Early Literacy Screening Assessment Benchmarks

What "At Risk of Reading Difficulty" Means

Mariann Lemke, Dan Murphy, Aaron Soo Ping Chow, Angela Acuña

Fall 2024

Background

Beginning with the 2020/21 school year, the Massachusetts Department of Elementary and Secondary Education (DESE) began an ongoing effort to collect and analyze literacy screening assessment data from schools and districts participating in certain state grants to inform improvement efforts. Grantee schools and districts that provide literacy screener data to DESE select their screening assessments from a list of state-approved, commercially available literacy screener products, and each assessment is typically administered to students three times per year (most commonly in the fall/beginning of year [BOY], winter/middle of year [MOY], and spring/end of year [EOY]).

Although all of the approved assessments are frequently used for early literacy screening, they vary in significant ways, including the content assessed, the technical characteristics of the assessments, the mode of administration, benchmark and risk definitions, and cut score calculations. This issue brief describes the benchmarks used to identify students as "at risk" and examines how those definitions of risk compare.

Defining Reading Risk

Most of the approved screening assessments provide several performance benchmarks or risk levels (e.g., "some risk"/"high risk"), which are intended to identify students at risk of reading difficulty. In this issue brief, we focus on benchmarks that DESE identifies in its *Early Literacy Screening Guidance* (DESE, 2023) for each approved screening assessment that it recommends schools and districts use to determine whether students are performing "significantly below relevant benchmarks" as

Data Included in 2022/23 Linking Analysis

- *More than 14,000 students in grades 2 and 3 from 211 schools within 69 districts*
- *Scores from nine literacy screening assessments: Acadience Reading, aimswebPlus, DIBELS 8th Edition, FastBridge aReading, i-Ready Diagnostic, Lexia RAPID, mCLASS, Star Early Literacy, Star Reading*

required by state regulation. For example, for DIBELS 8th Edition, DESE recommends using the "Well Below Benchmark" performance level to identify students as significantly below benchmark. These benchmarks, however, differ in how they were determined by the various test developers and in what they represent. The performance levels described by these benchmarks may also become easier or more difficult to meet over time (e.g., from BOY to EOY) and/or across grade levels.

Table 1 describes the meanings of "significantly below benchmark" performance levels for the most commonly used assessments in the Massachusetts early literacy screening data sample in 2022/23: DIBELS 8th Edition (24 percent of scores), mCLASS (18 percent of scores), i-Ready (17 percent of scores), and Star Early Literacy and Star Reading (32 percent of scores).[1] These screening assessments together represent about 91 percent of scores and three different approaches to determining whether or not students are at risk of reading difficulty.

Table 1. Benchmark descriptions for "significantly below benchmark" performance

Early literacy screening assessment	What does "significantly below benchmark" mean? How was the benchmark determined?
DIBELS 8th Edition and mClass	At BOY, MOY, and EOY, a score indicating performance significantly below benchmark ("well below benchmark" in DIBELS and mClass terms) identifies most students who would be expected to score at or below the 20th percentile on an EOY assessment. For kindergarten, the EOY assessment used in analysis was DIBELS Next, and for grades 1–3, it was the Iowa Assessment (Total Reading Score). The Iowa Assessment is described as "a published, group-administered, multiple-choice, norm-referenced measure of reading achievement," and technical documentation notes, "Whereas DIBELS Next includes letter naming and phonemic awareness component skills in the composite score, the Iowa Total Reading Score does not assess these same component skills, making it a more distal criterion measure." Based on studies carried out between 2017 and 2019 with about 7,000 K–3 students, the well below benchmark cut score will accurately identify 80 percent of students who would perform at the 20th percentile or below at EOY. In other words, being well below benchmark identifies students whose reading skills are still likely to be less well developed than those of most of their peers by EOY if they do not receive intensive intervention.
i-Ready Diagnostic	At BOY, MOY, and EOY, scores indicating performance significantly below benchmark ("at risk" on the i-Ready Diagnostic) describe the grade level associated with a student's performance in the context of college and career readiness standards. That is, "significantly below benchmark" generally means that students are performing one or more grade levels below their assigned grade. For example, at BOY, grade 3 students classified as "at risk" based on the i-Ready Diagnostic are performing at or below grade 1 standards; at MOY, grade 3 students classified as at risk are performing at grade 2 standards or below; and at EOY, grade 3 students classified as at risk are performing at or below a level that indicates partially meeting grade 3 standards.

1 Note that DIBELS 8th Edition and mClass are based on the same assessment tasks and use the same scoring approach. Also note that this brief refers to students scoring "significantly below benchmark" and being "at significant risk" interchangeably.

Early literacy screening assessment	What does "significantly below benchmark" mean? How was the benchmark determined?
Star Early Literacy and Star Reading	At BOY, MOY, and EOY, a score indicating performance significantly below benchmark ("intervention" or "urgent intervention" in Star terms) means that students are performing below the 25th national percentile, based on a 2014/15 study including more than 500,000 unique students who took Star assessments. At least 75 percent of students would be expected to perform better than students scoring at the intervention or urgent intervention level.

Sources: Curriculum Associates (2019); Renaissance Learning (2022a, 2022b, 2022c); University of Oregon (2020).

DIBELS 8th Edition and mClass used quantitative analysis to determine a cut score that accurately predicted performance on a different assessment of reading skills given to students at EOY. In this case, the assessment developers determined that performing at or below the 20th percentile at the end of the year on the DIBELS Next assessment (for kindergarten) or the Iowa Assessment (for grades 1–3) was a good indicator of risk. They then chose DIBELS 8th Edition cut scores that would accurately classify students who performed above and below the 20th percentile on those assessments.

i-Ready's publishers used a judgment-based method, informed by other data, to establish general performance level cut scores related to student grade-level performance. In this approach (a "contrasting groups" method of standard setting), test developers wrote descriptions of performance levels for students who partially met grade-level standards and students who just met grade-level standards, and the descriptions were reviewed by a panel of teachers. Panelists then rated their own students (who had also taken i-Ready) according to the performance level descriptions. Teacher ratings were matched to scores, and data were analyzed, to find a cut score that most closely matched teacher ratings. Teachers reviewed the draft cut scores and other data (such as the national percentages of students who would be classified in different performance levels, based on their cut scores), and final cut scores were established. i-Ready's publishers then used these cut scores to identify risk levels for Massachusetts.

Star Reading and Star Early Literacy used a normative approach to determining risk of reading difficulty. In this approach, student scores were compared to data from a national study conducted by the assessment publisher (Renaissance Learning, 2022a). Students whose scores put them in the bottom 25 percent were considered at significant risk of reading difficulty. This benchmark (the 25th national percentile) also corresponds to DESE guidance.

As previously noted, screener assessments also differ in other ways. Taken together, differences in test content, performance standards, and other aspects mean that there is no common definition of risk across screening assessments.

Using the Massachusetts Comprehensive Assessment System to Compare Benchmarks

Although there is no common definition of "reading risk," there is one assessment that all Massachusetts students in grade 3 take. The Massachusetts Comprehensive Assessment System (MCAS) English

Language Arts (ELA) assessment can provide a mechanism to use a single metric to compare screening assessment benchmark cut scores. Using a method called equipercentile linking, we mapped benchmark cut scores from grades 2 and 3 literacy screening assessments to MCAS Grade 3 ELA scale scores.

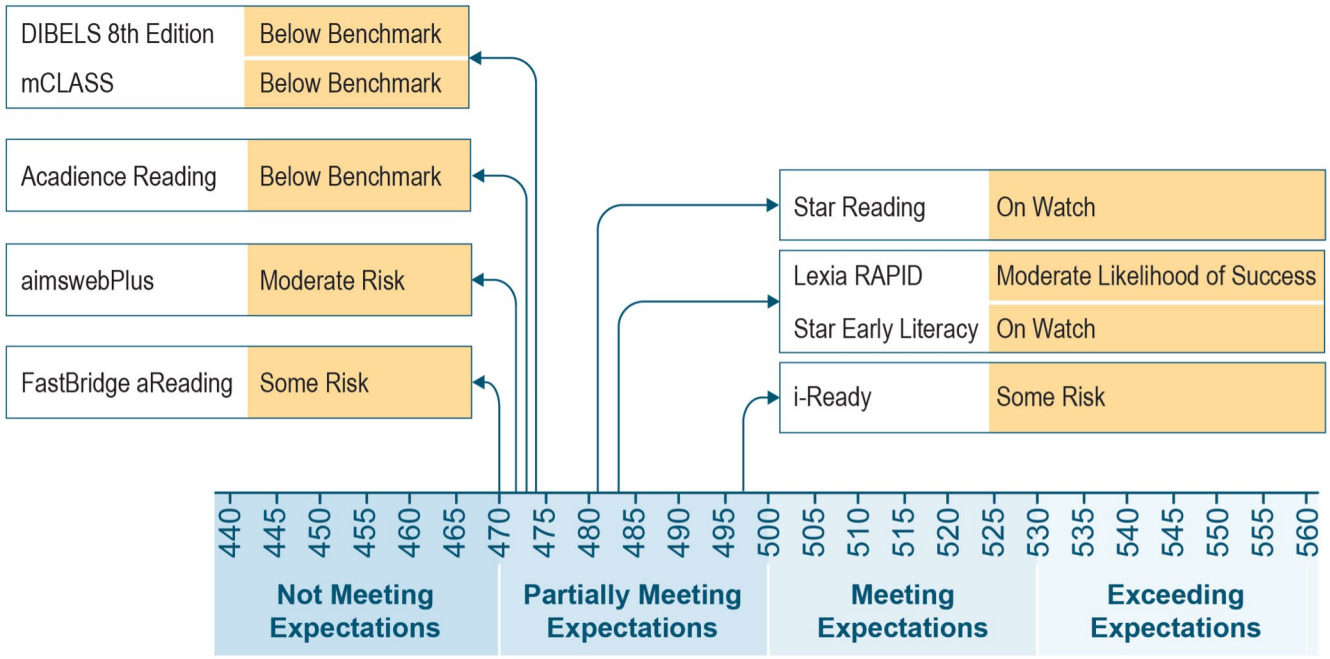
Publisher-provided benchmark categories representing performance significantly below benchmark at BOY and EOY largely link to MCAS scores in the Partially Meeting Expectations performance level, which ranges from 470 to 500 on the MCAS scale (Table 2). One assessment benchmark falls into the Not Meeting Expectations level at BOY. Figure 1 provides a visual representation of the grade 3 EOY "significantly below" screening assessment benchmarks, showing their relationships to the MCAS and to one another in MCAS scale score terms. Their mapping to the Partially Meeting Expectations MCAS performance level is not surprising, given that the benchmarks aim to identify students in need of additional support.

Table 2. Literacy screening assessment grade 3 BOY and EOY benchmark cut scores linked to MCAS Grade 3 ELA scale scores and performance levels, using equipercentile linking

Early literacy screening assessment	Benchmark	MCAS scale score		MCAS performance level	
		BOY	EOY	BOY	EOY
Acadience Reading	Below Benchmark	479	473	Partially Meeting	Partially Meeting
aimswebPlus	Moderate Risk	476	472	Partially Meeting	Partially Meeting
DIBELS 8th Edition	Below Benchmark	477	474	Partially Meeting	Partially Meeting
mCLASS	Below Benchmark	474	474	Partially Meeting	Partially Meeting
FastBridge aReading	Some Risk	469	470	Not Meeting	Partially Meeting
i-Ready	Some Risk	483	497	Partially Meeting	Partially Meeting
Lexia RAPID	Moderate Likelihood of Success	494	483	Partially Meeting	Partially Meeting
Star Early Literacy	On Watch	485	483	Partially Meeting	Partially Meeting
Star Reading	On Watch	487	481	Partially Meeting	Partially Meeting

Sources: District-provided screening assessment data and state-provided MCAS data.

Figure 1. Grade 3 literacy screening assessment EOY benchmark cut scores vary somewhat in how they map to MCAS ELA scores, but all are below the MCAS Meeting Expectations level



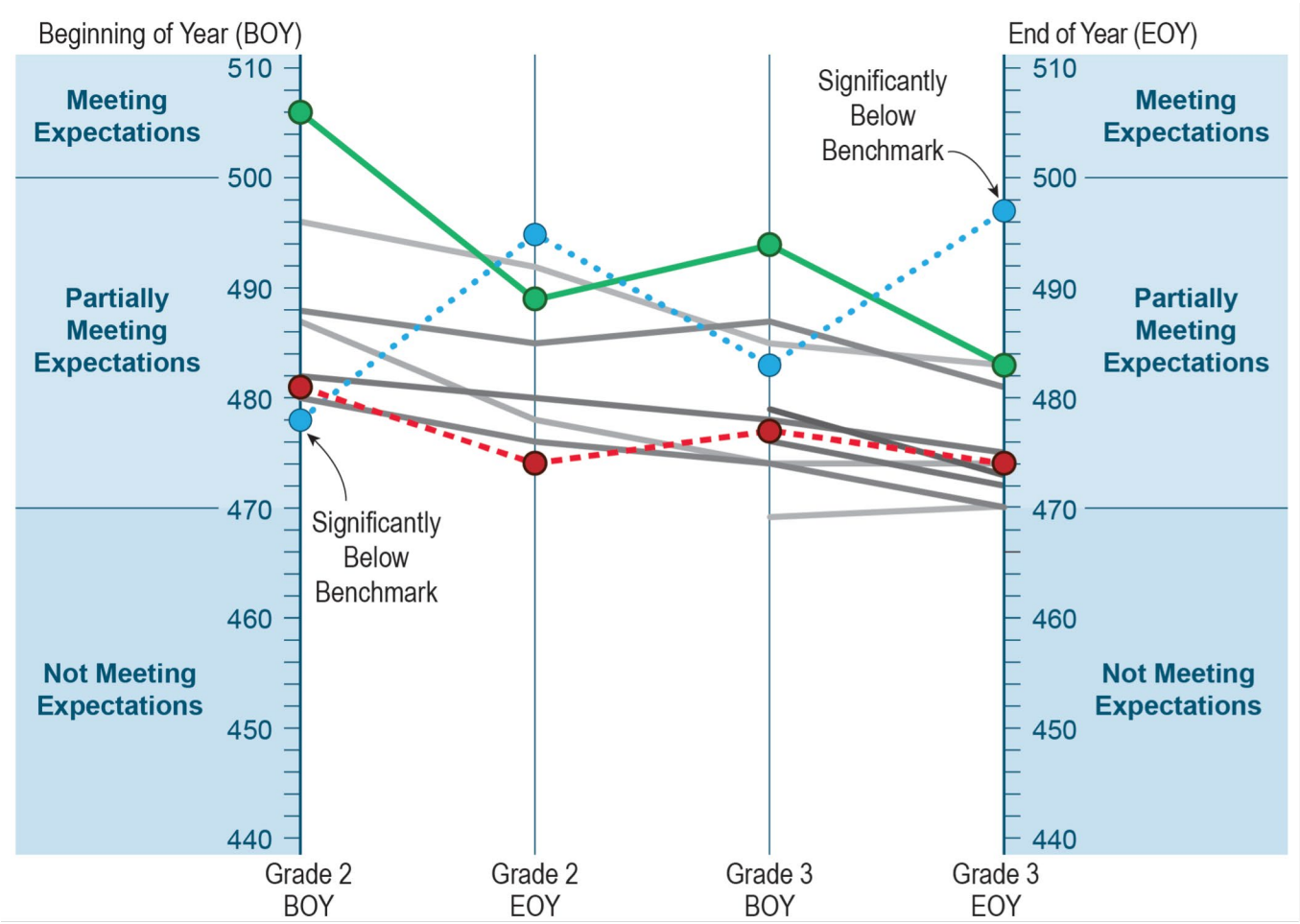
Sources: District-provided screening assessment data and state-provided MCAS data.
Note: Benchmarks indicate levels that will result in a student being classified as at significant risk.

Analysis using 2021/22 data examined only grade 3 EOY screening assessment benchmarks (see *A First Look at Early Literacy Performance in Massachusetts: Results of Initial Analysis Based on State Grantee Literacy Screening Assessments* [Lemke et al., 2023] for details). Using 2022/23 data, we also linked grade 3 BOY benchmarks and grade 2 BOY and EOY benchmarks to grade 3 MCAS scores. To ensure maximum comparability when examining BOY and EOY benchmarks, we matched students so that the equipercentile estimates represent the same students in each time period.

Screening assessment benchmarks indicating significant risk do not always map to the same MCAS or national percentile scores over time, which means that students with the same skills might be classified differently at different time periods.

If the interpretation of scores across time periods for a given screening assessment is intended to be the same, we would expect each time period's score to map to the same MCAS score. However, analysis shows some variation within assessments in where benchmarks link to MCAS between BOY and EOY, ranging from a decrease of 11 points between BOY and EOY to an increase of 14 points, whereas other benchmarks stay relatively constant over time (Figure 2). Additionally, some benchmarks shift in relation to MCAS across grades. This means that, depending on the benchmark patterns, students with the same screening assessment scores at BOY and EOY might be classified as significantly below benchmark at one time and not at the other.

Figure 2. Screening assessment benchmarks indicating significant risk vary within and across grade levels on the MCAS scale, meaning that identification as significantly below benchmark may be harder or easier at different times



Sources: District-provided screening assessment data and state-provided MCAS data.
Note: Figure includes all students with BOY and EOY scores and MCAS scores in 2022/23. Each line represents a screening assessment mapped to MCAS at four different time points (grade 2 BOY and EOY and grade 3 BOY and EOY). For some assessments, data were not available for grade 2 students and so the lines are incomplete. Several lines are highlighted in different colors (red, green, blue) to illustrate different patterns of change (e.g., increasing over time, decreasing over time, staying relatively constant over time).

Some change may be due to imprecision in linking estimates, but larger differences demonstrate variation in how benchmarks for different times of year were set. This variation may reflect differences in intended test purpose or use. Some assessments may prioritize measuring growth; others may prioritize growth toward a particular standard. As previously noted, screening assessments vary in many ways, from content to administration to benchmark-setting procedures. These differences do not indicate that one assessment is better than another; rather, they indicate that users must be aware of how their assessments were designed, because differences in benchmark-setting procedures can result in changes in the numbers of students identified as significantly below benchmark in each time period, which may not reflect changes in student knowledge and skills.

Changes in the percentages of students identified as significantly below benchmark may be due in part to changes in the benchmarks themselves.

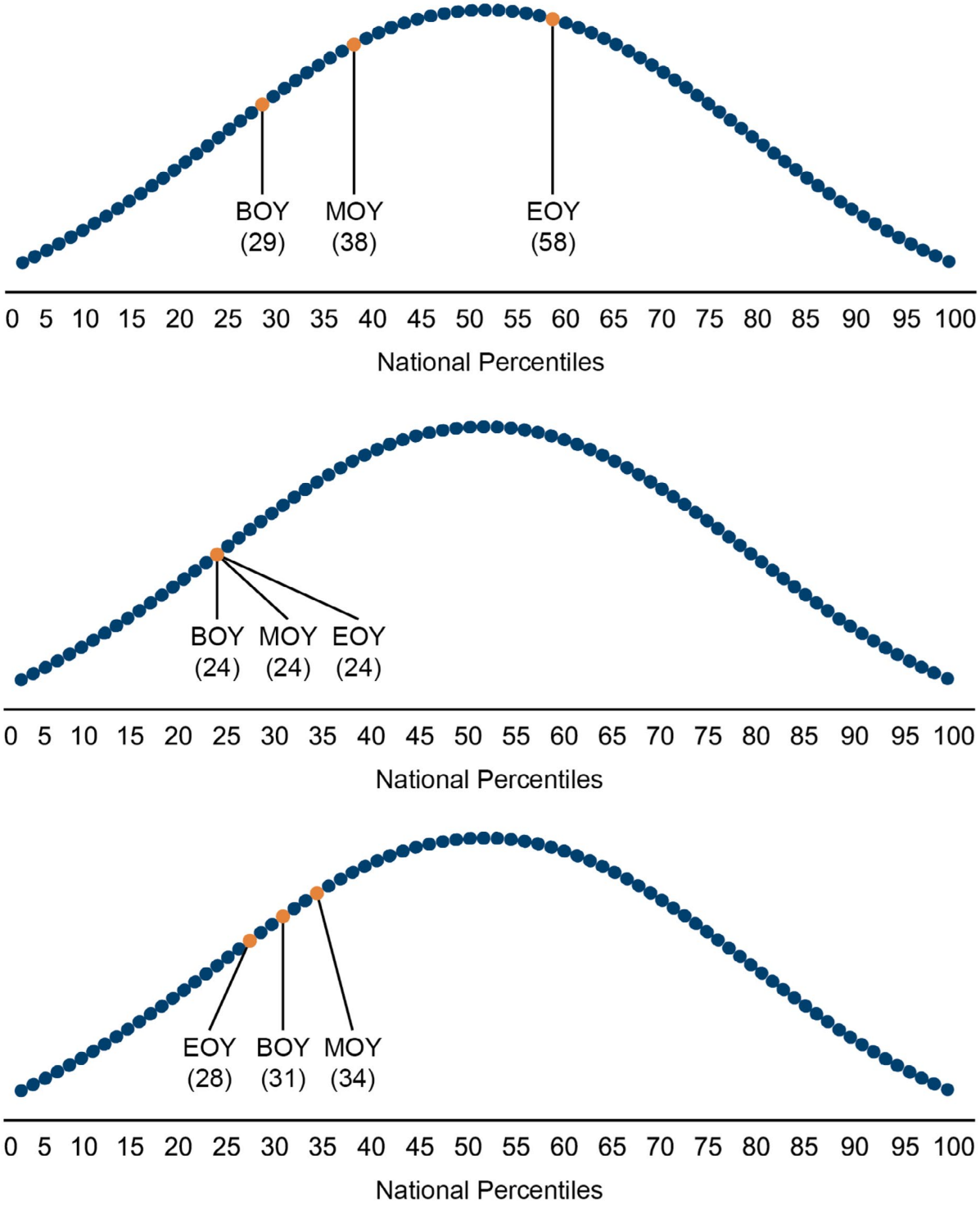
Benchmark-setting differences can also affect how student growth is understood. Students in schools that use an assessment with a benchmark that shifts from lower to higher (i.e., easier to harder) on the MCAS scale between BOY and EOY may improve their performance relative to the BOY benchmark but may appear to not show progress relative to the EOY benchmark. For example, if an assessment's benchmark indicating significant risk maps to an MCAS score of 478 at BOY, and a student scores below this level and then improves their skills by EOY, but the benchmark now maps to an MCAS score of 495 (as the blue dotted line does in Figure 2), the student may still be classified as significantly below benchmark. Whether such changes are intentional or unintentional, they affect how users should interpret performance changes in relation to benchmarks.

Conversely, students in schools that use an assessment with a benchmark that shifts from higher to lower (i.e., harder to easier) on the MCAS scale (as the red and green lines between BOY and EOY do in Figure 2) may appear to grow out of the significantly below benchmark category by EOY while still performing at a skill level similar to the BOY benchmark. Across grades, some students may appear to have lost ground over the summer if benchmarks shift upward from the end of one grade to the beginning of another, whereas other students may appear to progress over the summer if benchmarks shift downward. How benchmarks within assessments compare across time periods is important for schools to understand, so that they take these benchmarks into account when reflecting on student performance, and is especially important for analysis of growth within and across assessments.

Using national percentiles provides another way to examine screening assessment benchmarks over time, within and across assessments, as most assessments provide national percentile scores even if these scores are not used to identify students at risk of reading difficulty. The national percentiles associated with benchmarks indicating significant risk can vary from a national percentile of 13, meaning that about 13 percent of students nationally would be identified as significantly below benchmark and in need of intensive support, to a national percentile of 58, which would identify a much larger proportion of students.

The percentiles associated with significant risk tell the same story as the MCAS equipercentile linking—that benchmarks can shift across BOY, MOY, and EOY within assessments. Figure 3 shows BOY, MOY, and EOY benchmarks from three different screening assessments to illustrate how benchmarks can vary across assessments and across time periods within assessments. For example, in the first panel of Figure 3, about 29 percent of students would be identified as significantly below benchmark at BOY, but 58 percent would be identified at EOY. In the middle panel, the benchmark indicating significant risk is the same at each time period and would be expected to identify about 24 percent of students. Finally, the third panel shows benchmarks that would identify the largest number of students at MOY and the smallest number at EOY. As previously noted, these differences can affect the interpretation of student growth across time periods.

Figure 3. Grade 3 screening assessment benchmarks representing performance significantly below benchmark at BOY, MOY, and EOY, with corresponding national percentiles



Source: District-provided screening assessment data.
Note: Orange dots indicate BOY, MOY, and EOY benchmarks for three different screening assessments to illustrate how they differ in the percentages of students identified as significantly below benchmark at each time point.

Implications for Policy and Practice

- **Test developers and researchers should make how risk of reading difficulty is defined in screener assessments more transparent.** Clear definitions, in plain language, of what having different screening outcomes (e.g., being at risk of reading difficulty) means should be included in documentation and reviews of assessment tools, to help users make decisions about which tools to use and how to use them. Screening is intended to identify which students need support, but the kinds of support that are indicated for identified students will depend on what being "identified" actually means. For screening to best serve its role, users must know what kinds of interventions to design and implement to address the needs of students identified as at risk.
- **Educators should know what "reading risk" means for the assessments they use and how that meaning affects score interpretations, especially in relation to growth.** Making information about what scores mean available and understandable is an important first step, and publishers, states, and others will also need to work to communicate this information to educators. This issue brief provides examples of the different ways in which "reading risk" is defined for different assessments. Educators should understand the meaning of a student being flagged as significantly below benchmark on the screening assessments they use, because the details of how the student is flagged may imply different responses to the data. For example, normative benchmarks do not provide information about student knowledge or skills, and they assume that students in a given school are comparable to students in the sample used to set norms. Other kinds of benchmarks provide information about how students are predicted to perform later in the year on entirely different assessments—those benchmarks and normative ones may both measure "reading" or "reading skills," but in very different ways. Finally, measuring growth or progress in terms of the numbers of students who moved from one benchmark level to another may not provide the information that users expect in situations where the benchmarks themselves shift between time periods. As an example, if the percentage of students performing below benchmark decreases between time periods, but those decreases are due, in part, to benchmark attainment having become easier, users should be aware of that information.
- **Educators and policymakers must be cautious in particular about using scores from screening assessments for purposes other than identifying students who are in need of support, such as evaluating programs or schools, particularly if this use involves data from multiple assessments.** Early literacy screening assessments were initially intended to be used for one purpose: to identify students who may need additional support to gain the reading skills they need to be successful. Going beyond that purpose requires careful attention to differences within and between assessments, and openness about the potential limitations to the validity of such analysis.

References

Curriculum Associates. (2019). *i-Ready® Assessments technical manual*.

Lemke, M., Murphy, D., Soo Ping Chow, A., Spencer, H., & Zhang, A. (2023). *A first look at early literacy performance in Massachusetts: Results of initial analysis based on state grantee literacy screening assessments*. WestEd. https://www.doe.mass.edu/instruction/ela/research/first-look.pdf

Massachusetts Department of Elementary and Secondary Education. (2023). *Early literacy screening guidance*. https://www.doe.mass.edu/instruction/screening-assessments.html

Renaissance Learning. (2022a). *Defining benchmarks in Star Assessments*. https://star-help.renaissance.com/hc/en-us/articles/24424240092827-Defining-Benchmarks-in-Star-Assessments

Renaissance Learning. (2022b). *Star Assessments™ for Early Literacy technical manual*. https://star-help.renaissance.com/hc/en-us/articles/12483321397019-Star-Assessments-for-Early-Literacy-Technical-Manual

Renaissance Learning. (2022c). *Star Assessments™ for Reading technical manual*. https://star-help.renaissance.com/hc/en-us/articles/12542471051803-Star-Assessments-for-Reading-Technical-Manual

University of Oregon. (2020). *8th edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Technical manual*. https://dibels.uoregon.edu

WestEd
WestEd.org

Suggested citation: Lemke, M., Murphy, D., Soo Ping Chow, A., & Acuña, A. (2024). *Early literacy screening assessment benchmarks: What "at risk of reading difficulty" means*. WestEd.

WestEd is a nonpartisan, nonprofit organization that aims to improve the lives of children and adults at all ages of learning and development.

For more information, visit WestEd.org. For regular updates on research, free resources, solutions, and job postings from WestEd, subscribe to the E-Bulletin, our semimonthly e-newsletter, at WestEd.org/subscribe.