DISTINGUISHED SCHOLAR SERIES

# Indicators of Quality for Universal Screening Assessments of Early Reading: A Comparative Analysis

*Rachael Gabriel, University of Connecticut*

## About the Author

Dr. Rachael Gabriel is Professor of Literacy Education at the University of Connecticut. A former teacher and reading specialist, Rachael's research is focused on literacy instruction, leadership, and intervention, as well as policies related to teacher development and evaluation.

Author's note: The opening poem is a variation on *The Red Wheelbarrow* by William Carlos Williams (1962).

*Editor's note: Given the wide variance in screening tool naming conventions, we have used the National Center for Intensive Intervention (NCII) chart names for clarity.*

*So much depends*
*upon*
*oral reading fluency*
*screeners*

*Gateways to*
*sorting*

*and labeling*
*Also.*

Universal screening assessments are designed to be brief, efficient indicators of the potential for future conditions, including risk of difficulty, disability, and low achievement. Despite being designed to indicate, not diagnose, the effects of a universal screening assessment can be felt across all aspects of a reading program, and throughout the school's entire infrastructure for learning. A shift in screener influences all pillars of instruction for literacy improvement (Woulfin & Gabriel, 2020), and can cause shifts in the focus of teacher professional development and coaching, as well as the priorities of school leaders, alignment (or lack thereof) with curriculum, and the way data are generated and analyzed to inform grouping and instruction (see Figure 1). From a student perspective, a screener may be the first and most recognizable indicator of the need to add or change instruction, grouping and intervention, which impact an individual's social, cognitive, and emotional experience in school, as well as their unique set of opportunities to learn (Connor et al., 2009). Though positioned as a small part of a system of instruction and assessment, much depends on universal screeners and the data they routinely produce in schools.

As state laws increasingly require not only the use of universal screeners but the adoption of specific screening assessments, there is a need to understand the features and factors that could impact a use and interpretation. Therefore, in this paper I analyze publicly available technical manuals for six of the most commonly approved screeners in order to address the following questions: "What makes a good screener?" And, within each screener: "What counts as good reading?"

The results from those analyses are then used to compare the purpose and construction that differentiate state-approved options. I conclude with a discussion of the factors that may contribute to the selection of screeners and implications for the ways reading is constructed within them.

## Universal Screening: A Brief Overview

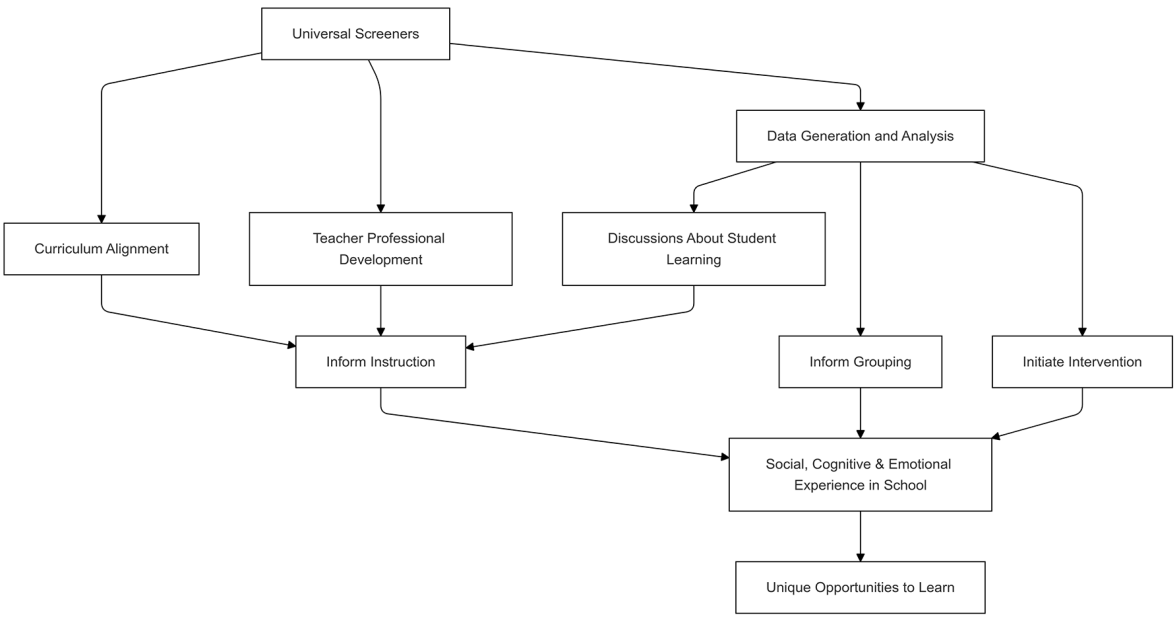
### *Asking and answering assessment questions*

The concept of universal screening was initially drawn from medicine, and later psychology, where an entire population is assessed to determine the risk or presence of difference, disease, or difficulty. It is used in particular where early identification is associated with more positive outcomes, as in the case of genetic disorders, where early identification can prevent serious difficulties (Ketter et al., 2014). The question a universal screener asks and answers in the case of newborn screening is, "Are any known indicators of blood, heart or hearing disorders present?" The value of a screener is to accurately identify those who might later be diagnosed with a small set of treatable conditions present at birth. Screenings are designed to be simple, inexpensive, and noninvasive. They are therefore not diagnostic, but they prompt the use of diagnostic testing that would be inefficient or impossible to apply universally. In medicine and psychology, they are most often part of a "two-gate" system, where the screener (first gate) triggers further testing (second gate) which identifies or confirms a physical or genetic difference requiring treatment (c.f. Walker et al., 2014).

Universal screenings have existed in education settings in varied ways for decades. However, they have become much more common, and are being used in increasingly standardized ways since 2004, when the Individuals with Disabilities in Education Act was reauthorized to allow responsiveness to intervention to be used in the identification of disabilities in lieu of an IQ-performance discrepancy. This provision placed greater importance on establishing systematic and equitable ways to identify students for intervention and monitor their progress with increasing levels of support. Still, researchers have demonstrated that predictions of difficulty within two-gate systems tend to outperform those made by a single screener (VanMeveren et al., 2020).

In some settings, screening all students to identify a subset of students for intervention was a relatively new practice. Thus, it initiated/encouraged a new set of assessments to address the question, "Which students need additional support to meet age or grade-related benchmarks?" Within response to intervention systems, the assessment question a universal screen is designed to answer is, "Which students are at risk of having a disability?" In this case, the value of a screener is its ability to accurately identify students who will later be

**Figure 1. The Influence of Universal Screeners Across Pillars of Literacy Improvement**



classified as having a disability. This requires comparison to a norm, rather than to the class, as reference. Screening assessments therefore may not focus on what was taught, but on what most students know and can do at a certain point in their development. If many students qualify, the size of the intervention or support setting may increase to accommodate them. If no students qualify, then no support needs to be offered. Though logical and familiar, this is not the only use of universal screening in early literacy or other areas.

For example, before 2001, schools using Reading Recovery® may have had a system of universal screening for first-grade reading to identify students for intervention. Within and outside of response to intervention systems, Reading Recovery is designed as a short-term intervention that either returns students to general classroom instruction without support, or generates data to inform a long-term system of support, including further testing. The universal screener used to determine eligibility for reading intervention has varied over time and across settings from teacher reports, to spelling, to one or more subtests of *An Observation Survey of Early Literacy Achievement* (Observation Survey/OSELA; Clay, 2019), or another locally approved screening assessment. The assessment question a screener is designed to answer where Reading Recovery is in use is, "Which students demonstrate the lowest literacy achievement in the group?" (North American Trainers Group, 2012). Though administrators might estimate that 20–25% of the first-grade class might enter intervention across the year (Clay, 2005, p. 7), the goal is to rank students by present achievement in reading and provide support to as many as is feasible in a one-to-one setting.

Over time and across contexts, the question that universal screening assessments are designed to ask and answer has varied. However, as schools (re)built systems of intervention that included multiple tiers of support, the practice of universal screening became more common and more standardized across settings (c.f. Mellard et al., 2009). In the 20-teens, a wave of legislation aimed at screening for dyslexia further standardized the tools used to conduct universal screens and brought increased emphasis on commercially developed, research-based assessments that could contribute to the identification of dyslexia. This was a key legislative goal of dyslexia advocates (Decoding Dyslexia, 2024) because, they argued, the goal of identifying dyslexia was not adequately met by current screening practices. Though screening was already included in many state laws, the assessments were rarely specified, and a range of tools were in use. This meant a range of assessments were used in varied ways, including the use of locally developed, or "homegrown" assessments, and assessment tools focused on varied areas of literacy development (e.g. spelling, phonics, word recognition, fluency).

> The concept of universal screening was initially drawn from medicine, and later psychology, where an entire population is assessed to determine the risk or presence of difference, disease, or difficulty.

### *Shifting assessment questions*

The focus on dyslexia added an analytic question to common uses of universal screening assessments. For example, when screening is to identify students who might benefit from intervention, assessment data is gathered in order to ask,"Which students perform differently from their peers?" When teachers aim to identify students at risk for difficulty, they might ask, "Which students perform differently from expectations for this point in time?" Or, more specifically, "Which students demonstrate characteristics of dyslexia?"

When screening is used to identify the risk of dyslexia, the question is no longer comparative, but categorical. Consequently, there becomes an increased focus on testing aimed at the areas implicated by current definitions of dyslexia, including phonics and phonemic awareness, and other areas related to characteristics of dyslexia like rapid automatic naming (Norton & Wolf, 2012; Torgeson, 2004). By asking and answering a slightly different assessment question, universal screeners aimed at early identification of dyslexia are associated with different outcomes than other universal screens (Odegard et al., 2020). Even without a specific focus on dyslexia, cut scores or the scores which draw the line between expected/unexpected or proficient/in need of support, are different across contexts. When cut scores

are decided by a national assessment company, some schools may have a large percentage of students qualifying for support, while other schools have comparatively few. If they are set locally, some students who would qualify for intervention in one school or district may not in another. It is the questions we use assessments to answer that determine the design, use, and impact of the assessment.

Though universal screening has a long history in education settings, legislation requiring the use of assessments capable of contributing to the identification of dyslexia were either premature or overstated: No such assessment had or has been identified as individually capable of providing the information necessary to determine risk. Current guidance suggests that a combination of indicators, not all of which could be included on a single screener, is needed to determine risk (National Center on Improving Literacy, 2019). These include considerations of family history and response to instruction. Some recent studies have concluded that a "two-gate approach," where multiple assessments are used in a particular order, is the most likely to ensure the accuracy of risk identification (c.f. VanMeveren et al., 2020). Other studies find that multivariate approaches, where two tests of different areas are used in combination, support better efficiency and accuracy (Klingbeil et al., 2017). Without a clearly identified screener or set of screeners to meet the requirement of evidence-based assessments for difficulty, disability, or dyslexia, options have proliferated, with few ways to judge their relative benefits and drawbacks.

### *State-mandated universal screening*

The majority of states now mandate forms of screening in the area of early reading. As of 2024, 42 states mandate screening for reading difficulty in Grades K–3 (Schwartz, 2022). Of the 42 that mandate universal screening, 36 maintain lists of approved screeners. Though there are significant points of overlap in lists of state-approved screeners, there is also a large range of options available. For example, Alabama and Iowa have 17 and 13 options, respectively. In contrast, Alaska requires the use of DIBELS exclusively unless a local authority applies to use an alternative. Nevada uses a state-created assessment exclusively. Other states provide guidelines for selection or link to clearinghouses like the National Center for Intensive Intervention (NCII), which maintains a list of screeners organized by amount and nature of evidence for their use.

For the purpose of this analysis, I collected the 36 lists of mandated assessments and identified the six that appear most commonly across states. These include DIBELS, Acadience, Aimsweb, Star, MAP, and iReady (see Table 1). Each of these has a technical manual that is publicly available for review except for AimsWeb, whose technical manuals are only available upon request. This analysis therefore replaces Aimsweb with the next-most common selection, NWEA MAP testing. I focus exclusively on first grade to create an apples-to-apples comparison. However, it should be noted that assessment systems often vary skills assessed and question types by grade, and may have separate testing programs for different combinations of grade levels (e.g. K–2, K–3, K–5). Likewise, the reliability and validity of tests and subtests varies across grade levels. In the section that follows, I describe the analysis used to compare the most frequently approved options for screening

**Table 1. Ten Most Common Entries on Lists of State-Approved Screeners**

| | Total Number | Percent |
|---|---|---|
| Acadience | 20 | 57 |
| Aimsweb or AimswebPlus | 18 | 51 |
| Amira | 12 | 34 |
| **DIBELS and/or MClass** | **32** | **91** |
| easyCBM | 8 | 23 |
| FastBridge | 18 | 51 |
| iReady | 20 | 57 |
| **ISIP or iStation** | **15** | **43** |
| **MAP (NWEA)** | **18** | **51** |
| **Star** | **18** | **51** |

NOTE: Red indicates NCII's highest possible rating for classification in fall of first grade.

in states with specific assessment mandates for first grade.

## Analysis

To complete comparisons of commonly approved tools, I engaged in a three-step process. First, I gathered the technical reports and the review of each. The NCII is one of only a few comprehensive sources of information about the evidence for varied reading screeners (see https://intensiveintervention.org/tools-charts/overview). Unlike other databases of available screeners, which may link to available evidence, NCII publishes its own ratings of screening tools based on classification accuracy, technical standards—including reliability and validity—and usability features. Five screening tools have the highest possible rating for classification accuracy in the fall of first grade. Of these, four are among the top 10 most popular choices, and the fifth is the Observation Survey, which does not appear on any state list. NCII found adequate reliability and validity for 18 screeners, including each of the screeners that appear most often on state lists, the Observation Survey, and seven others (see Table 1).

As a second step, I visited and reviewed the official product page of each assessment and noted any other product cross-advertised on its landing page (e.g. without any additional links or navigation). Then, I navigated the site to identify the publisher and any distribution partners. Table 2 shows the publisher and potential packaging of the six tools included in this study. Publisher information may be used to identify products that might be compatible and might be sold together as a bundle. In particular, iReady, MAP, and Star offer products for math in addition to reading, and MAP also has a science test for use in upper grades. I used Google searches to identify and corroborate (e.g. two or more sources used) information about parent companies and owners of each company as of 2024 in order to identify similarities across the flow of funding from mandated district contracts.

Finally, I read and coded the technical manuals to identify and confirm patterns in the topics covered and the ways that purpose, reliability, and validity were described. I extracted salient points from these manuals to facilitate comparison and worked to identify patterns in the ways each was used to indicate assessment quality.

In the section that follows, I present findings from the comparison of tools and analysis of quality indicators. I then discuss these findings in terms of decisions about assessment selection and implementation.

**Table 2. Packaging and Ownership Information**

| Screener | Advertised With | Publisher or Distribution Partner | Parent Company | Screener Ownership |
|---|---|---|---|---|
| Acadience Reading K–6 | Acadience Data Management | Acadience Learning & Voyager Sopris | Lexia, Cambium Learning Group | Veritas Capital |
| DIBELS 8th Edition | DIBELS Data System | Amplify | N/A | Private Investors |
| AimswebPlus Reading | Spell-links to reading and writing, DRA3, Review360 | Pearson Clinical Assessments | Pearson | Publicly Traded |
| FastBridge | FastBridge Reading, Math, and SEL screeners | Illuminate Education | Renaissance Learning | Francisco Partners |
| MAP | Integrated with 18 supplemental programs for first-grade reading | Northwest Evaluation Association (NWEA) | Houghton Mifflin Harcourt | Veritas Capital |
| Star | Star Reading & Star in Spanish | Renaissance Learning | N/A | Francisco Partners |

## Findings

### *What makes a good screener?*

This content analysis of technical manuals includes a comparison of descriptions of the purposes for each assessment, as well as the methods for demonstrating that it is valid and reliable. Technical manuals explain the structure, origin, and specifications of an assessment. In many cases, these are publicly available to download from the publisher's site. In other cases they are available by request. These manuals range from 28 to more than 100 pages of information about each assessment, how it was developed, and how its reliability and validity have been demonstrated. Such manuals also provide detail about some of the most important questions about a test: What is tested? How? How are final scores calculated? How do we know these questions and calculations are fair, valid, and reliable? For example, Illuminate's Fastbridge website states:

> A quality reading assessment should be grounded in the science of reading research to identify the specific literacy skills a student is struggling with and offer evidence-based recommendations on how to close skill gaps. Quality reading assessments should also be proven valid, reliable, and have multiple sources of high quality research indicating that they help teachers identify and solve reading problems that students have in schools. (2024)

One of the sources of research about an assessment is its technical manual, which makes a case for the technical adequacy of the tool. It is important to note that the purposes of each assessment type represented below are distinct, and these purposes inform differences in design and indicators of quality. For example, MAP and Star early literacy aim to align to standards and are designed as adaptive tests taken on a computer. Their alignment to standards is related to a goal of "provid(ing) schools with tests that match(es) their content standards" (NWEA, 2019, p.14).

While some screeners focus on early literacy skills in order to predict whether students will qualify for intervention in the future, others focus on standards for literacy in order to predict how students will score on state and other tests in the future.

Acadience is the newest version of DIBELS, developed by the same researchers, but published and distributed by an independent company. It therefore shares many of the same characteristics in overall design with DIBELS as marketed by Amplify in mClass — individual subtests designed to measure skills associated with early literacy. In contrast to the Fastbridge site, the Acadience technical manual indicates that the primary use of DIBELS/Acadience is to "identify students who may be at risk for reading difficulties" (p. 6). Therefore, while some screeners focus on early literacy skills in order to predict whether students will qualify for intervention in the future, others focus on standards for literacy in order to predict how students will score on state and other tests in the future. These are different ways of answering the same question, which point towards different references for determining what counts as a good or poor performance, and potentially different implications for any given score.

Though reports generated for each aim to make inferences about student reading ability that can inform instruction, intervention, grouping and other issues like curriculum design or analysis, the purpose of each directs its design and the measures that are used to indicate whether it is successful for the purpose(s) for which it was designed. In the section that follows, I compare the content measured, and measures of reliability and validity included in technical reports as evidence of the quality of each assessment.

### *Reliability*

Reliability is a feature of an assessment that relates to its stability and consistency across similar conditions. Reliability can be demonstrated using a range of tasks all designed to show that the test would produce similar results under consistent circumstances. In the case of reading assessment, an appropriate synonym would be "consistency"; the test consistently shows the same thing across time and context.

One way to demonstrate the reliability of tools is to show that the scores would be the same at different times

**Table 3. Reliability Measures**

| Screener | Alternate-Form Reliability | Test-Retest Reliability | Inter-Rater Reliability | Marginal Reliability | Split-Half Reliability |
|---|---|---|---|---|---|
| Acadience Reading K–6 | X | X | X | N/A | |
| DIBELS 8th Edition | X | X | X | N/A | |
| AimswebPlus Reading | X* | X | N/A | N/A | |
| FastBridge | | X | N/A | X | |
| MAP | N/A | X | N/A | X | |
| Star | N/A | X | N/A | N/A | X |

NOTE: *Called equivalency studies, these are designed to compare the stability of different forms of the same test. Up to 14 different forms of each Aimsweb subtest were studied.

(test-retest), no matter who is giving it (inter-rater) and/or on different forms if the test comes with more than one to use (alternate-form). In cases where a test is adaptive, there may not be forms and the same set of questions would not be given multiple times. Therefore, they may use split-half reliability as a way to show internal consistency. This method splits the test in half and compares the results from each half to demonstrate its consistency.

If a test is constructed using item response theory (IRT), which is a method of designing assessments that models the likelihood of a correct response for each item, instead of assuming each item is of equal difficulty and quality. Tests constructed within an IRT paradigm can calculate the amount of someone's score that is likely due to standard error vs. a true score. This can be used to calculate marginal reliability, which indicates how consistent the amount of estimated error is in scores across items in a test. In the same way that a test like Fastbridge or MAP can use its IRT features to calculate marginal reliability—but would not use something like alternate-form reliability—a test like DIBELS, not constructed in this way, couldn't calculate marginal reliability. The type of test determines the ways its consistency can be demonstrated.

Table 3 shows the ways that reliability is described in the six technical reports I analyzed.

My comparative analysis indicates that all of these assessments report some form of reliability. Beyond that, if we consider that Acadience is a version of DIBELS, no more than two assessments use the same indicators of reliability.

Among assessments that include multiple subtests, reliability actually varies for each subtest. So, a composite or overall score may be reliable, though individual subtests are less so. For example, Acadience's alternate-form reliability is lowest for correct letter sounds scored during the Nonsense Word Fluency subtest (r = .85). It is highest for the number of correct words read during Oral Reading Fluency (r = .98). When accuracy during oral reading is calculated as a percentage of words read, the reliability between test forms falls to .88. The potential sources of variation across forms shifts depending on what is being measured (e.g. nonsense words or passage reading).

MAP and Star early literacy are both computer-based adaptive screeners. As such they do not include different forms or require different (human) raters. Aimsweb is administered online but has fixed forms (nonadaptive), so there is also no need for inter-rater reliability. As adaptive tests, questions are generated based on the answers a student provides as they take the test, rather than providing a set of questions as form A and a different set as form B. Star uses split-half reliability to ensure that the first and second half of an assessment are equivalent and stable in a similar way that two forms of the same assessment might be.

As Lemke and colleagues wrote, "Lists of "universal screeners" even include different kinds of apples and different kinds of oranges" (2024, n.p.). The choice of reliability measure aligns with the way the test is constructed. They are similar, but

not equivalent, which makes it difficult to make a direct comparison between options on a state list. What can be said is that each of these has adequate reliability for its test type. And, each test type is designed to ask/answer a different assessment question.

### *Validity*

Validity is the degree to which a test measures what it is intended to measure. Similar to the finding that not all subtests within an assessment are equally reliable, technical manuals and related research find a range of validity across subtests and scores. For example, Johnson et al. (2009), found that "DIBELS Nonsense Word Fluency, Initial Sound Fluency, and Phoneme Segmentation Fluency measures show poor diagnostic utility in predicting end of Grade 1 reading performance" (p. 75). They found that Oral Reading Fluency had better classification accuracy than other subtests, but no better than assuming that no student had a disability. They further found that different cut scores were required to accurately classify students with varied levels of English proficiency. Therefore, scores for those with varying English proficiency required different interpretations to produce an equally valid measure of risk for reading difficulties. Table 4 shows the range of concurrent validity scores across subtests of each of the varied subtests. In contrast to tests like DIBELS with multiple subtests, the adaptive screeners, like MAP and Star, are associated with a single validity indicator. And, given their focus on predicting future test scores, their validity is measured, in part, by their correlation with future scores.

> The choice of reliability measure aligns with the way the test is constructed. They are similar, but not equivalent, which makes it difficult to make a direct comparison between options on a state list.

Table 5 shows the range of ways that validity is demonstrated across the technical manuals associated with the screeners listed above. As with reliability measures, despite some overlap, a direct comparison is difficult. Moreover, the assessments each use different reference points to demonstrate their validity. With concurrent validity, scores on one assessment are correlated with scores on a different assessment of the same construct. DIBELS 8 compared itself to the DIBELS Next Composite score (Acadience), as well as to the CTOPP-2, and Iowa Word Reading. FastBridge compared itself to the Gates MacGinitie Reading Test 4th edition. Star early literacy compared its scores to scores on DIBELS 8 and Group Reading Assessment Diagnostic Evaluation, as well as two state-specific tests: the Michigan Literacy Progress Profile, and Texas Primary Reading Inventory. Again, tests aimed at predicting future performance on tests tend to choose state tests for comparison and conduct studies using multiple tests. Those aimed at predicting classification or risk of difficulty compare themselves with tests more often used in classification or special education decisions.

Given the variation in ways different assessments demonstrate reliability and validity as indicators of quality and utility, it is important to consider what is actually measured

**Table 4. Validity Measures Used**

| Screener | Content or Construct Validity | Criterion-Related Validity | Discriminant Validitiy | Concurrent Validity | Predictive Validity | Classification Accuracy |
|---|---|---|---|---|---|---|
| Acadience Reading K–6 | X | X | X | | | |
| DIBELS 8th Edition | | | | X | X | |
| AimswebPlus Reading | | X | | | | X |
| FastBridge | X | | | X | X | X |
| MAP | | | | X | | X |
| Star | X | | | X | | |

in each. Just as choices about how to measure reliability and validity construct particular versions of what counts as a good screener, choices about what to measure in an early literacy screener construct particular versions of what counts as good reading. The areas assessed and classification schemes used to construct different versions of good reading are discussed in the following section.

### *What counts as good reading?*

Universal screeners approved for use across states are constructed with different purposes, properties, features, and indicators of quality. Though used to screen in the area of reading, they also assess different skill areas associated with reading (see Table 5).

The differences across tests confirm the variation evident in other comparisons and perhaps point to the impact of varied purposes and methods of assessment. Oral Reading Fluency is the subtest most likely to appear on fixed-form assessments, while other subtests seem to vary. Comparing composite scores from any of these things would be like comparing different multivitamins: all the tests measure aspects of reading, but none of the exact same aspects with the same focus or intensity.

There are few similarities across specific areas tested on computer adaptive tests, which cover more areas, with fewer questions. The differences in areas covered may only refer to a small handful of questions, especially in the case of adaptive testing. However, they represent the degree to which the assessment is designed to answer specific questions, either about typical/average development or future performance. These differences are also evident in the way scores are sorted or classified into various levels of performance, as shown in Table 6.

The terms used to classify performance on each assessment are similar in that they represent a

**Table 5. Areas Assessed**

| Screener | Sound- and Symbol-Level Skills | Word-Level Skills | Vocabulary | Fluency and Comprehension |
|---|---|---|---|---|
| **FIXED-FORM ASSESSMENTS** | | | | |
| Acadience Reading K–6 | Letter Naming Fluency, Phoneme Segmentation | Nonsense Word Fluency | | Oral Reading Fluency |
| DIBELS 8th Edition (15 min) | Letter Naming Fluency, Phoneme Segmentation | Nonsense Word Reading Reading Fluency, Word Reading Fluency | | Oral Reading Fluency |
| AimswebPlus* Reading (20–40 min) | Letter Word Sounds | Word Reading Fluency | Auditory Vocabulary | Oral Reading Fluency |
| **COMPUTER-ADAPTIVE ASSESSMENTS** | | | | |
| MAP** (20–40 min) | Phonological Awareness, Print Concepts | Phonics and Word Recognition | Context Clues and References, Vocabulary Acquisition and Use | Literacy: Key Ideas, Craft, Structure |
| Star; 27 items (12–15 min) | Alphabetic Principle, Concept of Word, Visual Discrimination, Phonemic Awareness, Phonics | Structural Analysis | Vocabulary | Sentence-Level Comprehension, Paragraph-Level Comprehension |
| **COMPUTER-ADAPTIVE ASSESSMENT WITH CBM** | | | | |
| FastBridge | Word Segmenting | Sentence Reading | | Oral Reading Fluency, Comprehension |

NOTE: *AimswebPlus is administered by computer, but is a fixed-form assessment. **MAP also includes items related to standards for writing, capitalization, writing process, and grammar.

**Table 6. Classifications Used in First Grade**

| Screener | Lowest Classification | Middle Classification(s) | Highest Classification |
|---|---|---|---|
| Acadience Reading K–6 | Well Below Benchmark (Likely to need intensive support) | Below Benchmark (Likely to need strategic support) | Well Below Benchmark (Likely to need core support) |
| DIBELS 8th Edition | Well Below Goal (At-risk) | Below Goal | Goal or Above Goal (On track) |
| AimswebPlus Reading | Well Below Average | Below Average/Average/ Above Average | Well Above Average |
| FastBridge* | 350 | | 750 |
| MAP | Low | Low-Average/Average/ High-Average | High |
| Star | Emergent Readers | Transitional Readers | Probable Readers |

NOTE: *See Reading Score Guide for level descriptions: https://fastbridge.illuminateed.com/hc/en-us/article_attachments/1260801069370

range of 3–5 levels of performance. They are different in terms of their implications for what counts as good reading. Aimsweb and Map relate to averages, while DIBELS and Acadience relate to benchmarks, and Star has specialized terms unique to their own tools to classify readers. Again, the goals of meeting benchmarks in development vs. predicting scores on standardized tests are evident across testing systems. This, combined with the differences in areas assessed, creates a varied picture of what counts as a good reader according to different screeners.

## Discussion and Conclusion

The current legislative focus on evidence-based or research-based assessment tools has not translated into transparency about the relative psychometric quality of available tools. Rather, a range of reliability and validity are reported within and across assessments. Though summaries and promotional materials indicate these are all viewed as adequate, the standards for different kinds of assessments are different, and thus adequate psychometric properties may not all be equal. As a result, presence on a state list is taken as the standard for selection, rather than some feature of the assessment itself. Yet, presence on a state list may change year to year without a change in the assessment. For example, since 2019, Nebraska has dropped four tests and added two, and noted updated versions of two tests on their list. Also since 2019, Connecticut dropped all three of the computer-adaptive options, leaving districts that had recently transitioned to these in the lurch.

In addition to changes over time, there are significant differences across state-approved lists. Though the average number of approved screeners is five, some states have only one, and others have more than 30 options on their lists. Moreover, each state has their own criteria or guidance for screener selection. Across states these criteria are expressed as rubrics, checklists, or lists of characteristics. In some cases, state departments of education make their own selections. In other cases, they oversee the creation of a committee or task force with varied composition to make selections. The range of criteria and processes leads to variation in approved screeners across states, and differences in the way reading difficulty and risk for difficulty are assessed and understood across state lines.

In addition to the 'apples and oranges' (Lemke et al., 2024) included on state lists, districts may be tempted or compelled to use combinations or hybrids. For example, the only available screener that includes a test of rapid automatic naming is EarlyBird. In order for other assessments to be in compliance with Michigan's 2024 bill, a standalone rapid automatic naming test or subtest is needed. Combining tests does not necessarily make them more effective, but it does add time and cost in terms of time, training and data

management. Developing—let alone validating—homegrown assessments or combinations require time and resources that very few districts have. This means that state lists do not only include apples and oranges in terms of the difficulty comparing different test types, but also quinces and asian pears — assessment systems that represent hybrids of features from multiple tests.

It is possible that homegrown assessments or bespoke combinations of subtests are as good or better than commercial tools, but demonstrating this would be time and cost prohibitive. In the 42 states that specify which screeners are approved for use, the most efficient way to come into compliance is to select from the list. And, it is impossible to assume that all options are equal, or that the choice of assessment brand or type is inconsequential. Though it is reasonable that state lists are updated periodically, this sets up a high-risk situation for districts that do not want to have to abandon one and buy another in a few years. This risk may explain the clear preference for DIBELS as an option on state lists because of its longevity and familiarity — features of the brand, not the test.

### *Weighing the value of varied versions of reading*

Good screeners are generally regarded as quick and easy to give, while reliably predicting scores on other screeners, and/or other tests, including state tests used for accountability purposes. Comparison to other or future screeners confirms the validity of an assessment, but creates a circular logic to the meaning of a score. Any given score means the student is likely to score similarly on similar assessments. If a parent or educator wants to draw conclusions about reading in general, now or in the future, such scores are less relevant. The ability to predict a state test or a future screening score is helpful for planning purposes because it foreshadows signals from accountability measures about how well a student, class, cohort, school, or district will be rated. Put simply: Universal screeners are only as valid as a measure of reading as other standardized measures of reading. It might be more accurate to consider that we are screening for or predicting difficulty *on reading tests* rather than difficulty with reading in general.

In the 42 states that specify which screeners are approved for use, the most efficient way to come into compliance is to select from the list. And, it is impossible to assume that all options are equal, or that the choice of assessment brand or type is inconsequential.

The focus on efficiency may also have a cost beyond validity. Speed is required for high scores on all subtests with "fluency" in the title, and time limits are applied to many of the quest2ons in adaptive tests. This not only sends the message that speed is a key component of reading, it systematically disadvantages those with differences in processing speed and rate of oral language. The American Speech and Hearing Association (2014) warns that oral reading fluency tasks are not appropriate for students who stutter or have other difficulties related to speed of articulation. However, these guidelines are not referenced in any of the technical manuals that provide guidelines for oral reading fluency test administration. Moreover, they may be too conservative in their estimation of the students who are impacted by a bias towards reading, responding, and talking quickly (c.f. Zipoli & Ramachandar, 2024).

The rate of speech and length of acceptable pauses is culturally bound. However, studies that examine whether screeners are biased against language learners do not collect or report information on the languages or cultures of students in the study. In some cases, multilingual students are treated as a monolith, and in others, they are sorted by level of English proficiency. In both cases, researchers identify the need for different cut scores to avoid bias against students with varied English proficiency (Cummings et al., 2021; Keller-Margulis et al., 2022; Vanderwood et al., 2014 ).

### *Final thoughts*

It is easy to lose sight of what screeners indicate — risk or likelihood of a particular test score. This is perhaps the largest difference between tests designed to screen for difficulty (e.g. quickly identify risk) and other test types. When a screener finds that students are progressing as *expected*, that expectation references past (as in norm-referenced) or future (predictive) performance on other tests. Assessments using classification accuracy as an indicator of validity

are calibrated to measure how well the test sorts students into groups, which may be different from how well a child reads. Once we understand what a test compares—student performance to other student scores, or predicted future scores—we can consider its quality in context.

It is easy to assume that any test with a strong NCII rating for reliability and validity can be considered reliable and valid. However, none of the screeners studied here have absolute reliability or validity across subtests. They each exhibit a range, and some subtests are consistently more/less reliable and valid than others. This means districts that use some, but not all subtests, should note whether the specific subtests they are choosing meet criteria that align with what they value. It also means that screening data should never be discussed as an authoritative measure to which others must be compared. Some subtests are stronger predictors than others. In particular, those subtests that require tasks more similar to real reading, such as Oral Reading Fluency, are consistently better predictors than others. The metaphor of a mirror may be useful in this case because of the ways that slight imperfections in a mirror can distort — especially at the edges of a distribution.

Finally, it is interesting to note that multiple assessments meet state criteria and earn similar NCII ratings. The differences between subtests of a screener may actually be larger than differences between screeners. It is difficult to tell because no two assessments use the same indicators of quality in their technical manuals. And no two tests claim to measure exactly the same things (see Tables 5 and 6). This is because the tests themselves are different—with different "theories of the test"—what is measured, why, and how. Choosing from among these options is not an exercise in selecting the best of a kind. On the surface, they are all equal, and beneath the surface, they vary in different ways.

Therefore decisions about which screener to purchase are likely to be influenced by factors outside the quality of the test or theory of the test, including pricing, packaging, and reputation.

## *References*


Acadience Learning. (2020). *Acadience reading K–6 technical manual.* https://acadiencelearning.org/wp-content/uploads/2020/01/Acadience_Reading_K-6_Technical_Manual.pdf

Clay, M. M. (2005). *Literacy lessons designed for individuals part one: Why? when? and how?* Heinemann.

Clay, M. M. (2019). *An observation survey of early achievement* (4th ed.). Heinemann.

Connor, C. M., Alberto, P. A., Compton, D. L., & O'Connor, R. E. (2014). *Improving reading outcomes for students with or at risk for reading disabilities: A synthesis of the contributions from the Institute of Education Sciences Research Centers* (NCSER 2014-3000). National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/

Cummings, K. D., Smolkowski, K., & Baker, D. L. (2021). Comparison of literacy screener risk selection between English proficient students and English learners. *Learning Disability Quarterly: Journal of the Division for Children with Learning Disabilities*, *44*(2), 96–109. https://doi.org/10.1177/0731948719864408

Decoding Dyslexia. *What is decoding dyslexia?* Retrieved October 7, 2024, from https://www.decodingdyslexia.net/

Illuminate Education. (2024). *Reading assessment: The key to science-based reading instruction. Reading assessment FAQs: What makes a qualilty reading assessment tool?* https://www.illuminateed.com/products/fastbridge/reading-assessment/

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice: A Publication of the Division for Learning Disabilities, Council for Exceptional Children*, *24*(4), 174–185. https://doi.org/10.1111/j.1540-5826.2009.00291.x

Keller-Margulis, M. A., Matta, M., Landry Pierce, L., Zopatti, K., Reid, E. K., & Schanding, G. T. (2022). A comparison of reading screeners in kindergarten: The Texas Primary Reading Inventory and Acadience Reading with English learners and monolingual English speakers. *Assessment for Effective Intervention: Official Journal of the Council for Educational Diagnostic Services*, 153450842211335. https://doi.org/10.1177/15345084221133559

Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (Eds.). (2014). An introduction to universal screening in educational settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 3–16). American Psychological Association. https://doi.org/10.1037/14316-001

Klingbeil D., Nelson P., Van Norman E., & Birr C. (2017). Diagnostic accuracy of multivariate universal screening procedures for reading in upper elementary grades. *Remedial and Special Education*, *38*(5), 304–320.

Lemke, M., Murphy, D., Chow, A., & Acuña, A. (2024, March 13). *Comparing early literacy assessments: What really matters.* WestEd. https://www.wested.org/blog/comparing-early-literacy-assessments-what-really-matters/

Mellard, D.F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research & Practice, 24*(4), 186–195. https://doi.org/10.1111/j.1540-5826.2009.00292.x

National Center on Improving Literacy. (2019). *Best practices in universal screening.* Retrieved from https://www.improvingliteracy.org/brief/best-practices-universal-screening

National Center on Intensive Intervention. (2024, January). *Academic screening tools chart*. Retrieved October 27, 2024, from https://intensiveintervention.org/resource/academic-screening-tools-chart

North American Trainers Group. (2015). *Rationales and guidelines for selecting the lowest-achieving first-grade students for Reading Recovery*. https://readingrecovery.org/wp-content/uploads/2016/12/student_selection_guidesheet_05-15-15.pdf

Norton, E. S., & Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annual Review of Psychology, 63*, 427–452. https://doi.org/10.1146/annurev-psych-120710-100431

NWEA. (2019). *MAP growth technical report.* https://www.nwea.org/uploads/2021/11/MAP-Growth-Technical-Report-2019_NWEA.pdf

Odegard, T. N., Farris, E. A., Middleton, A. E., Oslund, E., & Rimrodt-Frierson, S. (2020). Characteristics of students identified with dyslexia within the context of state legislation. *Journal of Learning Disabilities*, *53*(5), 366–379. https://doi.org/10.1177/0022219420914551

Pearson (2017). *AimswebPlus technical manual.* https://www.marshfield-schools.org/cms/lib/WI01919828/Centricity/Domain/82/Plus%20Technical%20Manual.pdf

Renaissance Learning. (2021). *Star Assessments for eary literacy technical manual*. https://renaissance.widen.net/view/pdf/yp69mwijgt/SELRPTechnicalManual.pdf?t.download=true&u=zceria

Schwartz, S. (2022, July 20). *Which states have passed "science of reading" laws? What's in them?* Education Week. https://www.edweek.org/teaching-learning/which-states-have-passed-science-of-reading-laws-whats-in-them/2022/07

Torgesen, J. K. (2004). Preventing early reading failure. *American Educator, 28*(3), 6–19.

University of Oregon. (2020). *Dynamic indicators of basic early literacy skills (DIBELS): Technical manual* (8th ed.). https://dibels.uoregon.edu/sites/default/files/DIBELS8-TechnicalManual_04152020.pdf

Vanderwood, M. L., Tung, C. Y., & Checca, C. J. (2014). Predictive validity and accuracy of oral reading fluency for English learners. *Journal of Psychoeducational Assessment*, *32*(3), 249–258. https://doi.org/10.1177/0734282913502937

VanMeveren, K., Hulac, D., & Wollersheim-Shervey, S. (2020). Universal screening methods and models: Diagnostic accuracy of reading assessments. *Assessment for Effective Intervention, 45*(4), 255–265. https://doi.org/10.1177/1534508418819797

Walker, H. M., Small, J. W., Severson, H. H., Seeley, J. R., & Feil, E. G. (2014). Multiple-gating approaches in universal screening within school and community settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 47–75). American Psychological Association. https://doi.org/10.1037/14316-003

Woulfin, S. L., & Gabriel, R. E. (2020). Interconnected infrastructure for improved reading instruction. *Reading Research Quarterly*, *55*(S1), S109–117. https://doi.org/10.1002/rrq.339

Zipoli, R. P., & Ramachandar, S. (2024). Oral reading assessment: Four conditions where caution is warranted. *Assessment for Effective Intervention: Official Journal of the Council for Educational Diagnostic Services, 49*(3), 171–176. https://doi.org/10.1177/15345084231220526