Benchmarks and Norms Interpretation and Use Guidelines

Version 4.0

July 2019

150 South Fifth Street
Suite 600
Minneapolis, MN 55402

612.254.2534
www.fastbridge.org

FastBridge Learning

THIS WAY TO FASTER RESULTS.

Formative Assessment System for Teachers™: Benchmarks and Norms Interpretation and Use Guidelines (version 4).

Citation:

FastBridge Learning. (2019). *FastBridge Learning: Benchmarks and Norms Interpretation and Use Guidelines (version 4).* Minneapolis, MN: Author.

150 South Fifth Street, Suite 600
Minneapolis, MN 55402
Email: help@fastbridge.org
Website: www.fastbridge.org
Phone: 612-254-2534

Table of Contents

Standard Setting

Standards are the content or skills that are expected of students (i.e., *content standards*), which are often defined by a score for purposes of measurement (i.e., *performance standards*). A number of terms are used to reference performance standards, including benchmarks, cut scores, performance levels such as basic, proficient and advanced, instructional or mastery levels, and thresholds. These terms each reference categories of performance with respect to standards.

Standards are often classified as either criterion-referenced or norm-referenced. Criterion-referenced standards are established using criteria external to a specific assessment. For example, states often use recommendations from expert educators as to what constitutes an acceptable level of performance in a particular subject and grade. Expert educators are asked to consider what proficient students know and can do. Then, each expert applies expectations of what constitutes proficiency to a set of items on a test and indicate which items the proficient student should get correct. From that the *test score* is determined such that students at or above the score can be considered proficient. These scores are typically obtained from a group of experts and then averaged. Most states define three of four levels of proficiency from demonstrating basic understanding of the content to an advanced understanding. Because this process involves human judgment it is common for states to evaluate proficiency levels relative to past performance and national norms. This step often results in adjustments to performance level cut scores. In practice, criterion-reference standards are often referred to as *benchmarks*.

Another form of *criterion-referenced* standard commonly used in educational settings is mastery. An example of a mastery criterion is that students can accurately identify all of the lower- and upper-case letters in the alphabet. This type of standard does not require human judgment and can be precisely articulated when the learning expectations represent a relatively small finite space.

Norm-referenced standards are expectations that are tied directly to normative performance levels. As noted by Sireci (2005), “scores are interpreted with respect to being better or worse than others, rather than with respect to the level of competence of a specific test taker” (p. 118). For many years in educational assessment, the *Grade Equivalent* was the most common example of a norm-referenced standard. A student with a grade equivalent equal to her enrolled grade was considered to be performing on-grade level. That value is equivalent to the 50th percentile in the enrolled grade.

Grade equivalents have largely been replaced by percentiles as the preferred way of reporting norm-referenced expectations. In education percentiles are almost always anchored to grade

level. A percentile indicates the percent of students in a population or group that scored at or below that level. Thus, the 40th percentile means that 40% of the population scored at or below that level.

Percentiles have several advantages. First, they are easy to explain and nearly universally recognized by educators and parents. Second, their interpretation remains the same across tests and grades, unlike scaled and raw scores. For example, a scaled score of 200 may be very high in one grade, but average in a higher grade; whereas, the 85th percentile means above average in every grade. Third, they describe the entire range of performance from very low (e.g., 1st percentile) to very high (e.g., 99th percentile). The main disadvantage of percentiles relative to raw or scaled scores is that their properties make them less suitable to statistical analysis and comparisons.

FAST Norms & Benchmarks

Norms

FAST reports national and local (class, school, and district level) percentiles by grade and season for each FAST assessment (see FAST National Norms Technical Report for details on the development of the norms). Percentiles range from 1 to 99 and are color-coded in the system to provide a quick visual representation of four levels of performance:

- 1 – 19th percentile (red)
- 20th – 29th percentile (yellow)
- 30th – 84th percentile (green)
- and 85th – 99th percentile (light blue).

These ranges were established to support early intervention and prevention within multi-tiered systems of support (MTSS; Brown-Chidsey & Bickford, 2016). MTSS research indicates that most schools can provide supplemental and intensive supports for 20% to 30% of their students and accelerated learning opportunities for about 15% of their students (Christ, 2008; Christ & Arañas, 2014). Schools rarely have resources to provide supplemental and intensive supports support for more than 30% of learners at-risk for low achievement, even if a larger proportion would benefit. The range of scores between the 30th and 84th percentile comprises more than one half of the national norm population and represents average or typical performance.

National norms are often used in combination with criterion-referenced benchmarks to guide school- and district-level decisions about instruction, curriculum, and system-wide services FAST also provides criterion-referenced cut scores that serve as benchmarks to classify

students as *high risk*, *some risk*, or *low risk* of not meeting end of year performance goals. These categories were derived from internal research and reflect common practices and recommendations from independent educational research and policy agencies such as the RTI Network and the National Center on Intensive Intervention (NCII). Here is a quote from the RTI Network.

> *"Reading screens attempt to predict which students will score poorly on a future reading test (i.e., the criterion measure). Some schools use norm-referenced test scores for their criterion measure, defining poor reading by a score corresponding to a specific percentile (e.g., below the 10th, 15th, 25th, or 40th percentile). Others define poor reading according to a predetermined standard (e.g., scoring below "basic") on the state's proficiency test. The important point is that satisfactory and unsatisfactory reading outcomes are dichotomous (defined by a cut-point on a reading test given later in the students' career). Where this cut-point is set (e.g., the 10th or 40th percentile) and the specific criterion reading test used to define reading failure (e.g., a state test or SAT 10) greatly affects which students a screen seeks to identify"* (RTI Network, 2015*)*

FAST research using CBMreading and a validated norm-referenced diagnostic reading test, the Group Reading and Diagnostic Evaluation (GRADE) test consistently demonstrated that the 15th and 40th percentiles serve as reliable cut scores for identifying levels of risk. Scores below the 15th percentile are classified as "high-risk." Those at or above the 15th and below the 40th percentile are classified as "some-risk;" and those at or above the 40th percentile are "low-risk." These levels are also consistent with findings from other CBM research.

Benchmarks

FAST Benchmarks are test-specific scores that indicate the student's risk of performing below a future (usually end of year) performance target. FAST defines two (or three) benchmark cut scores for each assessment in each season and grade, resulting in three (or four) levels of risk. The FAST default Benchmark settings are based on the national norms and correspond to the following percentile ranges.

- High-Risk: Below the 15th percentile
- Some-Risk: 15th – 39th percentile
- Low-Risk: 40th – 99th percentile

For aReading, aMath, and CBMreading the low risk range is divided into two levels.

- Low-Risk: 40th – 70th percentile
- Advanced: 71st – 99th percentile

The cut score at the 70th percentile represents the point above which students are likely to be on track for success on college and career readiness standards. Because these assessments measure the full spectrum of standards-aligned skills and have consistently shown to be strong predictors of future success, they can validly support the college and career readiness benchmark. See FAST Norms & Benchmarks for additional detail and Accessing FAST Benchmarks & Norms to see how to access the FAST benchmark and norms tables.

The FAST risk indicators (e.g., high-risk, some-risk, and low-risk) have been successfully used in school systems to identify a student's level of instructional need as part of MTSS data-based decision making. Students with scores in the *high-risk* category are performing well-below grade level expectations. Often these students have not received high quality instruction or lacked educational opportunities necessary for success. Without intensive intervention above and beyond core instruction, these students are likely to fall further behind. Consistent with usual practice, FAST recommends students with scores coded as high-risk receive intensive intervention and be placed on a weekly progress monitoring schedule. Because students often experience high growth when provided with a scientifically research-based intervention, an aggressive growth target should be considered. The FAST system provides guidance as to which growth rate is appropriately aggressive for each FAST assessment. An aggressive growth target is both necessary and justified to increase the likelihood that the student will reach grade-level performance. Depending on the severity of the performance deficit this goal may take more than one year of intensive support to achieve.

Prior research shows that students with scores in the *some-risk* category typically improve and get back on track with supplemental instruction or support. FAST recommends that students with scores coded as *some-risk* receive supplemental instruction and be placed on a weekly or bimonthly progress monitoring schedule with a performance target set high enough to ensure the student reaches the low-risk level by spring.

Students with scores in the low-risk or advanced level categories should be remain on track with high quality core instruction. Even so, careful attention should be given to students with scores in the low end of the range as these students are more likely to fall behind, which is why FAST recommends triannual screening of all students.

Table 1 lists the FAST assessments recommended for triannual screening by grade and domain.

Table 1. FAST recommended screening measures by grade and subject

Grade	Math	Reading
K	earlyMath composite	earlyReading composite
1	earlyMath composite	earlyReading composite
2 – 3	aMath	aReading & CBMreading
4 - 12	aMath	aReading & AUTOreading

Decision Accuracy with Benchmarks

Benchmarks and the associated risk categories simplify the process of evaluating the appropriate level of instructional support, as well as the impact of that support at the student classroom, school, and district levels. However, this simplification can lead to misinterpretation. As with any classification method using cut scores, it is important to consider measurement error at the cut score as well as the distance from the cut scores. For instance, a student just below the cut score on one occasion is likely to be just above the cut score if retested; whereas a student scoring well-below the cut score is not. The uncertainty of classification is even greater when the categories are based on predicting future outcomes. A method known as receiver operator characteristic (ROC) illustrates the amount of uncertainty in decisions based on performance categories. The ROC approach is described below (Swets, et.al., 2000).

Figure 1 illustrates the four possible ROC classification outcomes: true positive (TP; Zone 3), true negative (TN; Zone 2), false positive (FP; Zone 1), and false negative (FN; Zone 4). For illustration purposes, assume that the results are based on using the CBMreading Advanced cut score to predict who will score below the proficient level on a state reading test. In this illustration, CBMreading and state state test scores are represented on a z-score scale. This scale has a mean of 0 and a standard deviation of 1.0. Thus, scores above 0 are above the group average. The points on the

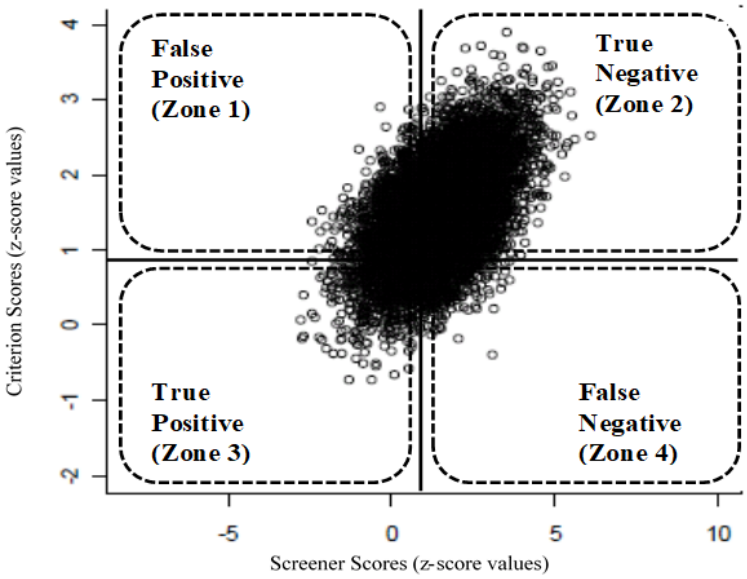

Figure 1

graph represent student scores, with CBMreading scores on the horizontal axis and state scores on the vertical axis.

Zone 3 shows all of the scores correctly classified by CBMreading as students who scored below the Advanced level and below proficient on the state test. Zone 2 represents the scores in the Advanced level on CBMreading and at or above proficient on the state test. Zones 2 and 3 represent correct classifications.

Zones 1 and 4 represent classification errors, or disagreements in classification between CBMreading and the state test. Scores in Zone 1 are students below Advanced but at or above proficient on the state test. They are false positives because CBMreading falsely indicated that the student would not achieve proficiency on the state test. Scores in Zone 4 are students in the Advanced level who scored below proficient on the state test. These are false negatives because CBMreading falsely indicated that the student would meet proficiency on the state test.

The overall accuracy of decisions (True Positive + True Negative) is primarily determined by the strength of the correlation between the predictor and criterion, and to a lesser degree on the position of the predictor cut score. Researchers use statistical models to identify cut scores on the predictor that minimize decision errors. The placement of the cut score can be adjusted to maximize accuracy or minimize decision errors. However, improving one index, such as true positives necessarily worsens another index such as false positives. Thus, researchers often use cut scores that balance accuracy and decision errors using metrics called sensitivity and specificity. These, along with secondary metrics such as positive predictive value and negative predictor value provide descriptions of findings. Here are examples for CBMreading and a state assessment.

> ***Sensitivity (true-positive proportion, TPP)*** represents the percent of all students scoring below proficient on the criterion who also score below the Advanced on CBMreading
>
> ***Specificity (true-negative proportion, TNP)*** represents the percent of all students scoring at or above proficient on the criterion who also scored at or above the Advanced cut score on CBMreading
>
> ***Positive predictive value (PPV)*** represents the percent of all students scoring below Advanced on CBMreading who also scored below proficiency on the state test.
>
> ***Negative predictive value (NPV)*** represents the percent of all students scoring at or above Advanced on CBMreading who also scored at or above the proficiency on the state test.

Cut Scores

A cut score (also called decision threshold) is established to maximize the benefits of a correct decision relative to an incorrect decision (Swets, Dawes, & Monahan, 2000). The cut score can be adjusted to balance decision criteria like true positives and false positives, to minimize under classification, or minimize over classification. Which approach to use depends on the costs of classification errors. In many educational applications, the cost of under identifying (that is, not identifying students who actually need help) is considered higher than over identifying students. As such, screening measures for reading often over-identify students by increasing the rate of true positives as well as false positives.

Thresholds that are more lenient (over-identify) increase sensitivity, thereby increasing the proportion of positive classifications (both TP and FP). Thresholds that are stricter (under-identify) increase specificity, thereby increasing the proportion of negative classifications (both TN and FN; Swets et al., 2000). The decision threshold is adjusted to obtain the optimal ratio of positive and negative classifications along with that of true and false classifications.

The strength of the classification prediction can be shown using the area under the curve (AUC). The AUC represents the total area under the curve of the plot of the false positive rate by the true positive rate for all possible predictor cut scores as illustrated in Figure 2. AUC graphs with a sharp bend as shown by the solid blue line in the figure result in more area under the curve relative to the diagonal dashed line, and thus are better predicters. AUCs range from 0.50 to 1.0. An AUC is of 0.5 would be depicted by a solid line that falls on the diagonal. Such an AUC indicates prediction is no better than chance and thus the predictor does not provide any useful information.

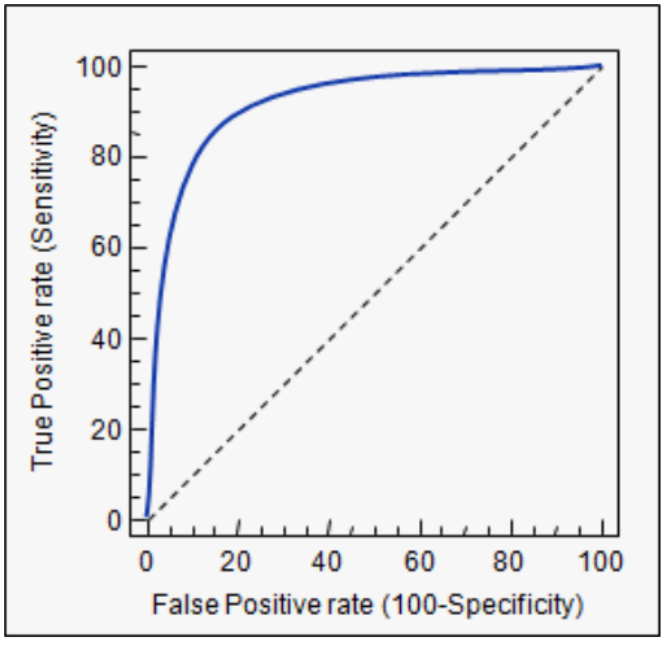

Figure 2

An AUC of 1 perfectly classifies all students into true positive and true negative categories and is almost never achieved. Generally, the criteria applied in the interpretation of AUCs consider values of .90 to 1.0 to be excellent, .80 to .89 good, .70 to .79 fair, and.69 and less poor. It seems reasonable and generally consistent with the standards outlined by the National Center

for Response to Intervention (NCRTI) that an AUC of at least .85 is required for low-stakes decisions (e.g., screening) and that an AUC of at least .90 is required for high-stake decisions (e.g., eligibility and decision making). An index called *Youden's J* identifies the point on the curve furthest from the dashed line. It is often used to select cut points.

Growth Norms

FAST also provides national and local growth norms to assist educators and administrators in interpreting student progress across the year. These norms are provided in the tables alongside the seasonal norms in the FAST *Training & Resources* tab (see *FAST Growth Norms*).

Growth is described in these ways:

- The shift across seasons in the score associated with each percentile
- The distribution of observed growth rates of the overall norm sample
- The distribution of growth rates at each point along the score scale

Growth based on seasonal shift is useful for determining how much growth a student at a given percentile needs to remain at that level throughout the school year. Due to the individual nature of these rates, they are not included in any FAST reports.

The other growth norms, aggregate and student growth percentiles (SGP) are used in the group growth report and used to guide goal setting decisions in the progress monitoring setup. The next section describes these growth norms and their use.

Aggregate Weekly Growth

Aggregate weekly growth norms are percentiles derived from the overall distribution of growth rates for a given assessment and grade. Seasonal growth, known as rates of improvement (ROI) are computed by dividing the overall gain across season by the number of weeks between administrations. For example, if the student has a CBMreading score of 125 in the fall and 150 in the winter, the calculation would be:

(150-125)/17 = 1.47 words per week

Thus, her ROI is 1.47 words per minute (wpm) per week. The aggregate growth rate is recommended for setting progress monitoring goals and evaluating student growth individually and in groups.

FAST researchers defined four growth rate levels which are anchored to the mean ROI growth rate for each FAST progress measure.

- Very Realistic: 80% of the mean rate
- Realistic: 100% of the mean rate (i.e., equal to the mean rate)
- Ambitious: 120% of the mean rate
- Very Ambitious: 150% of the mean rate

The FastBridge system defaults to the *Ambitious* rate for goal setting. This level was selected because research consistently shows that students with high-risk scores who receive intensive, scientifically research-based interventions grow about 20% faster than average.

Student Growth Percentiles

Student growth percentiles (SGPs) were originally developed for schools that use FAST to meet state requirements for reporting educator effectiveness ratings. SGPs are derived from a statistical method called Quantile Regression. This method produces ROI growth distributions for every observed baseline score and converts the distributions to percentiles. It also employs robust smoothing methods to account for sample variations. When used to compare growth rates across classrooms, grades, or schools, this method has an advantage over the *Aggregate* growth norms in that it accounts for ability differences between groups. This can be important because growth rates vary somewhat by ability, as is evident in the SGP table which shows a decreasing trend in ROI as percentile increases.

Group Growth Report

Leader Report Options

FAST provides a report for viewing both individual student and group growth. In this section we describe the elements of the leader (e.g., managers and specialists) Group Growth Report and provide some guidance for using and interpreting results. For more details about the elements in the Group Growth Report see 'Group Growth Report for Managers' in the Knowledge Base. Figure 3 is a snapshot of the Group Growth Report Benchmark View for aReading. This report provides four growth norms options:

- Growth by All: shows national aggregate weekly growth norm percentiles
- Growth by Start Score: shows student growth percentiles (SGP)
- Growth by District: shows district level aggregate weekly growth norm percentiles
- Growth by School from fall to winter: shows school level aggregate weekly growth norm percentiles

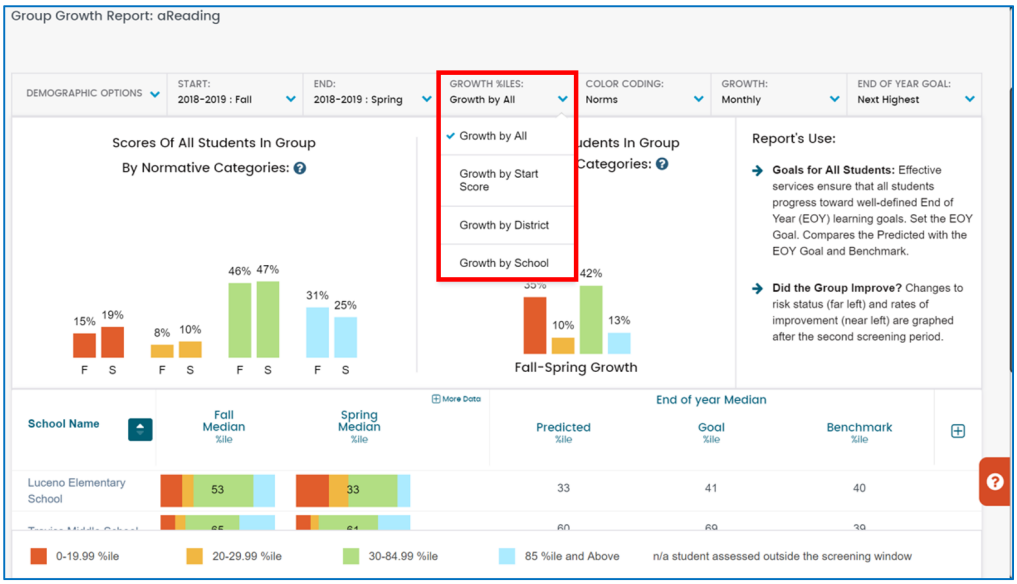

Figure 3

National, district, and school growth percentiles are derived from the distribution of ROIs of the entire (aggregate) sample of students for a specific FAST assessment and grade. Because these three norms use the same methodology, they are all calculated using the Aggregate Weekly Growth method indicated above.

The top section of the report indicates the users' selections. The *DEMOGRAPHIC OPTIONS* dropdown allows users to select demographic subgroups such as gender, race, or IEP status. The next two dropdowns are used to select the start season and end season and school year. Currently, fall is the only option for start season in this report. The *GROWTH %ILES:* dropdown enables selection of the norm group. The user can also select the color-coding scheme which is either the Norms scheme (1st – 19th, 20th – 29th, 30th – 84th, 85th – 99th) or the Benchmarks scheme:

- Flat growth: 1 – 15th percentile
- Modest growth: 16 – 40th percentile
- Typical growth: 41 – 70th percentile
- Aggressive growth: 71st – 99th percentile

These cut-points mirror the national percentile cut-points used for seasonal benchmarks. The growth interval will be weekly or monthly, depending on the assessment. Lastly the user has the option to select the end of year performance target (see Figure 4). The options are:

- By Benchmark (Advanced, Low Risk, or Next Highest)
- End of Year Score
- End of Year Growth Percentile

Figure 4

When the target is set, the report updates the growth rates needed to achieve the target as shown in Figure 4. The **Next Highest** benchmark is the default target because it sets individual goals for each student to reach the next risk or performance level. For example, a student whose fall score was in the high-risk range, would have a goal for a winter score in the some-risk range. This setting is designed to ensure that students' goals are not so high as to be impossible to reach.

Benchmark View (Color Coding), Aggregate Growth Rates

The Benchmark View with aggregate national norms includes a section in the report that summarizes growth of each group (e.g., teacher, grade, school) relative to the national norms. This information is provided by clicking on the plus symbol on the right side of the screen. Figure 5 shows the growth results in the middle section within the red outline for a hypothetical school district with three schools.

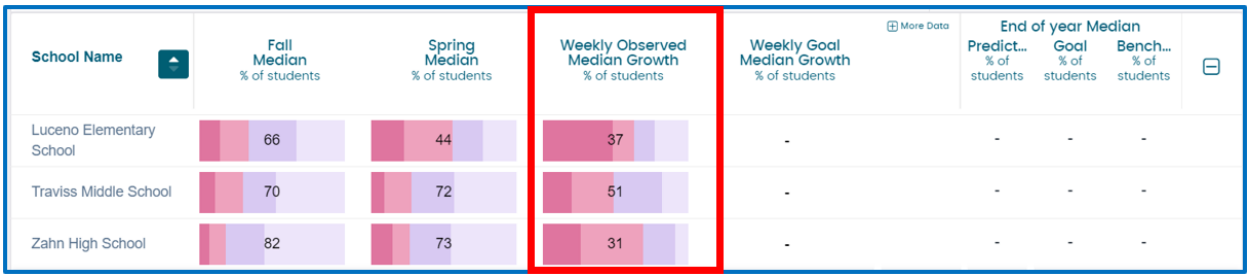

Figure 5

The horizontal stacked bar graph provides a visual representation of the percent of students in each growth category (flat, modest, typical, and aggressive). By clicking the plus symbol next to “More Data” in this section the report adds the percent of students in each of the four growth categories. The number on the graph is the percent of students with typical or modest growth. Values greater than 50 indicate average to above average growth rates.

Norms View (Color Coding), Aggregate Growth Rates

Figure 6 shows the results by norm categories. The information reported in this view is similar to what is reported in the benchmark view with some important differences. First, the performance and growth categories are defined by the following national percentile ranges:

- 1 – 19th percentile (red)
- 20th – 29th percentile (yellow)
- 30th – 84th percentile (green)
- and 85th – 99th percentile (light blue).

Second, the number in each horizontal bar is the **median** percentile. The national median is 50.

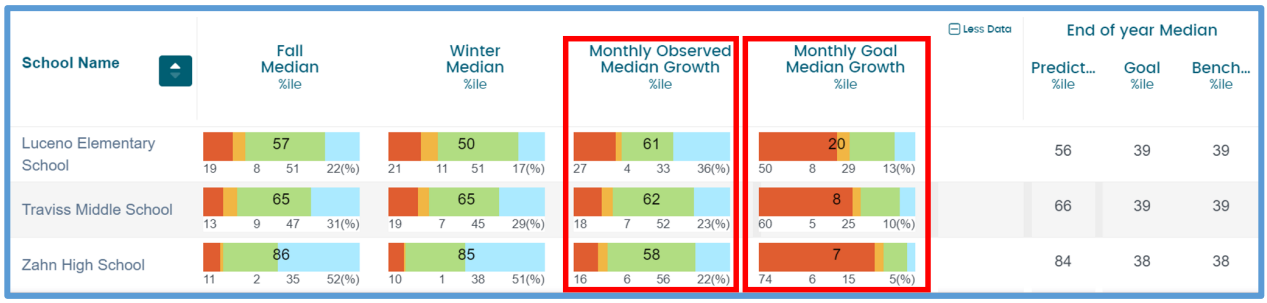

Figure 6

The section labeled “Monthly Observed Median Growth” displays the distribution of growth rates by the four normative categories. The numbers below each horizontal bar indicate the percent of the sample in each category. The larger number printed in the center of each bar indicates the median national growth percentile for that group. For example, the median growth percentile for Zahn High School is 58.

The section labeled “Monthly Goal Median Growth” indicates the amount of growth needed to obtain the end of year benchmark. In this example, the low-risk benchmark is the selected target. Thus, the average growth rate needed in Luceno Elementary for all students to achieve the benchmark is equivalent to the 20th national growth percentile. Because the national average is the 50th percentile, students in Luceno require only modest growth, on average, to achieve benchmark. It is important to note that for this calculation, students with winter scores at

or above the spring target score are assigned a rate of 0.0 because no additional growth is required to achieve the target. Because Zahn High School is high performing, with a median national percentile of 85% in the winter, the overall average growth rate required to meet the target is equivalent to the 7th percentile of national growth norms.

The rightmost section represents predictions of how the group will perform in the spring.

- *Predict %ile* indicates the predicted median end of year national percentile based on prior season results. For example, Zahn High School is predicted to perform at the 84th national percentile on average.
- *Goal %ile* indicates the spring national corresponding to the user defined by end of year goal, which in this example is low risk.
- *Bench %ile* indicates the benchmark percentile associated with low risk.

When the predicted percentile is at or above the goal percentile, the group on average is on track to meet the goal.

Growth by Start Score

The *Growth By Start Score* option provides different information for leaders. This version shows ROI distributions (and growth percentiles) that are derived separately for each fall score. Because growth rates can vary by initial performance, two students with the same ROI but different fall scores will have different growth percentiles.

The primary advantage of using *Growth By Start Score* is that it provides a more equitable comparison across subgroups (e.g., classes, grades, schools) with different overall performance levels. For example, suppose students who are at high risk in the fall tend to have higher growth rates than students who are at low risk in the fall. Other things being equal, low performing schools will have higher growth rates than high performing schools. The *Growth By Start Score* adjusts for these differences, which puts all schools on equal footing relative to national norms.

Growth by Demographic

It is important that all students are provided with the opportunity to make growth during each school year. Use the *Demographic Options* to evaluate growth rates by key demographic characteristics. This report provides access to all the demographic variables that were included in the district's FAST roster uploads.

Frequently Asked Questions

Why growth norms?

Growth rates on FAST measures are more sensitive to instruction than to the demographic composition of a group. Thus, even when the demographic composition of a group differs from the national sample, national growth percentiles still provide a useful context for interpreting progress and instructional effectiveness. The effect of demographic differences across groups can be further reduced by using "Growth by Start Score".

How does our growth compare to national averages?

With the "Norms View" color coding, compare the group's median growth rate in the section labeled "Monthly Observed Median Growth" to the national median (i.e., 50th percentile). Median growth rates near 50 indicate the group grew at about the national average. Values at least 10 points below or above 50 indicate the growth is significantly slower or faster than the national average.

With the "Benchmark View" color coding, compare the percentage of students at or above typical growth (use the number displayed on the bar graph) against the national average of 60%. Values greater than 60 indicate faster than average growth.

To compare to national growth rates among students with a similar overall ability level, use "Growth By Start Score".

How does growth vary across groups and key demographics?

To compare growth rates across schools, select Growth by Start Score (recommended) and the Norms or Benchmark View. With the Norms view, use the median percentile to compare growth rates. With the Benchmark view use the combined percent in the *Typical* and *Aggressive* growth categories to compare growth rates.

There are two ways to evaluate growth by student characteristic.

1. To determine the aggressiveness of the growth needed to achieve the end of year goal, select the demographic group of interest and use *Growth by All* to compute percentiles and *Low-Risk* (or Advanced when available) for the *End of Year Goal*. The combined total in the *Typical* and *Aggressive* growth categories is an indicator of the overall aggressiveness of growth needed to achieve the goal.
2. To compare growth across two or more demographic subgroups, select *Growth by Start Score* and Norms. Separately, select each demographic and compare the median national percentile in the *Observed Growth* column.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington DC: AERA Publications.

Brown-Chidsey, R. & Bickford, R. (2016). *Practical handbook of multi-tiered systems of support: Building academic and behavioral success in schools.* New York, NY: Guilford Press.

Christ, T. J., & Aranas, Y. A. (2014). Best practices in problem analysis. *Best Practices in School Psychology V*, *2*, 159–176.

Deno, S. L. (2005). Problem-solving assessment. In Brown-Chidsey, R. *Assessment for intervention: A problem-solving approach* (pp. 10-36). New York: Guilford.

Glass, G. V. (1978). Standards and criteria. *Journal of Education, 15*(4), 237-261.

Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards. *Educational Measurement, 4*, 433-470.

Jenkins, J. R. & Johnson, E. (2015). *Universal Screening for Reading Problems: Why and How Should We Do This?* Retrieved from: www.rtinetwork.org/essential/assessment/screening/readingproblems.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 426-461.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. (2006). Validation. In Brennan, R. L. *Educational Measurement, 4th Ed.* Westport, CT: American Council on Education and Praeger Publishers.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448.

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBMreading cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, *23*(4), 304–325.

Sireci, S. G. (2005). The most frequently unasked questions about testing. In Phelps, R. *Defending Standardized Testing.* (pp. 111-121). Mahwah, NJ.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest. 1*(1), 1-26.