.

Ultimately, the SDP Fellows found that NWEA MAP correlates with Texas STAAR performance, Smarter Balanced with math performance, Fountas & Pinnell with reading levels, and PARCC with reading and math performance.

Strategic Data Project Fellowship Capstone Reports

Strategic Data Project (SDP) Fellows author capstone reports to reflect the work that they led in their education agencies during the two-year program. The reports demonstrate both the impact fellows make and the role of the SDP network in supporting their growth as data strategists. Additionally, they provide recommendations to their host agency and may serve as guides to other agencies, future fellows, and researchers seeking to do similar work. *The views or opinions expressed in this report are those of the authors and do not necessarily reflect the views or position of the Center for Education Policy Research at Harvard University.*

Introduction

School districts across the United States use student assessment outcome data as a significant part of evaluating effectiveness at the program, school, and district levels. Agencies have a choice in how they leverage interim assessment data to evaluate and improve programs, with different agencies often using the same data in specific ways to meet the needs of their districts.

The four agencies described in this report use the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) test as a primary source of formative data. As an interim measure and part of a balanced assessment system, MAP provides on-track information about students' progress that agencies can use to adjust instruction and examine the effectiveness of programs. The agencies described in the sections that follow are the Howard County Public School System (HCPSS), Northern Michigan University Charter Schools Office (NMUCSO), Uplift Education (Uplift), and West Linn-Wilsonville School District (WLWSD). Together, they represent a variety of approaches to benchmarking expected performance on the MAP assessments by leveraging national and local norms to align multiple indicators of performance. Each uses these benchmarks to help inform stakeholders about progress and trends that give way to action planning and continuous improvement.

The four programs also reflect the work of agencies in the transition phase that is underway in the national accountability system, the move from No Child Left Behind (NCLB) to the Every Student Succeeds Act (ESSA).[1] This has implications for how agencies develop

[1] No Child Left Behind was signed into law in January 2002 as reauthorization of the Elementary and Secondary Education Act. The Every Student Succeeds Act was signed into law in December 2015 as reauthorization of the Elementary and Secondary Education Act. (See U.S. Department of Education, 2016a, 2016b.)

assessment systems and use the information obtained from them. The intent is to maintain an accountability system focused on using achievement outcomes to eliminate opportunity gaps. With the enactment of ESSA, assessment expectations are standardized to ensure that all students are making progress, and there is a more focused emphasis on preparing each and every child for college and careers. This type of national legislation should be implemented so that accountability for learning leads to meaningful work in schools and districts with high expectations and strategies to support them. This can be accomplished by developing a balanced assessment system, one that includes formative, interim, and summative assessments.

The four agencies included here have all emphasized the use of interim assessments because they recognized the importance of using more than summative measures to determine program effectiveness and student progress. Likewise, using more local and internal measures formatively and moving away from reliance on external summative assessments is an element of the shift from NCLB to ESSA. The agencies in this report believe that a focus on the interplay between formative and interim assessments, and between interim and summative assessments, is timely in terms of developing strategies to improve student achievement. The shifts that Darling-Hammond, Wilhoit, and Pittenger (2014) recommended (shown below in Figure 1) interest the four agencies and served as inspiration for their work with a balanced assessment system.

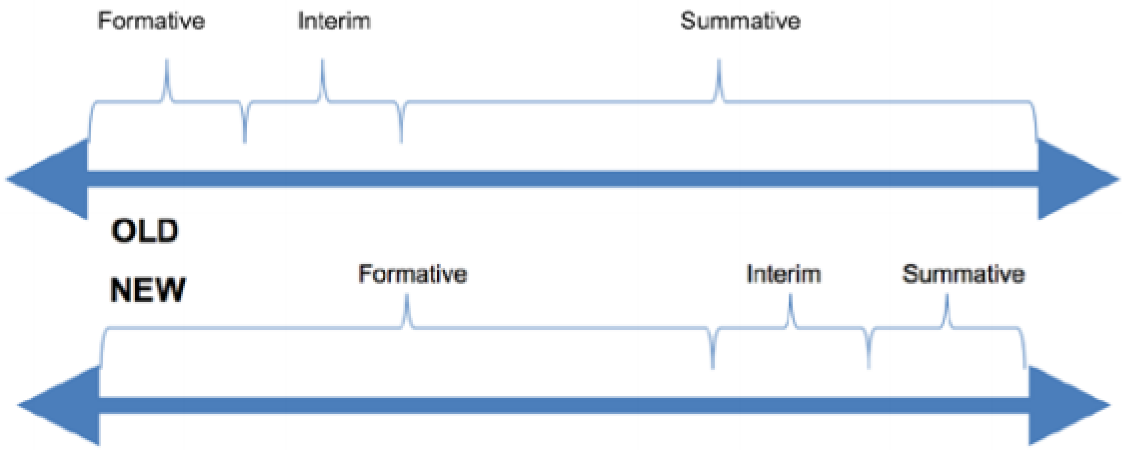


Figure 1. Relative emphasis of assessments within a balanced assessment framework. From Darling-Hammond et al. (2014).

Given what the NWEA publishes as nationally-normed MAP achievement and growth percentiles, we wondered: To what extent should local norms be considered in guiding schools to set benchmarks of expected performance? To answer this question, we first present a brief literature review on the relationship between standards and assessments, with attention to recommended practices on assessment data use and supporting data use. Then, using case studies, we illustrate how formative assessment data contributed to setting appropriate benchmarks and targets, as well as to monitoring progress toward these targets. We emphasize the importance of supporting data coaching to help stakeholders translate data into practices that have implications for teaching and learning. We end with key takeaways that have supported the success of this work, highlighting how appropriately visualized data has helped schools to move beyond the numbers to taking actions for continuous improvement.

Literature Review

Move to National Standards and Assessments

The plans enacted as a response to *A Nation At Risk* (National Commission on Excellence in Education, 1983) and reported in *Goals 2000* (National Education Goals Panel, 1994) started to describe how American students could gain “world class” skills to address concerns that they lagged behind their international peers and to reflect the agendas of educators, policymakers, and business leaders. To develop recommendations for national standards and testing, Congress established the National Council on Education Standards and Testing (NCEST). A NCEST (1992) report stated that the intent of national standards and testing was to both raise the ceiling for students who currently performed above average and to raise the floor for those who experienced the least success in school.

From the outset, gaining consensus about the purpose and use of national standards and assessment was difficult (Massell, 1994; Smith, Fuhrman, & O’Day, 1994) and agreeing on common state measures and ranking systems was problematic (Olson, 1998; Wixson, Dutro, & Athan, 2003). States set their own cut points for proficiency under NCLB, and while ESSA mandated annual standards-based assessments in reading and language arts, math, and science, individual states decide on the assessments they will use and on what constitutes proficiency. The differences among states’ definitions of proficiency and the increased sense of urgency around preparing students so they are successful beyond K–12 classrooms in each state has prompted local education agencies to recognize the importance of developing assessment systems that include and align formative, interim, and summative assessments (Black & Wiliam, 1998; Huebner, 2009; Stiggins, 2005).

The implementation of the Common Core State Standards (CCSS) in 42 states has been associated with small learning gains from 2009 to 2013, as measured by the National Assessment of Educational Progress. Moreover, the shift that states made from previous content-based assessments to those that assess the CCSS has pressed teachers to make major changes to instruction in order to prepare students to demonstrate the skills associated with college and career readiness (Kane, Owens, Marinell, Thal, & Staiger, 2016; Loveless, 2015).

A Balanced Assessment System

Within a comprehensive and balanced assessment system, interim assessments occupy a pivotal role because they allow students to show progress towards benchmarks that align with proficiency on summative assessments. They can also serve as the link between classroom-level measures of learning that build teacher, school, and district capacity for creating and using formative assessments effectively (Black & Wiliam, 1998; Perie, Marion, Gong, & Wurtzel, 2007). NWEA, an interim assessment provider that provides testing materials for many states, found in a 2007 study that there was significant variation in states' benchmarks for proficiency—at the lower end of the cutpoint range, 94% of students were proficient and at the upper part of the range, 23% were proficient (Cronin, Dahlin, Adkins, & Kingsbury, 2007). Because state assessments influence the instructional priorities and practices of local education agencies, both of which take time and significant effective professional learning to change (Darling-Hammond et al., 2014), it is likely that the benchmarks on a national interim assessment such as MAP will need to be considered in the context of the relative rigor of state standards that have been in place as the CCSS or their similar counterparts are fully adopted.

Among the wide-ranging factors present in public and charter schools, the NWEA MAP norms report notes that the factors on which NWEA schools differ from each other are strongly related to student achievement and school effectiveness (Thum & Hauser, 2015). To address this, in 2011, NWEA introduced the School Challenge Index (SCI), which includes variables that describe sociodemographic, organizational, and educational policy programming factors. The SCI has mostly given NWEA norms information about the differences between states, but because of the additional differences among their clients within states, it is suggested that district- or school-level weights also be used.

Purpose and Use of Interim Assessments

Interim assessments, including MAP, are used to inform instructional practice, evaluate programs and practices, and predict summative outcomes (Perie et al., 2007). Because they are given at intervals during a school year, they can become a source for ongoing analysis of student progress. Recommended practices for using interim assessments to improve achievement include: triangulating data by looking at the results alongside formative assessments; engaging teachers in developing hypotheses for improving the achievement that they see in the data and then modifying instruction to test them; and providing professional development that goes beyond use of the actual assessment and focuses on analysis and use of the results (Crane, 2010; Hamilton et al., 2009).

NWEA's MAP test is scored on a Rausch Unit (RIT) scale. The RIT scale is a stable equal-interval vertical scale that spans grades.[2] Thus, the difference between scores is the same regardless of whether a student is at the top, bottom, or middle of the RIT scale, and allows comparisons of students independent of grade level ("RIT [Rausch UnIT] scale," 2016). NWEA conducts a norms study every 3 to 4 years to account for improvement that occurs within assessed areas, for population changes such as increases in English language learners, and, as

[2] See https://www.nwea.org/assessments/map/scale-and-norms/

clients change, to reflect the characteristics of those who are using the assessments. MAP assessments provide status and growth norms.

The NWEA RIT Scale Norms Study includes student and school status, as well as growth norms in reading, mathematics, language usage, and general science based on 72,000 to 153,000 K–11 student assessment records from about 1,000 schools (Thum & Hauser, 2015). The norms are used to find a student's achievement status, given as a percentile rank. The MAP Conditional Growth Index (CGI) is a metric that shows how a student's growth compares to students across the nation. It compares a student's actual growth between assessment intervals to projected growth. Projected growth is derived from the mean growth of students with the same assessment scores across the nation. The CGI is expressed in standard deviation units, meaning that a CGI of zero shows that the student's growth matched the projection (NWEA, 2017).

Importance of Assessment Literacy

Professional learning that promotes assessment literacy and fosters a culture of data inquiry is an essential to the effective use of assessments within local education agencies (Hamilton et al., 2009). Leadership that introduces and develops regular practices for interpreting results, understanding them in the context of teaching and learning, and using them to guide instructional practice, evaluate programs, or predict outcomes determines the extent to which assessment information can be used to improve student achievement (Boudett, City, & Murnane, 2013). Similarly, professional development that provides teachers with information about how students learn and engages them in ongoing practices that let them see the effectiveness of the use of learned strategies has an impact on achievement when it is integrated with content-based learning (Blank & de las Alas, 2009). There is a reciprocal relationship between increasing assessment literacy as an integral part of professional learning and emphasizing the use of assessment information to determine areas of focus for professional learning that can improve student achievement (Blank & de las Alas, 2009; Boudett et al., 2013; Mizell, 2010). Providing principals and teachers with accurate and contextualized evidence of students' learning progress is important, but ensuring practices to make meaning from that information and use it for instructional decision making is essential for ongoing improvement.

Case Study 1:

Northern Michigan University Charter Schools Office

Northern Michigan University Charter Schools Office is committed to a course of action that enhances educational reform through authorization of public school academies (charter schools) founded on the principles consistent with Northern Michigan University (NMU). NMUCSO charters nine schools, only one of which is geographically close to the NMUCSO office (approximately three miles); the other eight are 85 to 458 miles away. Serving just over 4,000 students, the schools represent a wide range of diversity in terms of ethnicity, special education status (from 7.6%–18.1% of students, with an average of 12.7%) and socioeconomics (between 62.6% and 85.1% of students qualify for free or reduced price lunch, with an average of 73.2%). NMUCSO has one full-time director, one part-time professional staff for curriculum and assessment, two part-time school/authorizer liaisons, and two part-time office staff. The Strategic Data Project (SDP) Fellow at NMUCSO serves as a consultant, monitoring and evaluating the public school academies, strategic planning, and the assessment and analytics team.

In Michigan, charter school authorizers establish a contract with the charter school that is time-bound—it typically ranges from 3 to 5 years, but may range from 1 to 7 years, depending upon prior performance. Near the completion of the contract term, authorizers must determine whether the school should be offered a renewal of their contract, and why or why not. Multiple factors and types of evidence are evaluated when determining whether a school should be reauthorized.

Although the principles and standards of the National Association of Charter School Authorizers (NACSA, 2015) outlines five broad categories of charter school effectiveness, they can be summarized as academic, financial, and organizational. Within the broad category of academic performance, one approach authorizers commonly use is to compare the achievement results of students to another group of students. For example, some authorizers compare geographic composite feeder schools, others compare students with a host district or the state averages, and others compare students to national norms. The determination of a comparison group may or may not be appropriate depending upon shifting regional demographics or the mission of a specific school.

As part of monitoring academic performance, NMUCSO requires all nine schools to utilize the MAP. Given the ethnic and economic diversity of the nine schools, NMUCSO wanted at least one indicator comparing charter school students to other students most like them. More specifically, the goal was to find the "best indicator feasibly available" that describes how NMUCSO students compare to similar students. The MAP offers an optional product called the Virtual Comparison Group (VCG), which claims to be an "apples-to-apples" comparison that defines comparison students using the following criteria: (a) percentage of students who qualify for free or reduced-price lunch; (b) location type (i.e., urban, suburban, rural); (c) grade level; (d) starting RIT score; and (e) number of actual instructional weeks between the test events.[3]

[3] See https://www.nwea.org/resources/virtual-comparison-group-vcg/

Research Questions

The purpose of this project was to explore the feasibility of using the MAP VCG as a comparison indicator, possibly a benchmark indicator, and as evidence within the overarching framework of academic and school effectiveness collected to inform reauthorization decision making. The following questions guided this project:

1. To what extent do NMUCSO schools match the MAP norm group demographics and fall percentiles?
2. To what extent does the MAP norm-based CGI relate to the MAP VCG index?

Project Scope and Timeline

In summer 2014, NMUCSO embarked on a project to revise their reauthorization decision-making process. Over the two years that followed, the SDP Fellow played a role in all three of the following priorities that were part of the agency application to participate in SDP: (a) establish a formal and transparent reauthorization decision making matrix; (b) create an annual dashboard tool to provide for stakeholders; and (c) formalize the annual monitoring cycle/reports vis-à-vis five-year reauthorization contracts. More specifically:

1. Create a sustainable data collection/analysis process to improve data-driven decision making regarding reauthorization of K–12 charter schools authorized by NMU.
2. Establish appropriate comparison groups (i.e., districts) for the K–12 charter schools NMU authorizes based upon accurate subcategories of student populations.
3. Create performance indicators to use within reauthorization contracts that help to improve data collection/analysis and decision making at the site level (i.e., connecting the system to the specific schools).

The SDP Fellow took a lead role in determining the feasibility of using NWEA MAP VCG data as part of the third priority listed above.

Results

Key successes included creating the annual dashboard and working on Research Questions (RQ) 1 and 2. Key challenges included the scope of the overall project and the limited staff available to work on the proposed products. In the subsections that follow, we describe this work as well as anticipated next steps.

RQ1: To what extent do NMUCSO schools match the MAP norm group demographics and fall percentiles?

NWEA does not collect group demographics for all students; therefore, no comparisons by ethnicity/race are possible. However, when analyzing literature on the achievement gap, Rodrizguez, Nickodem, Palma, and Stanke (2016) posited: "(For achievement data) to be useful to schools and communities, we must provide descriptions to capture the full variability and complexity of student characteristics" (p. 9). To better understand the current ethnic demographics for the NMUCSO charter schools, Figure 2 documents the percentage of each school by ethnicity, as recorded by the Michigan Department of Education based upon test data from School Year (SY) 2015–2016.

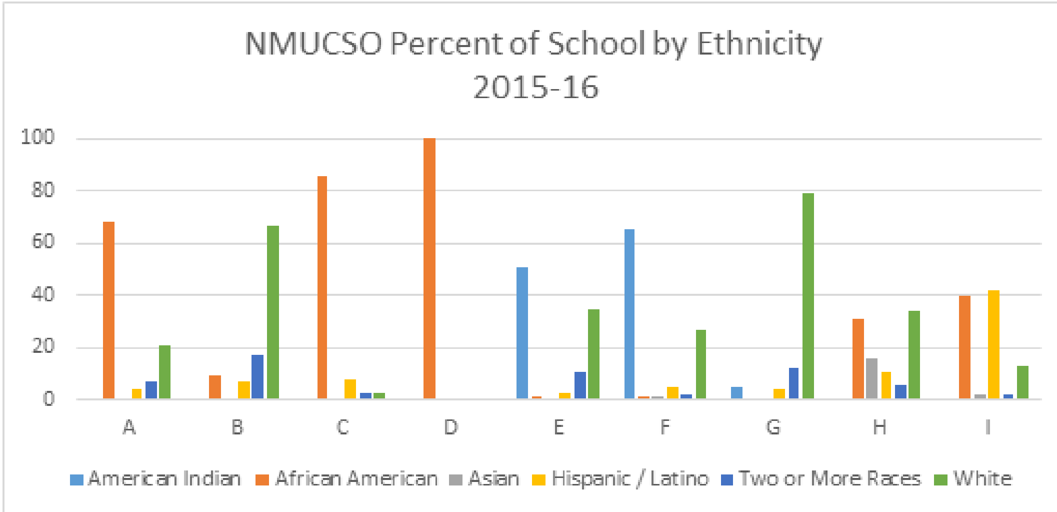


Figure 2. Ethnicity of NMUCSO charter schools by school for SY2015–2016.

According to NWEA (2015b), the percentage of a school matching the normed distributions would reflect the following criterion: "If the students selected matched the achievement distribution across the country, each bar [decile] would be at 10%" (p. 2). Figure 3 displays NMUCSO test takers' fall 2015 RIT scores converted to percentile, grouped by deciles for each of the nine schools. Clearly, no school matched the 10% test takers per decile criterion. For seven of the nine schools, the fall 2015 data suggest an over-representation of low-performing test takers.

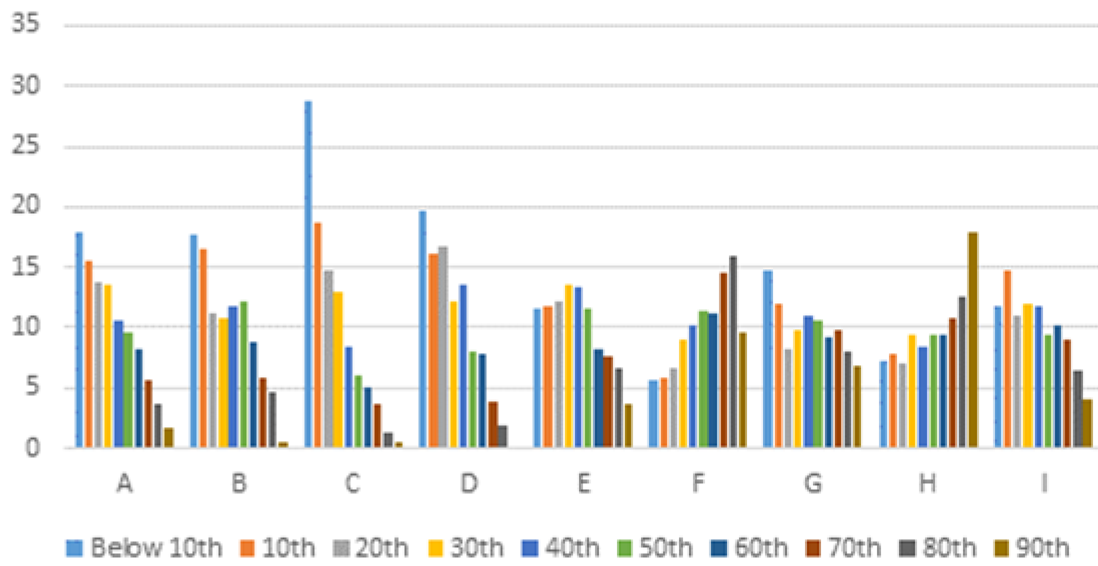


Figure 3. Percentage of NMUCSO students by school as deciles for fall 2015 RIT.

RQ2: To what extent does the MAP norm-based CGI relate to the MAP VCG index?

Figure 4 was created as a baseline to display grade-level performance for Grades K–7 mathematics MAP data for fall 2014 to spring 2015, comparing the VCG index data (vertical) with the fall 2014 to fall 2015 GCI data (horizontal) using index score differences. The individual grade-level graphs document the number of students over-performing (green) and under-performing (red) the expected value for the VCG index. The regression lines are included for future comparisons.

Further analyses will compare SY2015–2016 to the SY2014–2015 data below to establish trends for the percentage of students in Quadrant 1 (overperforming both GCI and VCG index), Quadrant 2 (overperforming VCG but underperforming GCI), Quadrant 3 (underperforming both GCI and VCG) and Quadrant 4 (overperforming GCI and underperforming VCG). These percentages will be used to pilot test benchmarks described in the section on next steps below.

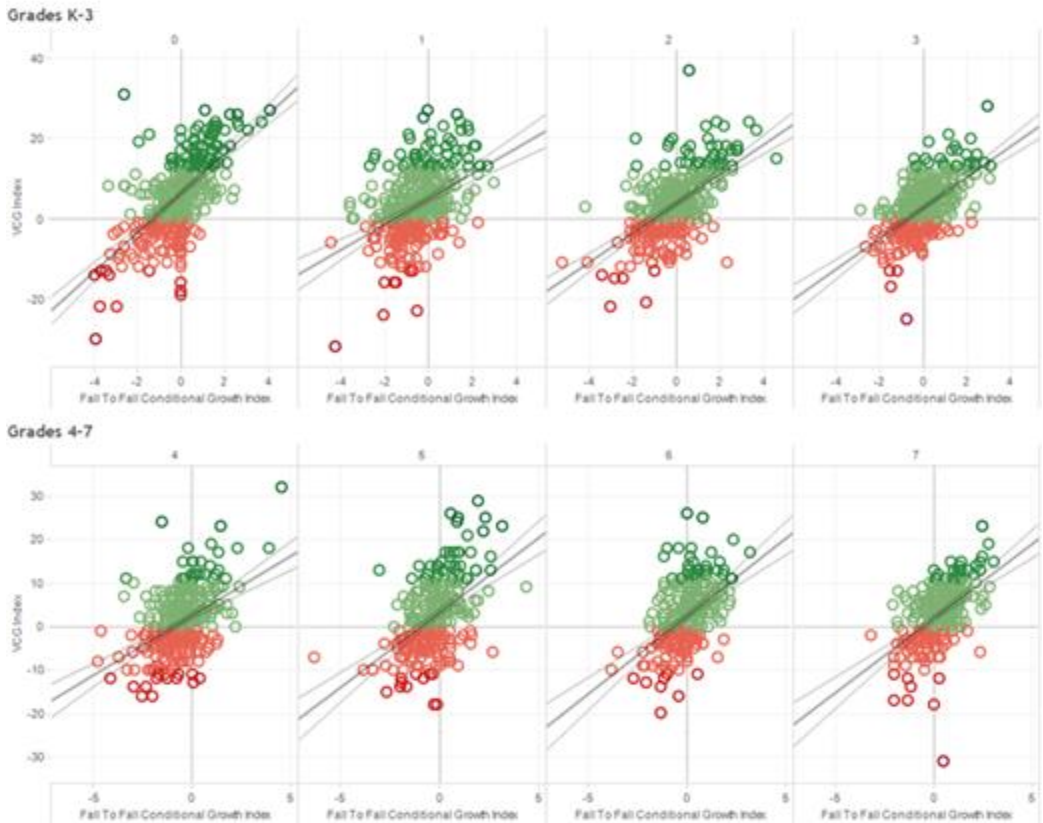


Figure 4. MAP VCG and CGI for mathematics, grades K–7 in SY2014–2015.

Key Impacts

The findings from the two research questions are making a significant difference for NMUCSO to better understand how MAP results may be used in determining educational effectiveness. Specifically, RQ1 documented school composition, based upon ethnicity, along with fall 2015 starting RIT scores that showed the percentage of students by school under the 50th percentile. This provided a valuable argument to analyze each school performance individually.

As mentioned above, NMUCSO is working on determining what comparisons and benchmarks are appropriate for the schools. In May 2013, NWEA researcher Dr. Nate Jensen encouraged caution when interpreting the percentage of students meeting or exceeding MAP growth projections: “[S]ince these projections provide estimates of average performance, it is important to note that on average, not every student will meet or exceed his or her growth projection—in fact, we generally observe that approximately 50%–60% of students nationwide meet or exceed their growth projections." This has impacted our stakeholders.

Next Steps

The ongoing work associated with SDP and answering these research questions not only laid the foundation for improved understanding of the specific schools, but it is also positioning NMUCSO to advocate for a more meaningful set of indicators of student achievement. Given the results and continued consultation with NWEA researchers, the following three targets will be piloted tested for SY2016–2017:

1. Sixty percent of students by grade level will meet or exceed a "zero" (median) conditional growth index from fall each academic school year to fall the subsequent academic school year in reading and mathematics for all grades tested.
2. Sixty percent of students by grade level will meet or exceed a "zero" (median) VCG index in reading and mathematics from fall to spring in the same academic school year.
3. By grade and subject, the school will meet or exceed the "Average" category for the median growth percentile.

Performance for this MAP component of the school dashboard will be reported using an appropriate color coordinated scale. For example, dark green is 60%+, light green is 50%–59.9%, yellow is 45%–49.9%, orange is 40%–44.9%, and red is below 40%. Further, as the SY2015–2016 MAP VCG data become available, the same descriptive analysis will be conducted to compare the two years and to run regressions to further compare local data to national norms. One especially noteworthy next step is to compare the percentage of students from SY2014–2015 to SY2015–2016 falling into each of the quadrants in Figure 4. This future analysis will help to determine how appropriate the 60% targets are for each school. One challenge still facing NMUCSO is to determine how sensitive using "percentage" as a benchmark is for smaller schools. Some grade levels within schools have very few students (10–18), and literally one or two students can cause performance category (i.e., color) differences in benchmark reporting.

Case Study 2:

Uplift Education

Uplift Education is a free public college preparatory charter school network serving 14,000 scholars in pre-K to Grade 12 at 34 schools on 17 campuses in the Dallas-Fort Worth region. Uplift students are 67.2% Hispanic, 18.4% Black, 8% Asian, 4.4% White, and 1.7% are of two or more races. About 74% of Uplift students qualify for free or reduced-price meals, 31.5% have limited English proficiency, and 5.9% receive special education services. Eighty percent of Uplift graduates are the first in their families to attend college, and 100% of Uplift Graduates are accepted into college, as this is a requirement for graduation.

In kindergarten through Grade 8, Uplift uses the MAP to measure student growth, and as a proxy to measure progress on the TEKS, the standards that are covered on the State of Texas Assessment of Academic Readiness (STAAR), the high-stakes end-of-year tests that students in Grades 3–8 take (as well as five end-of-course exams taken after specific high school level classes). Uplift administers MAP three times a year, in the fall, winter, and spring, and uses the fall data to set spring MAP goals for all students.

Over the past three years, Uplift has expanded by two schools or more a year, and added grade levels in existing schools, resulting in growth from approximately 7,000 students in SY2012–2013 to 14,000 students in SY2015–2016. To ensure that teachers are able to meet the needs of Uplift students, the program has prioritized providing teachers with strategic information about their students to help guide them to target skills gaps in their lesson planning and teaching. The result has been the deployment of data dashboards to display student and teacher data to all teachers, principals, and staff throughout the district.

Uplift Education enrolled one SDP Fellow, who is the director of data strategy. The fellow oversees assessment, internal reporting and analysis, custom application development, and database infrastructure for the district.

Research Questions

The purpose of this project was to determine how NWEA's MAP can be used to drive instructional decision making through predictive modeling and visualization. Uplift's research questions were:

1. To what degree are Uplift scores on MAP correlated with Uplift scale scores on STAAR?
2. Can correlations be used as a flag to identify students needing strategic intervention?
3. How can these data be presented to principals, assistant principals, and teachers?

Project Scope

In order to better serve the wide range of Uplift students, the MAP has been used to measure student growth across a year, and MAP performance was correlated with STAAR performance to create predictions to inform teachers and leadership of how likely a student is to achieve a certain level of proficiency on STAAR. The STAAR is based on the Texas Essential Knowledge and Skills (TEKS), which, while not the same as the CCSS, were designed to prepare students for college and careers. TEKS are estimated to have an 80% overlap with the CCSS (Texas Education Agency, 2009).

NWEA MAP publishes linking studies that give a correlation of STAAR performance with MAP achievement, based on the time of year that the MAP assessment is taken. Uplift has utilized these data in making predictions of student performance, but there was concern in SY2013–2014 that the linkage between STAAR and MAP was based on a population of students not representative of Uplift students. Uplift examined the correlation of MAP and STAAR scores for its students, and used these correlations to create predictions to inform teachers and leadership of how a student will perform on STAAR in Grades 3–8. The models were used as another data point in triangulating the type of support that a student, classroom, grade level, or school might need in order to increase the likelihood of STAAR passage.

Timeline and Results

A year before Uplift's partnership with SDP, Uplift entertained the idea of developing a predictive model. Uplift partnered with University of Dallas graduate students to build a growth model as part of a performance pay system build funded by a Teacher Incentive Fund grant from the Department of Education. Work began at Uplift in SY2013–2014, with investigation of MAP to be used as a predictor of STAAR Scores. The initial intention was that a growth model could use fall MAP scores to set a spring STAAR goal for each student. Work began by reviewing relevant research on early warning systems and predictive models. Additionally, Uplift reached out to other charter districts to learn more about their predictive model work.

In this first year, a model was created using fall and spring MAP data to predict STAAR scores, but it was not rolled out to the district, as the results of the model were noisy for many grade levels due to small student populations in SY2012–2013. However, results showed that in grades with large enough student populations, like Grade 3, the model was promising enough to entertain a rollout to principals and assistant principals in SY2013–2014 for Grades 3–8. At the conclusion of SY2013–2014, discussions began within the data team at Uplift and among senior staff around how a model like this might be explained to leaders and teachers.

In SY2014–2015, with the beginning of Cohort 6 of the SDP Fellowship program, work on the MAP-based predictive models began again. In the winter of SY2014–2015, a regression model using fall MAP and two Uplift-created district assessments was used to give predicted grade-level performance on STAAR, showing the percentage of students in each grade level at the performance bands of “Unsatisfactory,” “Satisfactory,” “Recommended,” and “Advanced.” This dashboard was released to principals and assistant principals just prior to the winter break. However, individual student-level results were not released with this prediction model, for fear of principals and assistant principals only focusing on “bubble students”—that is, those with predicted scores that were closest to the cut point for STAAR passage.

In the spring of SY2014–2015, student-level predictions were released for specific schools, in order to pilot the reception of these results and survey a group of principals and teachers on how they used the data. The models used fall and winter MAP scores, and included the Uplift district assessments. These predictions were largely used by principals to target intervention resources in particular classrooms. This was the first year that buy-in began for predictive models, as those principals using them saw their value and shared their learnings with peers within the Uplift network of schools.

In the winter of SY2015–2016, another linear regression model was created based on fall MAP and the two district assessments. These results were released at the student level for the subjects of mathematics and reading just prior to the winter break. In early February 2016, updates to the models were created, using fall and winter MAP, along with the two previously given internally developed assessments (Equation 1):

$$STAAR\ Scale\ Score = \beta_0 + \beta_1(Fall\ MAP\ RIT) + \beta_2(Winter\ MAP\ RIT) + \beta_3(Fall\ Uplift\ Assessment) + \beta_4(Winter\ Uplift\ Assessment) + \varepsilon \quad (1)$$

These models were also released at the student level for every school with Grades 3–8, including Algebra 1 test takers in Grade 7 or 8. In late February, performance on the MAP goal areas was visualized on a dashboard for all teachers and leaders in order to show the prediction, along with the skills gaps of the students. This allowed users to sort the data and more easily find groupings of students with similar skills gaps. The coefficient of determination for these models is displayed in Table 1.

Table 1

R^2 Values for Winter 2016 Predictive Models

Grade	Math	Reading
G3	0.75	0.72
G4	0.79	0.71
G5	0.79	0.69
G6	0.80	0.68
G7	0.74	0.68
G8	0.63	0.66
Algebra 1	0.65	

Key Impacts

Though not intended, Uplift's initial caution in rolling out the results helped build buy-in. Schools that were initially provided student-level results were those most in need of using the data to work with students unlikely to pass STAAR, given their current trajectories. Upon obtaining STAAR results and comparing them to predictions, principals developed confidence in these models. Coming into SY2015–2016, many principals and teachers were asking when the first round of predictions would be available so they could begin using the data to strategically intervene with students.

With results rolled out via dashboards, Uplift principals and teachers could interact with the data. Screenshots of the SY2015–2016 STAAR prediction dashboard can be seen in Figure 5 below, and with further detail in the appendix. The rollout via dashboards pushed principals and teachers to ask for a breakdown of performance on the TEKS for each student to be shown along with their prediction, so that teachers could group students with similar skill gaps and strategically intervene on particular skills.

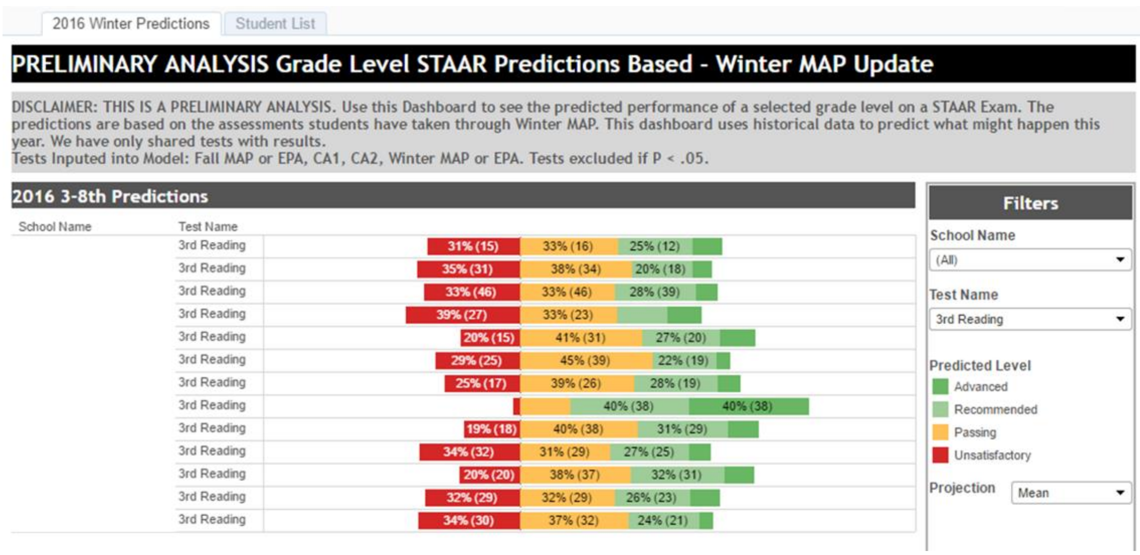


Figure 5. Grade 3 STAAR predictions dashboard, grade-level predictions.

Next Steps

Seeing their predicted performance and how that might impact their accountability ratings with the state pushed schools to carefully consider these data. A dashboard was created to show predicted accountability based on STAAR predictions. This served to engage those who were unsure of whether their student performance would result in not meeting the accountability standards set by the state of Texas.

Next steps are focused on three areas—model validation, refinement of the display of results, and encouraging the use of action based on the predictions. The first two steps will help users of the models better understand what the models are telling them. The final step encourages data-driven decision making based on the results of the models, and measuring whether action taken based on the models leads to improved student outcomes

As of July 2016, Uplift has been using the state assessment data to determine the classification rate of every model used by Uplift, as well as evaluate the predictions. This information will be published out to teachers and leaders in order to create transparency around the models, as well as for purposes of scientific reproducibility. Additionally, Uplift will analyze STAAR predictions by school, grade level, and classroom to get an understanding of where there are deviations from the models and to investigate lessons that can be learned about what may have caused performance to be different than what was anticipated, either from a modeling standpoint or actions taken in the classroom, grade level, and/or school.

Next, the data team at Uplift will adjust models based on the results to improve prediction accuracy for SY2016–2017. Uplift has recently created a MAP and STAAR predictive model based off of spring MAP data and prior year STAAR data to predict next year's STAAR scores, tentatively named the year-over-year model. Students in Grades 2–7 at Uplift in SY2015–2016 had a STAAR scale score generated for each tested subject they have to take in SY2016–2017. During the summer of 2016, Uplift worked to refine the dashboard display to make manipulation of prediction data easier and more simply visualized, as well as to place the year-over-year predictions on the SY2016–2017 version of this dashboard. The SY2016–2017 version of the dashboard can be seen in Figure 6.

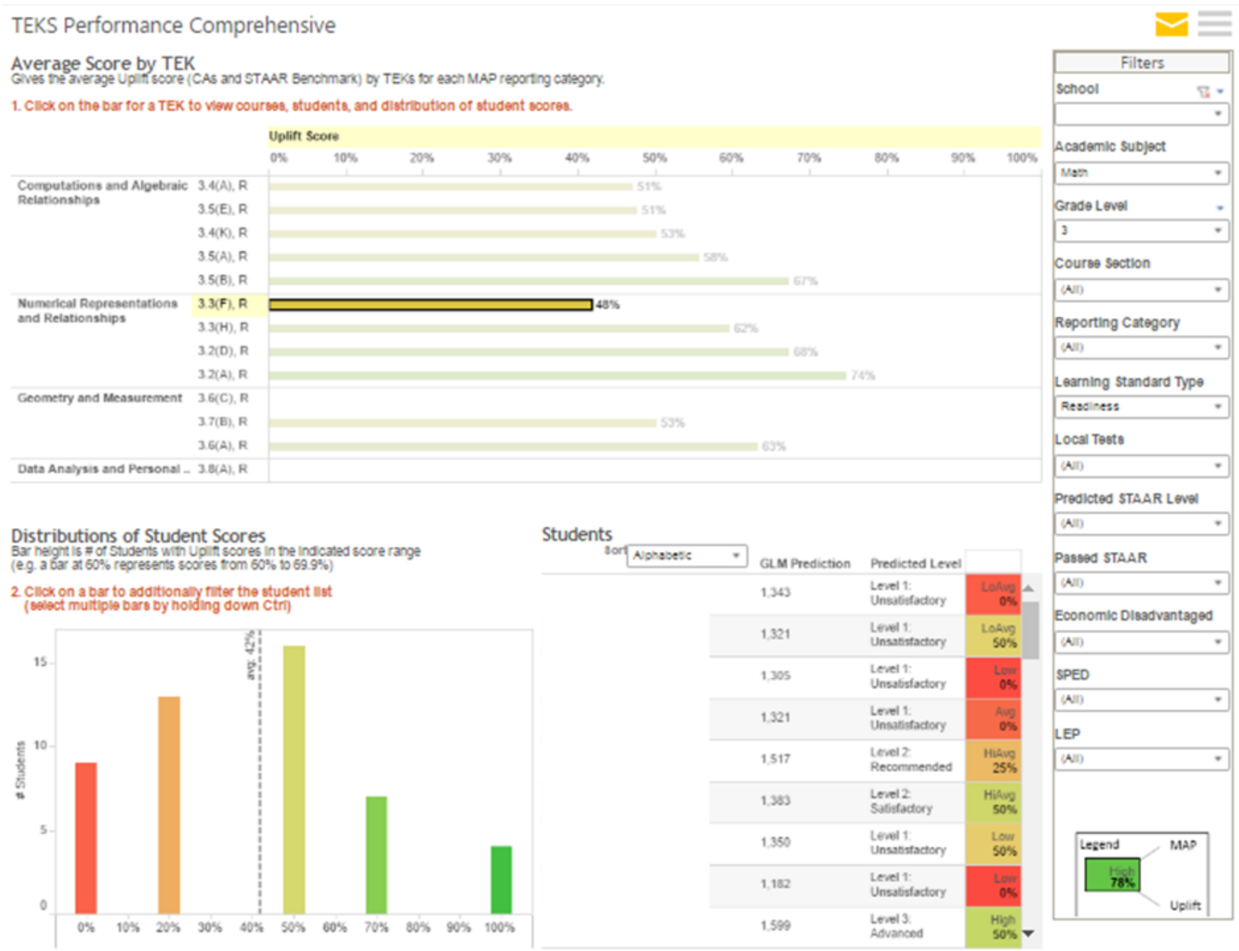


Figure 6. TEKS performance dashboard with STAAR predictions.

The final step is to take a more granular look at MAP and STAAR data. Over the fall of 2016, Uplift will examine the correlation of MAP goal areas to TEKS, using the most recent data released for the 2016 STAAR administration. An examination of the TEKS to MAP goal area alignment that Uplift and other Texas districts have created is necessary, as instructional decisions have been made based off of this linkage, and it has yet to be mathematically verified. A similar analysis has recently been completed for Uplift's internal assessments, the Common Assessments. By extending this to MAP goal areas, a potentially more granular prediction can be made. The result would be more targeted instruction based on results of MAP.

To encourage action based on STAAR results, Uplift spent the summer of 2016 developing trainings and meeting protocols to better inform teachers and leaders about the predictions. By providing trainings and in person meetings with teachers and leaders about the predictions throughout SY2016–2017, Uplift hopes to define plans of action that teachers and leaders can put in place. The Uplift data team will document these plans and see if they are correlated with students outperforming model predictions. Results from this work will be available in early fall of 2017.

Uplift has demonstrated thus far that teachers and leaders can be provided with predictive models based off of an assessment system. The challenge is to push the district to use these data to drive thoughtful instructional decisions that improve outcomes for students. Over the next year, Uplift will build upon what has been created thus far, and will continue to share out results, successes, and lessons learned.

Case Study 3:
West Linn-Wilsonville School District

The West Linn-Wilsonville School District serves more than 9,500 students and covers 42 square miles in the south metropolitan area of Portland, Oregon. There are 15 schools, including an alternative high school, in this pre-K–12 public school district. Students in WLWVSD are 12% Hispanic, 1% Black, 5% Asian/Pacific Islander, 7% multi-ethnic, and 76% White. About 7% of WLWVSD students are English learners, 10% receive special education services, and 23% qualify for free or reduced-price meals. The school district comprises two cities, West Linn and Wilsonville, each with a growing population between 20,000 and 30,000, and their surrounding more suburban and rural communities.

WLWVSD seeks to create conditions that promote a growth mindset, rigor, and supports for all students, as well as the systemic use of constructivist approaches to teaching and learning. The district uses assessment results along with evidence of learning based on participation and achievement in core academic courses that include advanced and accelerated options, as well as participation and performance in arts and STEM programs, to annually evaluate programs and set goals at district and school levels. Decisions about the use of resources and areas of focus for improvement are based on this information. The district's constructivist orientation and focus on growth mindset prompts interest in an assessment system that helps students, teachers, administrators, and parents see students' growth and achievement. The system strengthens the use of formative, interim, and summative evidence of learning for instructional decision making. Efforts are underway to link teacher learning and practice to student learning outcomes.

WLWVSD enrolled its director of curriculum, instruction, assessment, and research to be an SDP Fellow. The affiliation with SDP is WLWVSD's first step in developing a means of systematically using data visualization and providing sets of data across schools and grade levels. Although there have been cyclical approaches in place at the district and school levels to collect and analyze data, develop a theory of action, and determine responsive practices and strategies, implement actions, and then again examine outcomes, the data have consisted mostly of summative state assessment results for reading and mathematics at the district level and school-based F&P or the Developmental Reading Assessment 2nd edition (DRA2) for kindergarten through Grade 5 reading. Oregon adopted the Smarter Balanced Assessment Consortium (SBAC) assessments for reading and mathematics and administered them for the first time in 2015.

WLWVSD is working to develop a balanced assessment system that has aligned formative, interim, and summative assessments. This case study is looking at use of and alignment between interim and summative assessments. Specifically, analyzing the relationship between MAP and SBAC scores can serve as a starting point for looking at progress towards outcome measures and help with identifying strengths and areas for instructional attention. The district has integrated assessment literacy into professional learning for teachers and principals and used available data to inform program and instructional decision making for the past decade. The purpose of this case study is to examine the use of interim assessments as predictors of performance on summative assessments and to make recommendations about their use in informing instructional practice.

Research Questions

Five questions guide WLWVSD's work:

1. How well do interim (MAP) assessment results predict student performance on the SBAC assessments for mathematics?
2. What do MAP quintiles for cohort and grade-level mathematics results suggest about students' progress and growth?
3. What does the MAP CGI metric suggest about student learning in mathematics?
4. How could district-, school-, and grade-level teams use interim assessment results to inform instructional practice and the use of curriculum?
5. How could interim assessment data contribute to a district's evaluation of the effectiveness of professional learning in mathematics?

Project Scope and Timeline

The Oregon Department of Education adopted SBAC's English language arts and mathematics assessments and administered them for the first time in SY2014–2015. This was part of replacing content-based standards with the CCSS and implementing Senate Bill 253 in 2011, which set a goal for 100% of Oregon students to earn a high school diploma, 40% to earn an associate's degree or equivalent postsecondary certificate, and 40% to earn a bachelor's or advanced college degree. This goal was legislated with the intent to address Oregon's graduation rate, which increased from 66% to 77% from 2010 to 2015 but is still among the lowest in the United States.

The overall percentage of WLWVSD students who meet or exceed the benchmarks on state summative reading and mathematics assessments has consistently averaged about 10 points above the state average for the past decade, and this remained constant with the SY2014–2015 SBAC results. The graduation rate, which was 93% for SY2014–2015, is relatively strong. However, there is a need to know how well students are prepared for the college and career readiness focus of the SBAC assessments—WLWVSD results on SY2014–2015 SBAC English language arts (includes reading) and mathematics showed a 10% to 15% drop in the percentage of students whose scores met or exceeded benchmark cut scores in both subject areas across all grade levels from the previous Oregon assessment scores. Although this drop in scores was 15% to 20% greater for other school districts in Oregon—and the Smarter Balanced consortium predicted an initial drop in scores when states shifted to more rigorous assessments that align with the CCSS—the WLWVSD superintendent and central office staff guided the school board in developing goals in SY2013–2014 that included (a) improving achievement for all students by raising rigor and eliminating opportunity gaps, and (b) aligning assessment, professional learning, and implementation of strategies to ensure equitable outcomes and growth for students and educators.

WLWVSD initiated the use of NWEA MAP interim assessments for reading and mathematics in SY2013–2014 by inviting all kindergarten through Grade 8 principals, assistant principals, and instructional coordinators from each school to an initial overview and training. All nine primary (K–5) schools used the mathematics assessment at least once during the school year in Grades 2–5; six primary schools also tested K–2 students in mathematics. All nine schools used the reading assessment at some grade levels. One of the three middle schools assessed all students in Grades 6–8 in reading and mathematics twice during SY2013–2014; the other two middle schools used both assessments with a small number of students at each grade level at least once during that year.

The SDP Fellow coordinated the use of the reading and mathematics MAP assessments in SY2014–2015 and SY2015–2016, developed an assessment inventory so principals could engage teachers in identifying sources of evidence of student progress and track how that information was being used for instructional decision making, and coached principals in using MAP reports and data visualizations with teachers. In SY2014–2015, all primary and middle schools gave the MAP assessments in reading and mathematics in the fall, winter, and spring. Also in SY2014–2015, all primary schools began using the same assessment, the DRA2, to determine students' reading levels. That spring, students in Grades 3–8 and Grade 11 participated in SBAC assessments in language arts and math for the first time.

Although there was initial interest in the information about students' progress that would be available through the use of MAP assessments, principals were finding that following assessment schedules and procedures for administering the tests was becoming a focus point for teachers. In addition, as preparations for the new SBAC assessments got underway, some teachers, parents, and principals were concerned about the time given to assessments and how children would experience them. Engaging all stakeholders in seeing students' growth and how this information could be used for planning and addressing learning needs was critical for creating a culture that valued using evidence of learning to improve practices.

During the summer of SY2014–2015, each school leadership team met with the SDP Fellow to analyze MAP and SBAC results. Data visualizations for MAP results for reading and math and SBAC language arts and math results included: a comparison of school, district, and NWEA normed median scores by grade level (for MAP); disaggregated comparisons by school and district grade levels for English learners and students served through special education; and fall/winter/spring mean growth comparisons by school and district grade level that also showed NWEA normed median scores (for MAP).

Throughout SY2014–2015 and SY2015–2016, principals and teachers were guided in looking at the percentage of students who scored within each quintile on MAP reading and math assessments so trends for grade levels and cohorts could be identified. NWEA MAP divides normed percentile rankings into quintiles, which provides a view of levels of performance that can be used to track growth and make instructional decisions. To add another indicator of growth, beginning in SY2015, the median grade level CGI metric, based on NWEA MAP norms, was included in analysis.

The cut score and probability tables, available from NWEA (May 2015), show the MAP RIT scores and percentile ranks associated with each of the four performance levels on SBAC for fall, winter, and spring assessments. These tables suggested that WLWVSD students would have performed better at all grade levels on both SBAC assessments based on fall, winter, and spring MAP scores. This raised questions about how well MAP performance predicts SBAC performance. Principals were especially concerned because of the 7 to 8 hours of testing time for SBAC in addition to the class time taken for giving MAP assessments three times a year. During SY2014–2015 and SY2015–2016, the items on MAP assessments changed so that the depth of knowledge levels moved towards matching SBAC, which specifies those levels in the assessment blueprints. This increased the time required for MAP assessments.

Because WLWVSD was embarking on the selection of K–5 math curriculum materials that aligned with CCSS in SY2015–2016, and had completed preliminary collection and coding of all K–8 teachers' reflections about professional learning in mathematics during SY2014–2015, this case study focuses on math results. The products that principals, teachers, and district administrators received and used include:

1. Correlation between winter 2015 MAP mathematics scores and spring 2015 SBAC mathematics scores for Grades 3–8.
2. Charts showing each school's fall, winter, and spring mathematics quintiles in K–8 and the district grade level quintiles.
3. Charts showing district CGI grade-level means for K–8.
4. Assessment Inventory/Assessment Use tool.
5. PowerPoint presentations with protocols to guide principals in interpreting and making decisions based on reading and mathematics data with the staff at each school.

Results and Key Impacts

SBAC and MAP. Linear regression analysis was used to look at how well winter MAP mathematics scores in Grades 3–8 predicted SBAC mathematics scores. The correlation coefficient ranged from .65 to .74 for Grades 3–8. Regarding RQ1, this analysis suggested that

MAP mathematics scores for Grades 3–8 were a moderate to strong predictor of students' SBAC math performance. Figure 7 shows general consistency in the relationship between MAP mathematics and SBAC scores.

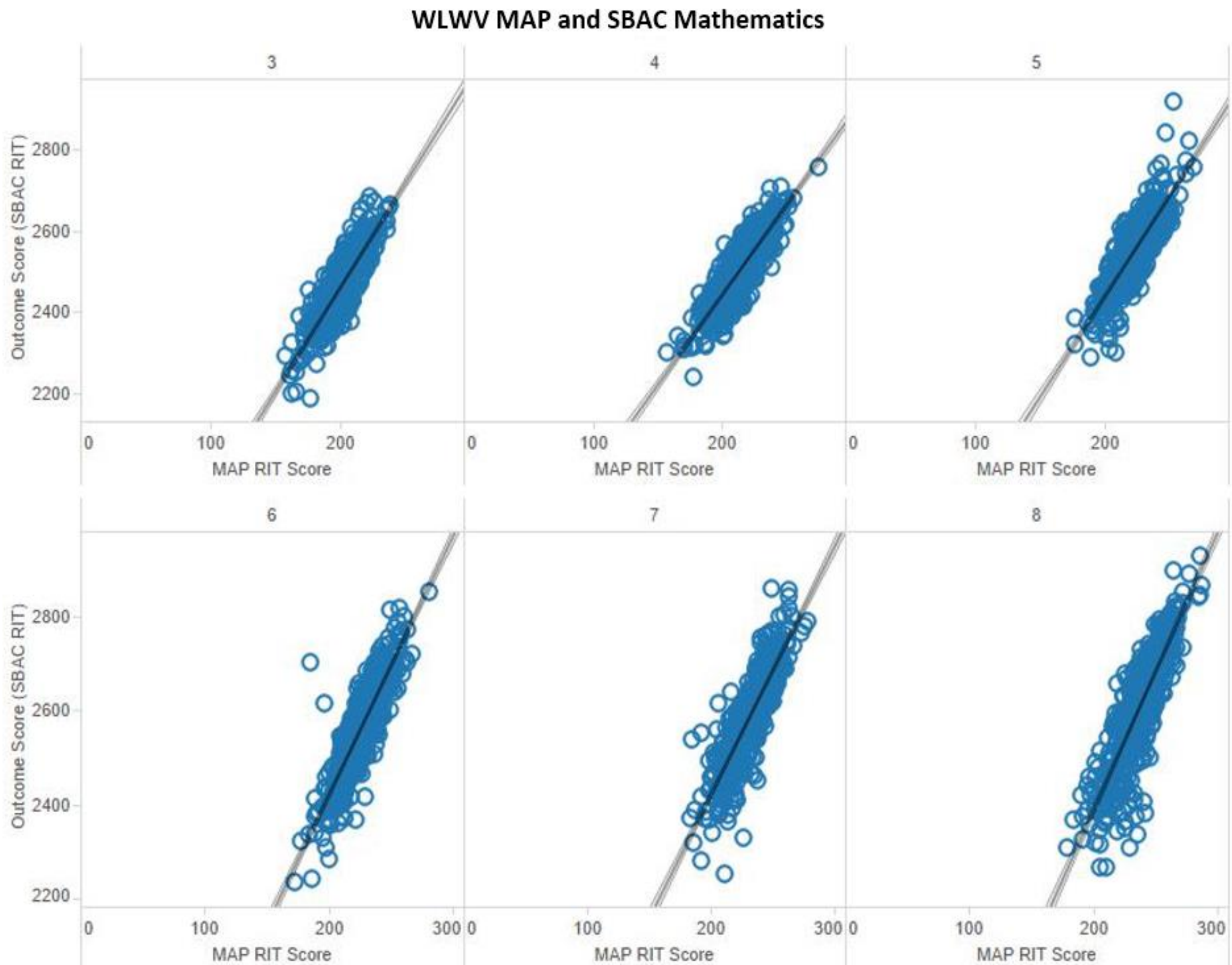


Figure 7. WLWVSD MAP mathematics and SBAC mathematics scores for SY2014–2015.

The correlation between MAP and SBAC scores contributed to affirming district and school leaders' commitment to aligning formative, interim, and summative assessments and to integrating instruction, curriculum, and assessment. This led to principals' and teachers' increased interest in and work with developing math performance tasks using the SBAC blueprints so that teachers would have rigorous, standards-based tasks and could use them formatively and respond with instructional strategies. WLWVSD will compare MAP reading and math scores to SBAC scores again when the SY2015–2016 SBAC scores are available. Likewise, the district will compare the NWEA MAP fall, winter, and spring predictor scores at each grade level to district scores to continue to establish WLWVSD predictor scores.

MAP quintile analysis. The evidence for RQ2 suggested that more could be found from examining the relationships between quintile scores, professional learning, and the subsequent implementation of practices than could be implied about students' progress and growth from just this information. The analysis of MAP math quintile trends for fall, winter, and spring in SY2014 and fall and winter in SY2015–2016 showed that the greatest decrease in the percentage of students whose math scores were below the 21st percentile was in the cohort of students followed from Grades 2–3. This overall district result was repeated in each primary school. The largest increase in the percentage of students whose math scores were above the 80th percentile was also seen in the Grade 2–3 cohort in each primary school. Figure 8 shows the trend that is predicted for the next two assessments using the linear trend algorithm.

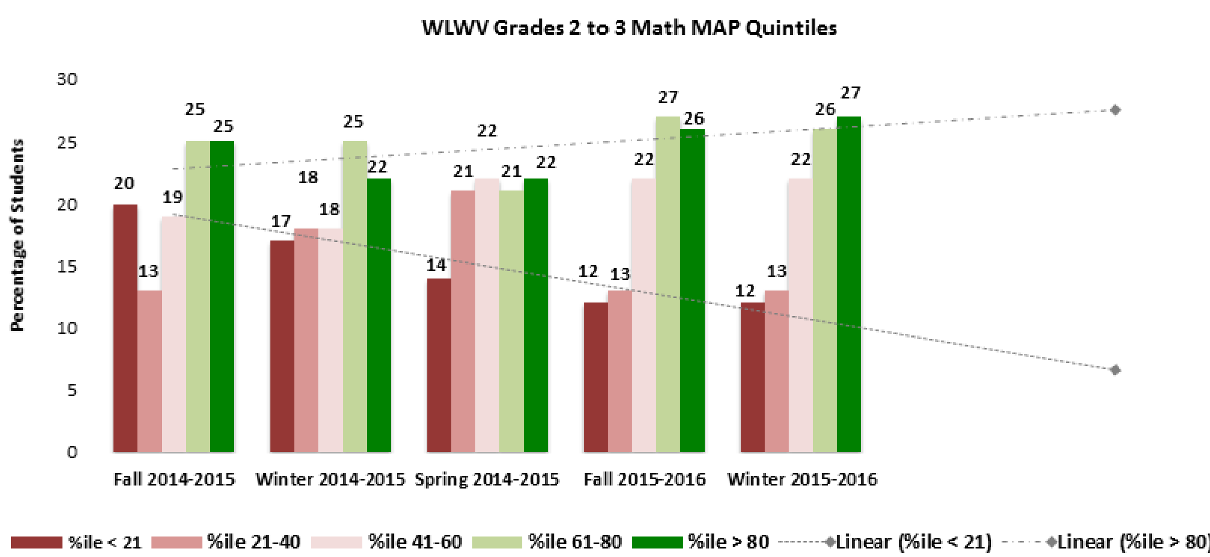


Figure 8. WLWVSD Grade 2–3 cohort mathematics MAP quintiles for SY2014–2015 and fall and winter of SY2015–2016.

Aside from the Grade 2–3 results, the quintiles for reading and math at each grade level varied across schools by as much as 15% for each quintile. The quintile charts served as a starting point for principals, guided by district office staff and the use of data interpretation protocols (Boudett et al., 2015), to identify grade-level and cohort patterns. Instructional strategies introduced and practiced in math professional workshops in SY2014–2015 were associated with results that showed a decrease in the percentage of scores below the 21st percentile and increase in the percentage of scores above the 80th percentile.

District office staff and principals integrated the quintile charts with mid-year cycle-of-inquiry classroom walkthroughs and used observation of practice tools to look for implementation of math practices. There are only associations between the practices that teachers identified for administrators to look for in their classrooms, the observations of implementation, and the quintile data. Nevertheless, there are trends that will be checked through the collection and interpretation of quintile percentages, evidence of implementation of practices, and collected and coded statements about what teachers identify as practices that will be seen and heard in their classrooms during SY2016–2017. Data displays that relate professional learning participation, observed practices, and MAP math quintile scores will be developed.

The quintile results for fall, winter, and spring of SY2014–2015 and fall and winter of SY2015–2016 show that more than 50% of Grade 5 students scored above the 60th percentile in mathematics. WLWVSD used these results, along with SY2014–2015 SBAC scores and Grade 5 formative assessments in mathematics, to determine that the Grade 5 math curriculum would include standards and units of study from Grade 6. In a process where teachers and principals used the CCSS math practices and kindergarten through Grade 8 MAP quintile results, they selected a set of math curriculum materials with higher depth of knowledge levels within the lessons and problem sets.

Conditional growth. The MAP CGI, the metric used in RQ3, suggested that the greatest amount of negative conditional growth from fall to winter in mathematics was in kindergarten, followed by Grade 1. Students in Grade 7 showed the most growth according to the CGI. Figure 9 shows that there is also positive conditional growth in Grades 2, 3, and 6, and negative growth in Grades 4, 5, and 8.

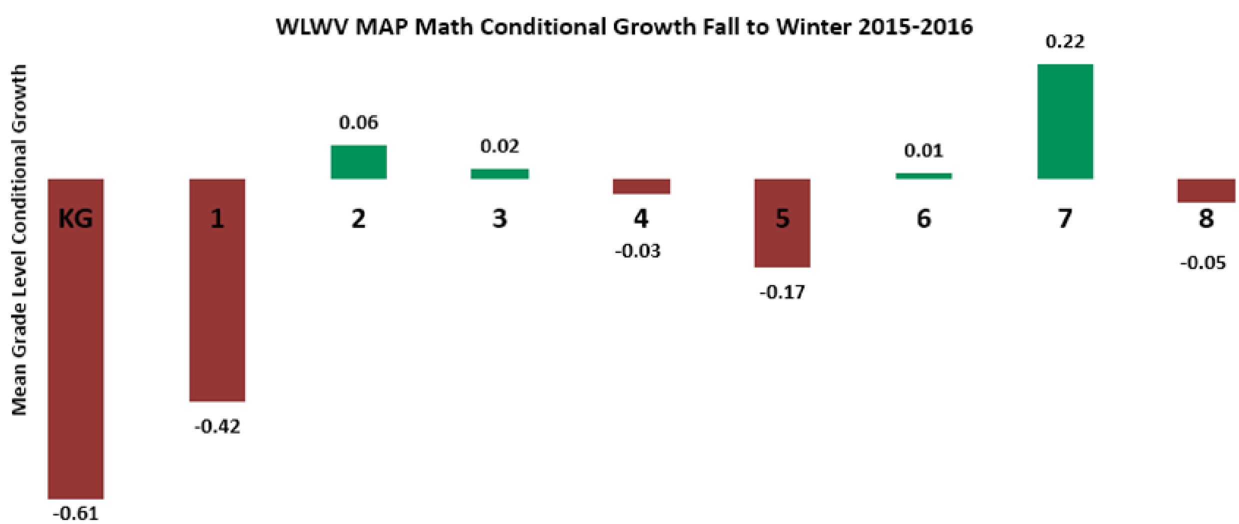


Figure 9. WLWV MAP mathematics conditional growth in fall to winter, SY2015–2016, by grade level.

WLWVSD will include the CGI in MAP analyses for each school year and as NWEA MAP researchers recommend, look at conditional growth from fall to spring or over the course of three testing periods. Using the CGI along with quintile cohort results provides a starting point for looking at group growth. The results raised questions about the new learning in mathematics in kindergarten and Grade 1. WLWVSD has focused on professional learning in these grades that promotes practices that use approximations of learning and the regular introduction of new math learning at a more rapid pace and that reflects higher academic expectations for students in their first 2 years of school, especially in mathematics. The CGI will be used with grade-level and cohort quintile results in mathematics and reading to generate questions about curriculum and instructional practices.

Next Steps

WLWVSD will continue to use multiple indicators of students' achievement and growth and to strengthen assessment literacy in the following ways:

1. Using students' performance on goal areas within MAP reading and mathematics assessments, the domains of the assessments, to identify areas of strength and growth that can be addressed through instruction.
2. Including the goal areas (within MAP) and the similar claim areas (within SBAC) in the annual regression analysis.
3. Coding implementation of instructional practice data and including it in the analysis of MAP and SBAC results.
4. Creating simple dashboards for principals to use regularly to look at formative, interim, and summative assessment results.
5. Regularly practicing the data analysis protocols that have been introduced so principals use them independently in their triad groups.
6. Tracking the introduction of new learning in mathematics at each grade level and comparing that information to quintile and CGI trends.
7. Engaging principals and teachers in using MAP and SBAC results to embed formative assessment practices into daily instruction.

By providing district staff, principals, and teachers with simple data displays that show the relationship between national norms and the local context, student achievement information is more credible. It becomes more usable when those who make instructional decisions are also regularly engaged in data analysis. The initial analyses in this case study focused on checking how well an interim assessment predicted achievement on a national summative assessment. Questions about growth offered more insight into students' learning, and led to looking at quintiles and the growth index, which gave a more complete picture of all students' progress. If other districts are interested in this type of analysis, identifying and correlating benchmarks—by examining the formative, interim, and summative measures they use—is a starting point.

Because the strands or domains with the interim and summative assessments (referred to as *goal areas* in MAP and *claim areas* in SBAC) break down an overall score into areas of learning, it would be helpful to also run analyses that show how they correlate. It would be worthwhile to set up a means of collecting data about the implementation of research-based practices emphasized in professional learning. This will allow districts to simultaneously examine factors that contribute to student achievement and allow for a more thorough analysis of outcomes. This, in turn, will give principals and teachers more information to use for making decisions about their next areas of focus.

Case Study 4:
Howard County Public School System

The Howard County Public School System is located in Maryland between Baltimore and Washington, DC. Howard County itself is one of the wealthiest counties in Maryland. The school system serves over 53,000 pre-K–12 students in 76 schools with 8,200 staff members. As of 2015–2016, HCPSS students were 40.8% White, 22.5% Black, 20.3% Asian, 9.9% Hispanic, and 6.2% of two or more races. About 20.6% of HCPSS students qualify for free or reduced-price lunch, less than 5% have limited English proficiency, and 8.5% receive special education services (HCPSS, 2016).

In this investigation of how local norms inform benchmark setting, SDP Fellows from HCPSS examined MAP scores in the context of state- and county-mandated assessments to align benchmarks across multiple measures—specifically, Fountas & Pinnell (F&P) text levels and Partnership for Assessing Readiness for College and Careers (PARCC) performance levels. HCPSS teachers use the F&P Benchmark Assessment Systems to assess K–2 students' text levels quarterly. Students' text level information is used to inform instructional groupings. The state-mandated PARCC assessments are used to measure students' progress in Grades 3–11 toward becoming college and career ready, with proficiency levels of 4 and 5 indicating college-ready performance. In addition, approaches in using these multiple measures to inform actions at various levels throughout the system, such as assessing program effectiveness or informing classroom practices, are discussed in this case study.

Research Questions

The following two questions and their related subquestions guided this case study:

1. How well aligned are the benchmarks used by HCPSS that inform students' progress toward college and career readiness? Specifically:
    a. What is the correlation between MAP reading RIT scores and F&P text levels?
    b. What is the correlation between MAP reading RIT scores and PARCC reading performance?
    c. What is the correlation between MAP math RIT and PARCC math performance?
2. How does HCPSS support staff throughout the school system in using these multiple data sources to answer questions specific to different levels of action?

a. How might school-level MAP performance be used at the central office level in assessing program effectiveness?
b. How might school administrators leverage multiple data sources to inform school improvement planning processes?
c. How might teachers access the same information to inform their professional and instructional practices?

Project Scope and Timeline

HCPSS aims to prepare students to be ready for college or a career by the end of high school so that they become successful, productive citizens. To achieve this goal, the district must make data relevant and meaningful to educators at all levels within the pre-K–12 continuum. The SDP Fellows at HCPPS, under the direction of agency supervisors, were tasked with finding ways to leverage longitudinal student data to support high-impact instructional and programmatic decisions. As a result of the work of prior SDP Fellows (in Cohort 4 of the program) on high school indicators of college readiness and fall college enrollment, in SY2013–2014 the HCPSS Division of Accountability developed and shared with system and school staff an empirically-based K–12 trajectory with grade-specific, multi-measure benchmarks aligned with college readiness and college enrollment. Major data points common across schools include F&P, MAP, and PARCC data.

The F&P reading assessments are given to all students in K–2, and at a minimum to all below-grade level readers in Grade 3 and above. MAP assessments were introduced in select schools in SY2012–2013 and SY2013–2014, and census administration in Grades 1–8 began in SY2014–2015. Schools are required to administer the fall and spring MAP assessments; the winter assessments are optional for students at or above grade level in reading. Also, whereas all

Grade 1–8 students take the MAP reading assessment, only those students not in Algebra 1 or beyond are required to take the MAP math assessment. Starting in SY2014–2015, the PARCC assessments replaced the Maryland School Assessments in reading and mathematics as the state accountability summative measure.

We emphasize MAP as an instructional tool for teachers to make better decisions around lesson planning and student groupings, and as one indicator for principals to gauge the percentage of the students at their schools who are meeting grade-expected benchmarks for being college ready. To make MAP results meaningful to and trusted by teachers and administrators, the Division of Accountability actively collaborates with the Division of Curriculum, Instruction, and Administration to support professional learning around instructional use of MAP results. Dedicated staff from Accountability collaborate with schools to facilitate the application of MAP results in the context of existing data to inform instruction and school improvement strategies. In addition, the SDP Fellow analyzed MAP performance for alignment with other measures already in use by teachers—namely, the F&P text levels. Of importance at a systemic accountability level is performance on the PARCC, which is an indicator of how prepared HCPSS students are for college. For this reason, we also examined MAP's ability to predict PARCC performance.

For central office leaders, we emphasize how MAP data serve to benchmark program effectiveness and inform program planning. In particular, we use school-level MAP performance data in evaluating the effectiveness of the Elementary School Model (ESM), a comprehensive and innovative initiative of HCPSS. The ESM aims to narrow achievement gaps between student groups by focusing on early childhood education and improving instructional practices. It comprises five components: (a) departmentalized instruction in Grades 1–5; (b) daily Spanish instruction in pre-K through Grade 5; (c) strengths-based education focused in Grades 4 and 5; (d) full-day pre-kindergarten at six schools; and (e) telehealth services for eligible students.

Starting in SY2014–2015, the ESM was implemented at six elementary schools. These schools were among those with the highest percentage of students receiving free and/or reduced-priced meals within the district. Two more schools were added to the ESM initiative in SY2015–2016. With the resources invested into the ESM, we must investigate the effectiveness of this initiative to bring about desired changes in student achievement in order to inform decisions around continuation or termination of the program. We plan to use MAP results at the school level to inform program effectiveness.

Results and Impact

Spring MAP reading and F&P were positively correlated. Students' spring MAP reading RIT mean scores for each F&P text level were computed, and they are charted in Figure 10. Since F&P levels are not on an equal-interval scale, we visualized the correlation between MAP and F&P performance to estimate the relationship between the two measures. We found a positive relationship between them.

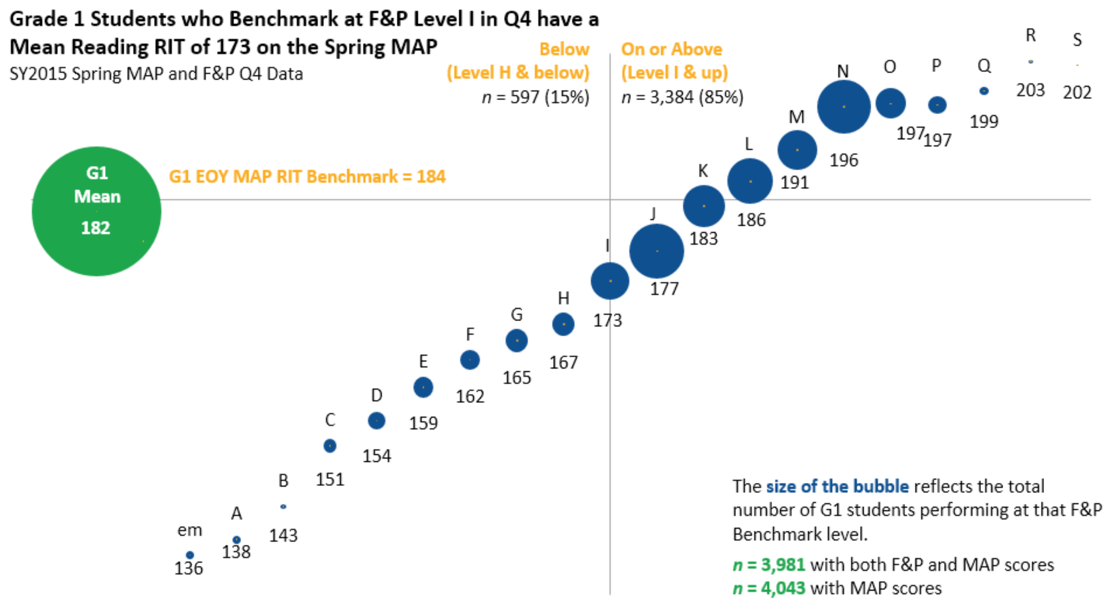


Figure 10. Grade 1 spring 2015 MAP RIT mean scores for each F&P text level.

In the HCPSS preliminary performance benchmarks for college and career readiness, the expected end-of-year (EOY) Grade 1 spring MAP RIT benchmark is 184 points, which is similar to the observed Grade 1 mean spring MAP reading RIT of 182 points. According to Figure 10, Grade 1 students who benchmarked at F&P Level K had a mean spring MAP reading RIT of 183; students at Level I achieved a mean of 173 on the MAP.

As of SY2015–2016, the Grade 1 expected F&P performance was Level I; however, based on information from the publisher of the F&P assessments and other local education agencies in the state, combined with the results of this local analysis, the HCPSS Office of Elementary Language Arts moved the EOY Grade 1 F&P text level up to Level J starting in SY2016–2017. Other reading trajectory work that occurred simultaneously to investigate whether the F&P text level should be up-adjusted for kindergarten corroborated this decision to revise the EOY Grade 1 reading level expectation. Specifically, the expected F&P performance for kindergarten was recommended to be up-adjusted, from Level C to Level D, to better align expected performance longitudinally.

MAP and PARCC performances were highly and positively correlated. MAP and PARCC scores were analyzed for both reading and math by grade level for students who had both scores. Visual inspection of the scatterplots of the scores indicates a linear relationship between the MAP and PARCC scores. The fall and spring MAP RIT scores were strongly and positively correlated to each other for both reading and math ($r \geq .86$), as well as with PARCC scale scores in each respective content area, with the spring MAP slightly more strongly correlated than the fall MAP with PARCC scores ($r > .80$ for reading in Grades 3–8; $r \geq .90$ for math in Grades 3–6). Grades 7 and 8 are excluded from the math analyses because these students may not take MAP math depending on their math courses.

A scale score of 750 (Level 4: Met Expectations) on the PARCC assessments has been determined as indicative of college- and career-ready performance (PARCC Inc., 2016). MAP scores were used to predict college-ready performance on the PARCC by regressing PARCC scale scores on MAP RIT scores to determine how the HCPSS preliminary performance benchmarks align with the PARCC indicator of college readiness (Equation 2):

$$PARCC\ Scale\ Score = \beta_0 + \beta_1(MAP\ RIT) + \varepsilon \qquad (2)$$

Table 2 displays the results of the regression analyses calculated for the MAP RIT score that would result in a PARCC scale score of 750. About 95% of the students who scored between the lower and upper estimates are predicted to perform at college and career levels as measured by the PARCC (i.e., Level 4 or higher).

Table 2

MAP RIT Scores Associated with Scoring a 750 on the PARCC

	Math	Reading

Grade	Lower Bound	MAP RIT	Upper Bound	Lower Bound	MAP RIT	Upper Bound
Fall						
G3	190.9	197.6	204.5	189.7	198.2	207.1
G4	204.1	210.8	217.7	196.4	205.2	214.5
G5	216.8	224.0	231.5	203.9	213.5	223.4
G6	221.8	228.7	235.9	212.0	221.5	231.4
G7	226.0	237.3	249.2	213.2	223.3	234.0
G8	226.4	241.7	258.1	215.9	226.4	237.5
Winter						
G3	195.4	202.3	209.5	193.3	202.1	211.2
G4	208.3	215.1	222.1	199.0	208.1	217.6
G5	219.7	227.1	234.8	205.9	215.4	225.4
G6	219.8	234.2	249.5	209.4	226.9	245.9
G7	222.7	238.5	255.6	209.2	227.8	248.1
G8	225.0	241.8	259.9	213.0	231.4	251.3
Spring						
G3	202.2	208.3	214.5	196.9	205.1	213.8
G4	215.7	221.7	227.8	201.8	210.3	219.2
G5	224.8	231.4	238.2	207.6	216.5	225.7
G6	227.1	233.5	240.1	215.1	224.1	233.5
G7	230.3	241.3	252.8	215.0	225.0	235.5
G8	229.8	244.5	260.2	217.6	228.0	239.0

Note. Analyses only include students with both scores. Upper and lower bounds are computed from the 95% confidence interval for β_1.

Results were shared with the Division of Curriculum, Instruction, and Administration staff and reviewed against the existing HCPSS preliminary performance benchmarks for alignment purposes. Preliminary comparisons suggest that the MAP benchmarks shared are in line with PARCC indicators for on-track progress toward college and career readiness. Findings suggest that MAP scores correlate with major data points we use in the HCPSS to measure academic achievement—MAP scores serve as an additional data point to reflect student progress toward being college-ready in both reading and math. MAP items themselves align to objectives in the Maryland College and Career Ready Standards, which are being taught throughout HCPSS. Thus, both local empirical findings and MAP instructional planning resources exist to support the use of students' MAP results in instructional planning as teachers prepare students to be ready for college.

With empirical support for the validity of MAP scores in benchmarking students' college readiness at a programmatic level, MAP may also be useful as an indicator of program effectiveness in increasing student achievement. We will explore whether MAP spring achievement and/or MAP fall-to-spring RIT growth may be sensitive to changes that directly result from a second year of full implementation of the ESM. The amount of growth at the school level after Year 2 of ESM implementation compared to that observed after one year of ESM implementation will also be examined. We expect that ESM schools will demonstrate significantly higher growth in RIT mean scores than comparison schools because of the professional learning and supports provided to ESM teachers and students. We will also explore how MAP may benchmark program effects for the ESM initiative after its second year of implementation. As part of these analyses, we will need to explore root causes for any progress or lack thereof in MAP performance between ESM and comparison schools. Any information gleaned will continue to be shared with relevant stakeholders to maximize collaborative opportunities and support alignment in practices throughout the school system.

Translation into Practice: Aligning Accountability

As noted earlier, the Division of Accountability collaborates with staff from the curricular and administration offices to provide ongoing professional learning to teachers and administrators on how to connect MAP results with instructional planning at both the classroom and school improvement levels. This collaboration has been essential in facilitating the use of a data protocol to provide a common language throughout HCPSS in conducting data conversations for continuous improvement.

To support staff throughout the school system in using the rich data sources to answer questions specific to different levels of action, the SDP Fellow engages with central office and school staff in conducting school- and classroom-level data discussions that allow stakeholders to see their students beyond the numbers. With different audiences, the same data require different analytical angles to make the results align conceptually with diverse staff roles. For example, central office leaders might look for system-wide patterns to identify county-level strategic and programmatic needs; principals may focus on school trends in relation to county-wide performance to set school improvement targets; teachers may hone in on specific classes or students to look for potential skill gaps in planning their daily lessons. Ultimately, even though the resulting actions may differ, all stakeholders are using available data to prepare students to be ready for college or a career and college when they graduate.

To develop a culture of consistent utilization of multiple measures for informing school improvement planning and classroom instruction, the SDP Fellows and HCPSS colleagues developed a data protocol to guide data conversations, from (a) assessing the gap between the current and desired states; (b) identifying root causes within the system's control that explain the observed gaps; (c) planning strategies that target those root causes; (d) acting and monitoring actions, and (e) back to assessing progress using newly collected formative data. We first piloted the data protocol steps with central office leadership to examine county-level performance, and then asked school administrators for feedback on the process, prior to introducing the data protocol to school improvement teams to guide their planning work.

In supporting the translation of data into action, we found that visualizing data in different ways helped stakeholders understand the information. For example, Figure 9 made a lasting impression on them. Previously, tabulations of data were the main method by which data were shared with schools and central office leaders. When the data were presented visually as in Figure 9 and supplemented with a verbal explanation of the information in the data, we noticed that stakeholders became more engaged with what the numbers were communicating. In another case, we are able to demonstrate how data may be analyzed at a group level to illustrate group-level patterns (e.g., to inform school-level planning) before drilling down to the teacher or individual student (e.g., to inform classroom-level planning). In turn, stakeholders also learned how the same set of data could inform different actions needed to address root causes at the various levels of analyses. This drill-down exercise also helped reduce the impersonal feeling around data and numbers to make data more personal to the students we, as educators, serve.

Next Steps

At the time of writing, SY2015–2016 EOY PARCC results have just become available. We plan to apply the regression equation to examine how well MAP performance predicted PARCC scores using the estimated model. Regardless of whether the prediction model yields accurate results, we must be mindful of using such predictions at an individual level. Predicted scores may contribute to guiding group-level practices—for instance, in providing guidance to school teams in setting appropriate school-level targets. We also plan to replicate the visualization of MAP against F&P text levels to examine whether similar patterns are observed for the SY2015–2016 F&P year-end performance and spring MAP scores data.

Lessons Learned

Local norms and prior achievement must be examined in the context of national norms when considering benchmarks for expected achievement and growth. For example, Figure 11 displays the distribution of each agency's Grade 1–6 MAP math winter 2016 percentile ranks. The difference in the distributions across agencies makes the contextualization of the local landscape important in order to understand how each district must serve its populations, and how nationally normed data can only serve as a partial guide to evaluating how successfully a district is meeting the needs of its students.

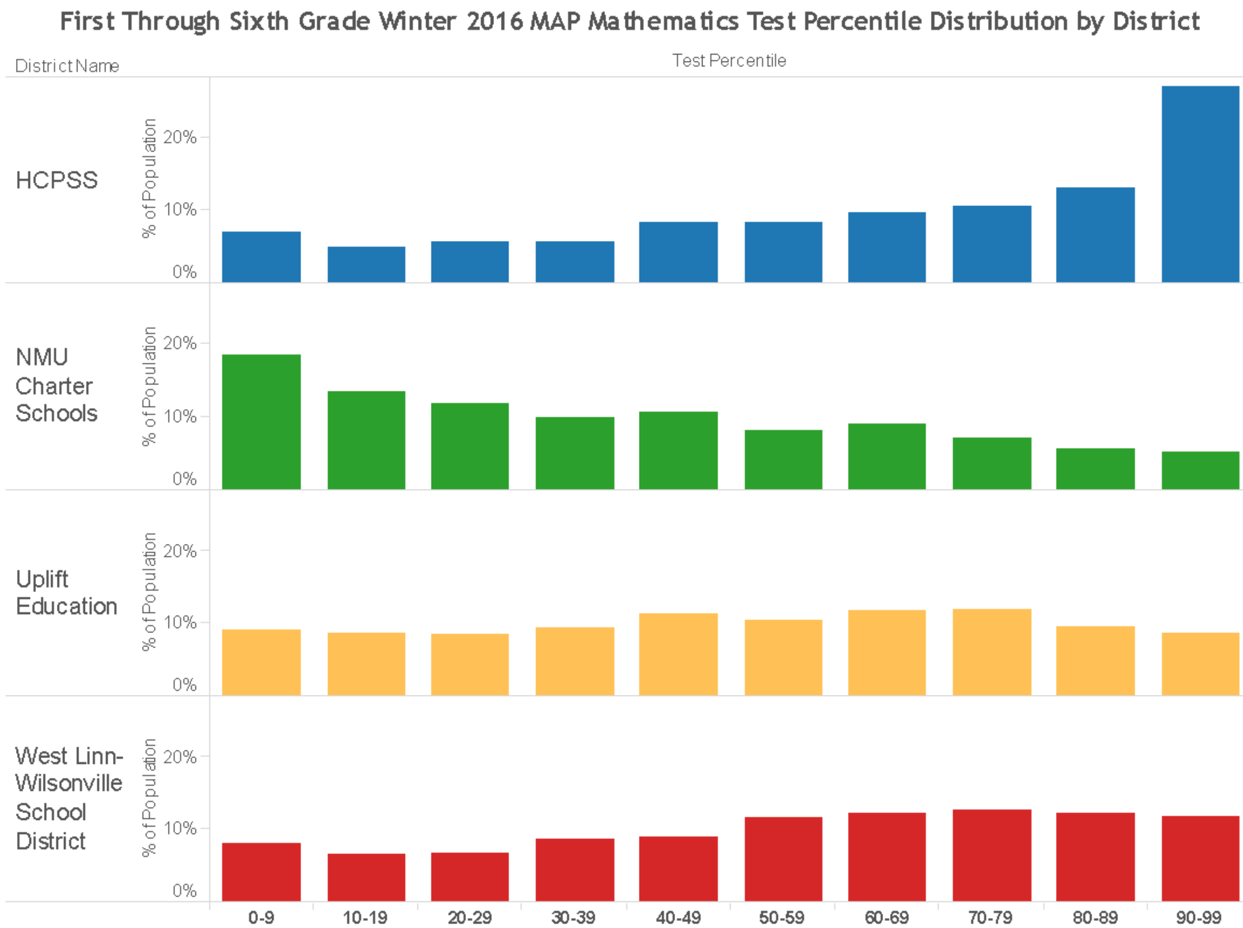


Figure 11. Grades 1–6 winter MAP mathematics decile distribution by district for SY2015–2016.

According to the NWEA's nationally-normed scores, when the population of scores is categorized by deciles, each decile (i.e., each vertical bar in the chart above) should be approximately the same height. Put another way, if the sample district performance reflected nationally-normed performance, one would expect the same number of test takers to fall within each decile. As seen in Figure 11, the agencies all found that MAP score distributions did not align with expected national norms. Thus, setting targets for MAP achievement levels may differ for local education agencies, depending on each district's current performance levels.

In another example, when using MAP to measure student growth, setting expectations that all students meet normed growth can create unrealistic expectations. NWEA set expected growth based on their national norming study, such that the 50th percentile reflects the growth observed for 50% of their norming sample. In other words, the MAP growth norms were developed in such a way that NWEA expects 40% to 60% of students would meet typical growth. In addition, by virtue of regression to the mean, a student who starts the year at the 30th percentile is likely to make greater growth on the MAP RIT scale than a peer in the same grade level who begins at the 85th percentile. Thus, proposing the same RIT growth target for all students—or proposing that more than 60% of students would make typical growth—may create unattainable benchmarks. Instead, individual students and individual schools need to consider their current starting points in the context of aspirational benchmarks in order to make realistic yet challenging short- and long-term targets for growth.

Alignment of standards and results from different data sources is crucial in fostering a culture of consistent data use that supports the shared mission of preparing students to be college- and career-ready. Agencies found strong, positive correlations between students' performance on the MAP and other assessments, which contributed to staff recognition of making use of the MAP for formative instructional purposes. In the case of HCPSS, the visualized positive correlation between the MAP and a heavily-used reading assessment not only allows educators to see concretely how the two measures are aligned, but also pushes for adjusting the existing reading benchmark so that it better aligns with college-ready expectations. In the case of Uplift Education, using predictions as part of the assessment system allows the district to triangulate multiple assessments in order to inform instructional decision. Still, caution is warranted in using NWEA MAP predictions at the student level, as student-level predictions, particularly for high- and low-scoring students, tend to lack the accuracy of a school-level prediction. Such statistical analyses may be more appropriate for making decisions at a group or programmatic level.

In using assessment results, support for data literacy of principals and teachers is key. In addition to having multiple data sources to review, test results such as those generated by the MAP assessments may not be easy to digest; understanding the data and making the information actionable is a role the district must play. Any analysis must be presented as instructionally relevant, and doing so in innovative ways may increase data use. For example, using visualizations to contextualize the data may help teachers see the relevance of one data point in relation to other data points—assessment results are more powerful when viewed in conjunction with other measures. To this end, these agencies have found that the clarity of data presentation is most effective when a teacher does not have to perform additional analyses on their own to make the information instructionally relevant. Likewise, when teachers do need to perform their own analyses, the interactive feature of the data is built-in; analyses are straightforward and the students behind the numbers immediately jump out. The various figures presented in this report are examples of visualizations that can be used to present such data.

Helping teachers and administrators craft questions they want to answer is another way to connect data to teaching and learning. School and central office staff alike have questions they want to answer. For example, which students may benefit from more rigorous academic coursework? Which students may just need an extra push to meet college readiness standards? Which groups continue to under-perform and what can be changed about instruction to meet students' needs? To answer these action-oriented questions using a data-driven process, data coaching supports and a data protocol to guide data discussions are essential. These types of resources increase data literacy and allow meaningful interpretations of information from multiple data points at all levels in the system.

In summary, policy makers and school leaders can be supported to make decisions about curriculum, instruction, and programs by triangulating evidence of progress pulled from multiple data sources. Based on the actions taken by the districts in this report, districts should consider multiple strategies to realize the stated outcome above. Validating MAP as a predictor for success on a national- or state-level college and career readiness assessment (PARCC, SBAC, STAAR) successfully created this triangulation for HCPSS, Uplift, and West-Linn Wilsonville. Examining how results aligned with formative assessments used to make daily instructional decisions (DRA2, F&P) made MAP easier to contextualize for elementary educators at HCPSS.

All of the districts have, in one way or another, compared district and national achievement norms in order to set meaningful goals with principals. All of the districts have also engaged a combination of students, teachers, principals, and district leaders in tracking individual, classroom, grade-level, and school growth on the MAP assessment over time. All of have cut their data to look at levels of achievement by examining status and growth of quintiles and deciles across grades, schools, and districts. With a combination of these strategies, districts using the MAP assessment can help inform stakeholders about progress and trends that give way to action planning and continuous improvement.

References

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London, England: King's College.

Blank, R. K., & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement: How meta-analysis provides scientific evidence useful to education leaders*. Washington, DC: Council of Chief State School Officers.

Boudett, K. P., City, E. A., & Murnane, R. J. (2013). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard University Press.

Bruce, M., Bridgeland, J. M., Fox, J. H., & Balfanz, R. (2011). *On track for success: The use of early warning indicator and intervention systems to build a grad nation*. Washington, DC: Civic Enterprises. Retrieved from http://new.every1graduates.org/wp-content/uploads/2012/03/on_track_for_success.pdf

Cliffordson, C. (2008). Differential prediction of study success across academic programs in the Swedish context: The validity of grades and tests as selection instruments for higher education. *Educational Assessment*, *13*(1), 56–75.

Crane, E. (2010). *Building an interim assessment system: A workbook for school districts*. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Resources/Publications/Building_an_Interim_Assessment_System_A_Workbook_for_School_Districts.html

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G. (2007). *The proficiency illusion*. Washington, DC: Thomas B. Fordham Institute. Retrieved from http://eric.ed.gov/?id=ED498573

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). *Accountability for college and career readiness: Developing a new paradigm*. Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from https://edpolicy.stanford.edu/sites/default/files/publications/accountability-college-and-career-readiness-developing-new-paradigm.pdf

Elementary and Secondary Education Act of 1965, 20 U.S.C. § 6301 et seq. (2012).

Every Student Succeeds Act of 2015, 20 U.S.C. § 1177 et seq. Retrieved from Retrieved from http://www.ed.gov/essa?src=rn

Goals 2000: Educate America Act of 1994, 20 U.S.C. § 5801 et seq. (2012). Retrieved from http://www2.ed.gov/legislation/GOALS2000/TheAct/index.html

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Howard County Public School System. (2016). *Howard County Public School System profile 2015–16*. Retrieved August 12, 2016 from http://www.hcpss.org/f/aboutus/profile.pdf

Huebner, T. A. (2009). What research says about…/ Balanced assessment. *Educational Leadership, 67*(3), 85–86.

Jensen, N. (2013, May 29). Interpreting the percentage of students meeting or exceeding their growth projections [Blog post]. Retrieved from https://www.nwea.org/blog/2013/interpreting-percentage-students-meeting-exceeding-growth-projections/

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R. C., Staiger, D. O. (2016). *Educators' perspectives on Common Core implementation*. Cambridge, MA: Center for Educational Policy Research. Retrieved from http://cepr.harvard.edu/files/cepr/files/teaching-higher-report.pdf

Loveless, T. (2015). *The 2015 Brown Center report on American education: How well are American students learning?* Washington, DC: Brookings Institution. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2015/03/bcr/2015-brown-center-report_final.pdf

Massell, D. (1994). Three challenges for national content standards. *Education and Urban Society, 26*, 185–195.

Mizell, H. (2010). *Why professional development matters*. Oxford, OH: Learning Forward.

National Association of Charter School Authorizers. (2015). *Principles & standards for quality charter school authorizing*. Chicago, IL: Author. Retrieved from http://www.qualitycharters.org/wp-content/uploads/2015/08/Principles-and-Standards_2015-Edition.pdf

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Retrieved from http://www2.ed.gov/pubs/NatAtRisk/index.html

National Council on Education Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the secretary of education, the National Education Goals Panel, and the American people*. Washington, DC: Author.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 (2002). Retrieved from http://www2.ed.gov/nclb/landing.jhtml

Northwest Evaluation Association. (2013). *Texas linking study: A study of the alignment of the NWEA RIT scale with the State of Texas Assessments of Academic Readiness (STAAR).* Portland, OR: Author. Retrieved from https://www.nwea.org/content/uploads/2013/12/Texas_Linking_Study2013.pdf

Northwest Evaluation Association. (2014). *Proficiency guidance from NWEA on new state summative assessments.* Portland, OR: Author. Retrieved from www.nwea.org/resources/proficiency-guidance-on-new-state-summative-assessments-fromnwea

Northwest Evaluation Association. (2015a). *Smarter Balanced preliminary performance levels*. Portland, OR: Author. Retrieved from www.nwea.org/content/uploads/2015/01/SBAC-Preliminary-Cut-Scores-MAY15.pdf

Northwest Evaluation Association. (2015b). *Virtual Comparison Group reporting: Customized growth analysis for district and school leaders*. Portland, OR: Author. Retrieved from https://www.nwea.org/content/uploads/2013/06/Virtual-Comparison-Group-One-Sheet-JUL15.pdf

Northwest Evaluation Association. (2017). *Conditional growth index: Definition and calculation.* Portland, OR: Author. Retrieved from https://community.nwea.org/docs/DOC-1642

Olson, L. (1998, April 15). An 'A' or a 'D': State rankings differ widely. *Education Week*. Retrieved from http://www.edweek.org

PARCC, Inc. (2016). *2014-2015 tables of cross-state and state-specific PARCC results*. Retrieved from http://www.parcconline.org/images/Consortium_and_State_Tables_FINAL_3_7_16.pdf

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system*. Aspen, CO: Aspen Institute. Retrieved from http://www.achieve.org/files/TheRoleofInterimAssessments.pdf

Pinkus, L. (2008). *Using early-warning data to improve graduation rates: Closing cracks in the education system*. Washington, DC: Alliance for Excellent Education. Retrieved from http://ndpc-web.clemson.edu/sites/default/files/uploads/par_advantage/EWI.pdf

RIT (Rausch UnIT) scale. In *Mount Pleasant community school district maps 101 home page* (Wiki). Retrieved May 16, 2016, from https://mpcsdmaps.wikispaces.com/file/view/NWEA+MAPS+RIT+Defined.doc

Rodriguez, M. C., Nickodem, K., Palma, J., & Stanke, L. (2016). *Analyzing and reporting achievement gaps: Guidance for Minnesota schools.* Minneapolis, MN: University of Minnesota.

Smith, M. S., Fuhrman, S. H., & O'Day, J. (1994). National curriculum standards: Are they desirable and feasible? In R. F. Elmore & S. H. Fuhrman (Eds.), *The governance of curriculum: 1994 yearbook of the Association for Supervision and Curriculum Development* (pp. 12–29). Alexandria, VA: Association for Supervision and Curriculum Development.

Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*, 324–328.

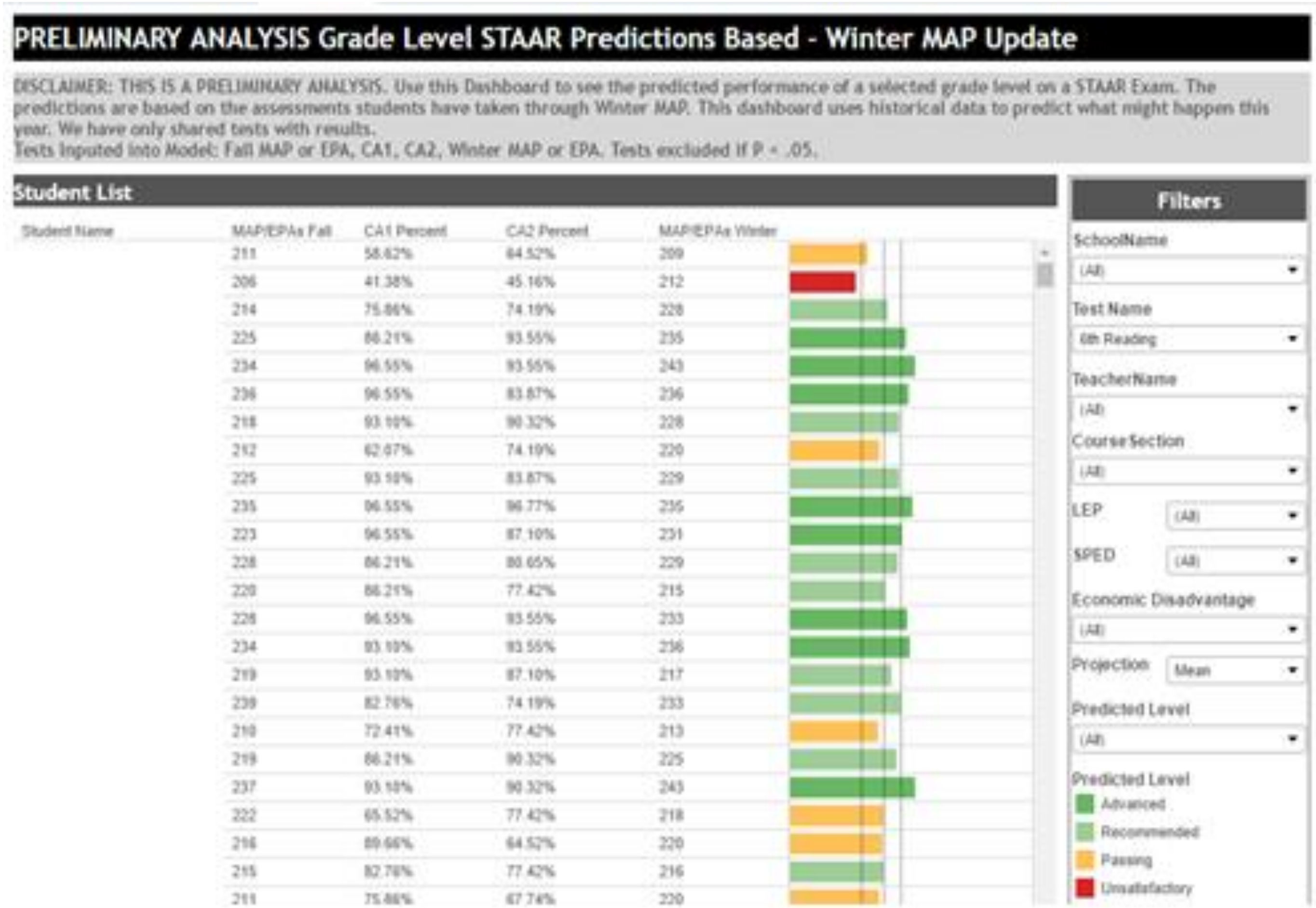
Texas Education Agency. (2009). *Texas college and career readiness standards*. Retrieved June 10, 2016 from http://tea.texas.gov/curriculum/teks/

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. Portland, OR: Northwest Evaluation Association.

Wixson, K. K., Dutro, E., & Athan, R. G. (2003). The challenge of developing content standards. In R. E. Floden, (Ed.), *Review of research in education* (pp. 69–107). Washington, DC: American Educational Research Association. Retrieved from http://www.colorado.edu/education/sites/default/files/attached-files/Wixson%20et%20al_The%20Challenge%20of%20Developing.pdf

Appendix:

Uplift Education Grade 6 STAAR Predictions Dashboard, Student List



PRELIMINARY ANALYSIS Grade Level STAAR Predictions Based - Winter MAP Update

DISCLAIMER: THIS IS A PRELIMINARY ANALYSIS. Use this Dashboard to see the predicted performance of a selected grade level on a STAAR Exam. The predictions are based on the assessments students have taken through Winter MAP. This dashboard uses historical data to predict what might happen this year. We have only shared tests with results.
Tests Inputed into Model: Fall MAP or EPA, CA1, CA2, Winter MAP or EPA. Tests excluded if P < .05.

Student List

Student Name	MAP/EPAs Fall	CA1 Percent	CA2 Percent	MAP/EPAs Winter
	211	58.62%	64.52%	209
	206	41.38%	45.16%	212
	214	75.86%	74.19%	228
	225	86.21%	93.55%	235
	234	96.55%	93.55%	243
	236	96.55%	83.87%	236
	218	93.10%	90.32%	228
	212	62.07%	74.19%	220
	225	93.10%	83.87%	229
	235	96.55%	96.77%	236
	223	96.55%	87.10%	231
	228	86.21%	80.65%	229
	220	86.21%	77.42%	215
	228	96.55%	93.55%	233
	234	93.10%	93.55%	236
	219	93.10%	87.10%	217
	239	82.76%	74.19%	233
	210	72.41%	77.42%	213
	219	86.21%	90.32%	225
	237	93.10%	90.32%	243
	222	65.52%	77.42%	218
	216	89.66%	64.52%	220
	215	82.76%	77.42%	216
	211	75.86%	67.74%	220