Georgia Journal of Literacy
2024, Vol. *46*(1), pp. 4–19
https://doi.org/10.56887/galiteracy.133
ISSN: 2833-7611

Research Article

A Psychometric Review of Universal Reading Screeners Approved by the State Board of Education

Lindee Morgan
Sandra Dunagan Deal Center for Early Language and Literacy, Milledgeville, Georgia and Department of Professional Learning and Innovation, Georgia College and State University, Milledgeville, GA

Joseph Wenke
Sandra Dunagan Deal Center for Early Language and Literacy, Milledgeville, GA

Kristina Dandy
Department of Psychological Science, Georgia College and State University, Milledgeville, GA

ABSTRACT

In 2023, the Georgia Legislature passed the Georgia Early Literacy Act (HB 538), representing a sweeping reform effort to improve the quality of early reading instruction in the state. HB 538 requires schools to screen children in kindergarten through third grade three times yearly. Related to this requirement, HB 538 requires that the State Board of Education (SBOE) approve a list of universal reading screeners that can: 1) provide relevant information to target instruction, 2) measure foundational literacy skills, 3) identify students who are struggling to acquire reading skills, and 4) be used to monitor progress. The purpose of this review is to provide a supplement to the SBOE's approved list so that Local Education Agencies (LEAs) can assess the relative psychometric strength of each screener as they select the most appropriate screener for the students they serve. We compiled information regarding each screener's reliability, validity, sensitivity, and specificity to create an exposition of their strengths and weaknesses. We found that GaDOE's approved list contains numerous tools with acceptable psychometric properties; however, large variability in the amount of psychometric data available for each screener is problematic. LEAs are recommended to consider psychometric strength as a critical factor when selecting an early literacy screener.

KEYWORDS

psychometric review; universal reading screeners

In 2023, the Georgia Legislature passed the Georgia Early Literacy Act (HB 538), representing a sweeping reform effort to improve the quality of early reading instruction in the state. Among other aspects of the law, HB 538 requires schools to screen children in kindergarten through third grade three times yearly. Related to this requirement, HB 538 requires that the State Board of Education (SBOE) approve a list of universal reading screeners that can: 1) provide relevant information to target instruction, 2) measure foundational literacy skills, 3) identify students who are struggling to acquire reading skills, and 4) be used to monitor progress.

CONTACT **Dr. Lindee Morgan**, Executive Director, Sandra Dunagan Deal Center for Early Language and Literacy, Professor of Education, Department of Professional Learning and Innovation, Georgia College and State University, Milledgeville, GA; email lindee.morgan@gcsu.edu. **Joseph Wenke**, Data Analyst I, Sandra Dunagan Deal Center for Early Language and Literacy, Milledgeville, GA; email joseph.wenke@gcsu.edu. **Dr. Kristina Dandy**, Professor of Psychology, Department of Psychological Science, Georgia College and State University, Milledgeville, GA; email kristina.dandy@gcsu.edu.

The Georgia Department of Education's (GaDOE) policy division coordinated a Request for Information (RFI) process beginning in May 2023. The RFI application required vendors to include evidence in several areas, including how their screener addresses the requirements listed in HB 538 as indicated above. Following this, publishers of screeners prepared and submitted information about their screener to the SBOE. The SBOE approved a list of 16 screeners on July 19, 2023, and shortened this list on February 22, 2024. The current approved list can be found here.

The purpose of this psychometric review is to provide a supplement to the SBOE's approved list so that Local Education Agencies (LEAs) can assess the relative psychometric strength of each screener as they select the most appropriate screener for the students they serve. This independent review is meant to clarify several psychometric properties of each approved screener and provide LEAs with additional context regarding the tools included in this list.

Literature Review

Understanding the psychometric composition of a screener is critical when determining what populations a screener will effectively target. Psychometrics enables us to analyze the instruments we use to measure behaviors and traits; it also provides us with objective rules for scoring the results of tests (Raykov & Marcoulides, 2011). It is important to note that screeners differ from childhood assessments. Universal reading screeners identify students in need of additional evaluation, while assessments give insight into specific abilities and competencies (Moodie et al., 2014).

Psychometric Constructs

Reliability. Reliability is an index of whether students' scores on the screener will be stable despite extraneous factors, including when, who administers it, and where it is administered (Moodie et al., 2014). Reliability is impacted by variables such as test length, homogeneity of items, test-retest interval, variability of scores, student guessing, testing situation variance, and sample size (Sattler, 2020). For a psychometric test to be supported by evidence of reliability, the measure must be consistent across raters, time, and items (White et al., 2022).

While many reliability indices are available, the most common types used in our review were interrater reliability, test-retest, and internal consistency. Interrater reliability indicates whether the test's scores will vary when assessed by different raters (Cook & Beckman, 2006). Interrater reliability can demonstrate the objectivity of an assessment's scores (Sattler, 2020). Test-retest reliability demonstrates that an assessment yields stable results when administered to an individual at two or more time points (Sattler, 2020). Internal consistency demonstrates that the items in a test measure the same construct or concept or that the items in the test are homogenous (Cronbach, 1951; Tavakol & Dennick, 2011).

Validity. Validity is a measurement of "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (American Educational Research Association [AERA], 1999, p. 9). That is, validity must be inferred from multiple sources of evidence and stated within the context of a specific purpose (Cronbach & Meehl, 1955). A screener may have strong evidence of validity for identifying dyslexia but weak validity in screening for difficulty in acquiring reading skills or vice-versa. At the same time, a screener may obtain perfect sensitivity by flagging every student it assesses as at-risk for reading difficulty. However, its specificity would become incredibly weak for the number of false positives. Thus,

multiple sources for evidence of validity are required to state that a tool has evidence of validity (Cronbach & Meehl, 1955). Criterion validity demonstrates that a screener is accurate and precise by measuring it against an already accepted assessment. Using accepted assessments as our criterion measure allows us to advance the field by expediting the review of new tools. Concurrent validity indicates the accuracy of an assessment by comparing its results to another well-tested assessment administered at about the same time. Predictive validity demonstrates a screener's ability to predict a child's scores on another well-tested assessment at a later date.

Sensitivity and Specificity. Sensitivity and specificity are additional forms of validity that indicate a measure's capacity to correctly identify which students are at risk and which are not (Council on Children, 2006; Swift et al., 2020). The sensitivity indicates a tool's accuracy in identifying students with or at risk for a condition (i.e., true positives), in this case, reading difficulty or dyslexia (Parikh et al., 2008). Specificity demonstrates the tool's capacity to accurately rule out students who are not at risk for a specific condition (i.e., true negatives). It is essential to acknowledge that sensitivity and specificity only apply to the tested group. Sensitivity and specificity do not provide the probability of an individual student's test result being correct; that probability is more appropriately assessed by positive predictive values and negative predictive values (Trevethan, 2017). Sensitivity focuses entirely on the percentage of the population with the condition caught by the screener; it is not impacted by false positives. Positive predictive values, on the other hand, show the percentage of true positives out of all positive test results (Trevethan, 2017). False positives are less concerning than false negatives when evaluating reading screeners because a false positive will only result in a student receiving extra assistance, while a false negative results in a student who needs assistance not receiving it (Classification Accuracy, n.d.). Adequate sensitivity and specificity are crucial to show that a screener appropriately identifies the students who need extra assistance without overburdening the screening and response system by flagging children who are not indeed at risk for reading difficulties.

Psychometric Review Process

The authors reviewed all 16 screeners approved by the SBOE to provide LEAs with an answer to the question: What are the relative psychometric strengths of the universal reading screeners approved by the Georgia SBOE? The review used data published by independent experts when available and information provided by the screeners' publishers. The publishers of each screener submitted a report to GaDOE in response to a call for universal reading screening tools for students K–3. These reports contained information about how each screener works, the domains it assesses, and evidence of its efficacy. The National Center for Intensive Intervention's (NCII) Academic Screening Technical Review Committee (TRC) reviewed eleven of the sixteen screeners. The TRC comprises individuals with expertise in measurement and research methodology in academic screening. In addition, the TRC included committee members with expertise in culturally and linguistically diverse groups. Members of the TRC evaluated screeners for classification accuracy, reliability, and validity. Screeners not evaluated by NCII's TRC included *Amira, Battelle Early Academic Survey, aimswebPlus, Predictive Assessment of Reading, MindPlay Universal Screener*, and *Exact Path Diagnostic Assessment.* For each screener, we utilized reports submitted to GaDOE by the publishers of each screener as part of the RFI process and searched for additional studies on the screeners. Although several screeners on our list were developed for use beyond the third grade (e.g., as high as grade 8 or 12), we restricted our review to grades K–3 to align with the requirements of HB 538.

Method

When evaluating each screener's psychometric strength, we focused specifically on metrics of reliability, validity, sensitivity, and specificity. These metrics provide robust indicators of a tool's value in educational settings, enabling the communication of meaningful information through precise psychological measurements (Sattler, 2020). We identified statistical tests performed in evaluating each screener and reported the strength of evidence each statistical test provided. Together, these metrics provide insight into whether an early literacy screener can accurately and consistently indicate a child's reading status.

Each screener is a norm-referenced tool (i.e., these tools compare each student to a sample population) using grade-level norms (Ornstein, 1993). NCII was the first source of information used for our evaluation. For screeners not evaluated by NCII, the primary source of information was publisher reports submitted to GaDOE. Additional information from publishers' websites, journal articles, and technical manuals was also used. In their reports to GaDOE, each screener's publisher determined acceptable cutoff scores for the psychometric tests they used. The cutoffs used in our review are applied to all screeners based on relevant literature and standard research guidelines. Thus, they may vary from what was used by the publisher and generally provide a more conservative assessment of each tool's performance.

Reliability

For reliability, this review focused on interrater reliability, test-retest reliability, and internal consistency metrics. NCII gave ratings of convincing evidence of reliability to each screener that demonstrated the following: a model-based approach to reliability, at least two types of reliability that are appropriate to the tool, and for each type of reliability, the median lower bound of the confidence interval around the estimate had to meet or exceed 0.70 (Academic Screening Tools, n.d.). NCII's ratings were applied separately for each grade level targeted by the tools evaluated.

Interrater Reliability. Interrater agreement was only reported on screeners evaluated by NCII. Therefore, our determination of acceptable levels of interrater agreement corresponds with that deemed by NCII. Interrater reliability can be demonstrated by percentage agreement, kappa, intraclass correlation coefficient, or product-moment correlation coefficient (Sattler, 2020).

Test-Retest. Correlation coefficients calculated from the test-retest reliability depend on the type of data used and can include Pearson's r or Spearman's p correlation coefficients. A test-retest coefficient below 0.5 is considered weak, 0.5 to 0.7 is moderate, while above 0.7 is strong, and above 0.9 is very strong (McDaniel & Ziniel, 2023).

Internal Consistency. Internal consistency can be shown using Cronbach's alpha, Omega, or split-half reliability; scores of 0.7 are considered good, and scores of 0.8 are considered excellent, but 0.9 or higher may suggest redundancy more than consistency (McDaniel & Ziniel, 2023). A score between 0.6 and 0.7 could be considered adequate in limited situations, but anything below 0.6 is considered poor reliability.

Other reliability tests used in reviewing the screeners but not included in our results include IRT-Score-based reliability and EFA/CFA Model-based coefficient Omega. While both tests are acceptable ways to measure reliability, too few screeners used these tests to justify including them in our results table.

Validity

For the current review, we include metrics of criterion validity: concurrent and predictive. A correlation between the two measures evaluates both. For predictive and concurrent validity, a median coefficient of 0.49 or less is considered weak, 0.5 to 0.69 is considered moderate, and anything over 0.7 is considered strong (McDaniel & Ziniel, 2023).

Sensitivity and Specificity. Acceptable sensitivity and specificity depend on what is being assessed and the population in which it is being assessed. Sensitivity and specificity are expected to vary with changes in the prevalence of the condition being screened for (Parikh et al., 2008). For example, the sensitivity of a screener meant to detect reading difficulty should be higher than a screener meant to detect dyslexia because reading difficulty is more prevalent than dyslexia (Catts et al., 2012; Yang et al., 2022). Scoring systems must attempt to minimize under-referring or over-referring, which is why sensitivity and specificity scores of 0.7–0.8 are generally acceptable (Council on Children, 2006). Sensitivity and specificity are reported with a range from 0 to 1, with 1 indicating perfect measurement (Swift et al., 2020). Given that NCII gives high ratings to sensitivities of 0.7 or greater and to specificities of 0.8 or greater (Classification Accuracy, n.d.), we modified the rating scale for the current review, citing sensitivity and specificity ratings of 0.8 and above as acceptable; this modification was made to highlight the importance of accurately identifying reading difficulties in K–3 children. Given the interdependence of these measures within the context in which they are assessed, interpretation of specific scores should be made with consideration for the purpose of the assessments.

Screener Ranking

Based on the information available to us and in examination of screener features and psychometric indices, we derived an informal coding system to generate relative rankings of the approved screeners. In this coding system, we ascribed weighted points for each area to derive a total score so that these tools could be considered relative to one another. Seven aspects were included in the coding system, including: (1) screener scope, (2) psychometric breadth, (3) reliability, (4) criterion validity, (5) sensitivity, (6) specificity, and (7) sensitivity and specificity composite.

The screener scope, determined by the domains that each screener assessed, was rated as a 0 or 1. The domains assessed by each screener were indicated by publishers in their reports to GaDOE; details can be found in Table 2. Screeners that assessed at least 12 domains received a score of 1. Psychometric breadth also rated a 0 or 1, was based on the range of grades psychometric data were provided for each screener. Screeners that provided data for at least three grades between K–3 received a score of 1. Reliability was also rated a 0 or 1, based on the abundance of assessments supported by evidence of reliability. Screeners that tested at least two types of reliability received a score of 1. Due to the nature of the data and its intended use in correctly identifying children with reading delays accurately, criterion validity was weighted more heavily than reliability at a maximum of 3 points. Points given for criterion validity were based on the cutoffs described above. Sensitivity weighting was calculated as a factor of the screener's reported sensitivity across K–3. We multiplied each screener's mean sensitivity by 6 to give it heavier weighting due to the relative importance of sensitivity as a psychometric feature for screeners in early education. Specificity was rated from 0–3, with 3 points for a specificity of 0.9 or more, 2 points for a specificity between 0.8 and 0.9, 1 point for a specificity between 0.7 and 0.8, and 0 points for a specificity of less than 0.7. A score of 0 was also given if specificity data were not provided. Finally, we calculated a mean sensitivity score and a mean specificity score and created

a composite score by adding them together for each screener. The exact composite score was added to the ranking score of each screener, for a minimum of 0 and a maximum of 2.

Using each of the aspects described above, we summed scores to rank each screener relative to one another and organized them into three categories: strong, moderate, and weak. The maximum possible score was 17 points. Scores above the median were categorized as strong. Scores below the median were categorized as moderate or weak. Screeners that received less than 60% of the maximum score were categorized as weak.

Results

The results of our review are presented in Tables 1–4. Table 1 provides an alphabetical listing of each screener, its publisher, and the grades for which the tool is intended. Table 1 also indicates whether the tool shows convincing evidence of reliability and validity for each grade analyzed. An acceptable reliability rating was required for a tool to be determined to have convincing evidence of reliability in each grade, and a moderate validity coefficient was required for a tool to be determined to have convincing evidence for validity. When possible, NCII's judgment on evidence was used. For screeners not evaluated by NCII, cutoff points for reliability and validity were used, as indicated in the section above.

Two screeners (*iSTEEP and MindPlay Universal Screener*) did not provide grade-specific metrics of reliability and validity. One notable finding is that nine of the sixteen tools do not have convincing evidence for either reliability or validity at kindergarten. Four of these tools do not have strong evidence for both reliability and validity at kindergarten. Additional information on each screener's supporting evidence can be found in Table 3. It is also worth noting that for reliability and validity to have real meaning, the intended population must be the same as the group tested in the tool's development (Moodie et al., 2014); however, an assessment of the test population was beyond this review's scope. Generally, this information can be found on publisher's websites or in screener technical manuals. Table 1 also states whether a screener requires administrator/teacher training or technology to administer, with most tools requiring both.

Table 1: Overview of Literacy Screeners Approved by the SBOE

Measure Name	Vendor	Grades Developed For	Convincing Evidence of Reliability by Grade	Convincing Evidence of Validity by Grade	Administrator Training Required	Technology Required for Administration
Acadience Reading K–6	Acadience Learning, Inc.	K–6	K–6	1, 2, 4, 5, 6	Yes	No
aimswebPlus	Pearson	K–8	K–8	K–8	Yes	Yes
Amira	Houghton Mifflin Harcourt	K–3	K–3	K–3	No	Yes
Battelle Early Academic Survey	Riverside Assessments	K–2	K–2	K–2	Yes	Yes
Classworks Reading Universal Screener	Classworks	K–10	2–8	2–8	Yes	Yes

EasyCBM for Reading	Riverside Assessments	K–8	K–5	2, 3, 4, 5	Yes	Yes
Exact Path Diagnostic Assessment	Edmentum	K–3	K–3	K–3	Yes	Yes
FastBridge aReading	Renaissance Learning	K–8	K–8	2–8	Yes	Yes
i-Ready Assessment for Reading	Curriculum Associates	K–8	K–8	K–8	Yes	Yes
ISIP Reading with RAN and ORF	Istation	K–8	K–8	K–8	Yes	Yes
iSTEEP	iSTEEP, LLC	K–12	N/A*	N/A*	No	Yes
MAP Reading Fluency	NWEA	K–3	K–3	1, 2, 3	Yes	Yes
mCLASS	Amplify Education, Inc.	K–8	K–8	K–5	Yes	No
MindPlay Universal Screener	MindPlay	K–12	N/A*	N/A*	No	Yes
Predictive Assessment of Reading	Red E Set Grow	K–3	K–3	1–3	Yes	Yes
Star Assessments	Renaissance Learning	K–3	1–3	K–3	Yes	Yes

Note. *Not specified by grade.

Table 2 (a & b) lists the domains assessed by each screener. Table 2 is split into two parts for readability, with each part including eight screeners. Screener domains are sets of related skills or information classified together for assessment purposes. GaDOE provided two categories of screener domains: foundational literacy skills and characteristics of dyslexia. GaDOE provided these for publishers to indicate what screeners purportedly assess. The grades at which each domain is assessed are also indicated. Although GaDOE has listed each of these domains separately, the domains are not necessarily mutually exclusive. With very few exceptions, this group of screeners assesses each of the domains listed at K–3. It is worth noting that the Predictive Assessment of Reading evaluates only one out of the seven domains of dyslexia.

Table 2(a): Domains Assessed

	Predictive	*Acadi-ence*	*Aimsweb Plus*	*Amira*	*Battelle*	*Class works*	*Easy CBM*	*Exact Path*
Foundational Literacy Skills								
Phonological Awareness	K–3	K–1	K–1	K–3	K–2	K–2	K–1	K–1
Phonemic Awareness	K–3	K–1	K–1	K–3	K–2	K–2	K–1	K–1
Phonics	K–3	K–3	K–1	K–3	K–2	K–3	K–1	K–3
Fluency	K–3	1–3	K–3	K–3	K–2	Not assessed	K–3	K–3
Vocabulary	K–3	K–3	K–3	K–3	Not assessed	2–3	2–3	K–3
Reading Comprehension	K–3	1–3	2–3	K–3	Not assessed	1–3	2–3	K–3
Spelling	K–3	K–1	K–3	K–3	Not assessed	3	Not assessed	K–3
Oral Language	K–3	K–3	K–3	K–3	K–2	K–3	K–3	K–3
Intersection of Reading and Writing	K–3	K–1	1–3	Not assessed	K–2	1–3	3	K–3
Characteristics of Dyslexia								
Sound Symbol Recognition	Not assessed	K–2	K–1	K–3	K–2	K–2	K–1	K
Alphabet Knowledge	Not assessed	1,2	K–3	K–3	K–2	K–3	K	K–3
Decoding Skills	Not assessed	K–3	K–1	K–3	K–2	K–3	K–1	K–3
Encoding Skills	Not assessed	K–1	K–3	K–3	Not assessed	K–3	K–1	K–3
RAN	K–3	K–1	K–3	K–3	K–2	Not assessed	K–1	K–3
Accuracy of Word Reading	Not assessed	1–3	K–3	K–3	K–2	K–3	1–3	K–3
Sight Word Reading Efficiency Skills	Not assessed	1–3	K–3	K–3	K–2	K–2	K–1	K–1

Table 2(b): Domains Assessed

	FastBridge aReading	*i-Ready Assess-ment for Reading*	*ISIP Readingwi th RAN and ORF*	*iSTEEP*	*MAP Reading Fluency*	*mCLASS*	*MindPlay Universal Screener*	*Star Assess-ments*
Foundational Literacy Skills								
Phonological Awareness	K–1	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Phonemic Awareness	K–1	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Phonics	K–3	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Fluency	K–1	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Vocabulary	K–3	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Reading Comprehension	K–3	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Spelling	K–3	1–3	K–3	K–3	K–3	K–3	K–3	K–3
Oral Language	K–1	1–3	K–3	K–3	K–3	K–3	K–3	K–3
Intersection of Reading and Writing	K–3	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Characteristics of Dyslexia								
Sound Symbol Recognition	K–1	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Alphabet Knowledge	K	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Decoding Skills	K–1	K–3	K–3	K–3	K–3	K–3	Not assessed	K–3
Encoding Skills	K–3	K–3	K–3	K–3	K–3	K–3	K–3	K–3
RAN	K	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Accuracy of Word Reading	1–3	K–3	K–3	K–3	K–3	K–3	K–3	K–3
Sight Word Reading Efficiency Skills	K–1	K–3	K–3	K–3	K–3	K–3	K–3	K–3

Table 3 summarizes the strength of each psychometric index evaluated as well as the source of information for these metrics. Specifically, Table 3 identifies the reliability, criterion validity, sensitivity, and specificity of each screener, specifically in grades K–3. While publishers may have reported these results for specific grade levels, the metrics in Table 3 are based on an average of

scores provided from K–3. These metrics provide insight as to whether a screener can accurately and consistently indicate children's performance in the domains listed in Table 2. The metrics in Table 3 were analyzed against specific cut points to represent varying levels of reliability, criterion validity, sensitivity, and specificity. Our classifications represent the inferred strength of available evidence for each of the aforementioned psychometric indices. These are not absolute judgements; thus, exact values are not included in Table 3. A key is provided in the table that indicates these cut points, from low to acceptable, weak to strong, and weak to acceptable.

Table 3: Reliability, Criterion Validity, Sensitivity, and Specificity of Screeners at Grades K–3

Screener Name	**Source**	**Reliability**			**Validity**		
		Interrater	*Test-Retest*	*Internal Consistency*	*Criterion*	*Sensitivity*	*Specificity*
Acadience Reading K–6	Intensive intervention	Acceptable	Acceptable	Acceptable	Strong	Weak	Acceptable
aimswebPlus	Pearson	Not assessed	Acceptable*	Acceptable	Moderate	Acceptable	Acceptable
Amira	HMHCO Amira Learning: Research Evidence Base	Not assessed	Acceptable	Acceptable	Strong	Acceptable	Acceptable
Battelle Early Academic Survey	Riverside	Not assessed	Acceptable	Acceptable	Strong	**	**
Classworks Reading Universal Screener	Intensive intervention	Not assessed	Acceptable	Acceptable	Strong	Weak	Acceptable
Easy CBM for Reading	Intensive intervention	Not assessed	Acceptable	Not assessed	Moderate	Weak	Acceptable
Exact Path Diagnostic Assessment	Edmentum Research	Not assessed	Not assessed	Acceptable*	Strong	Acceptable	Acceptable
FastBridge aReading	Intensive intervention	Not assessed	Acceptable	Not assessed	Strong	Acceptable	Acceptable
i-Ready Assessment for Reading	Intensive intervention	Not assessed	Acceptable	Acceptable*	Strong	Acceptable	Acceptable
ISIP Reading with RAN and ORF	Padlet	Not assessed	Acceptable	Acceptable*	Strong	Acceptable	Weak
iSTEEP	Intensive intervention	Acceptable	Acceptable	Not assessed	Moderate	Weak	Acceptable
MAP Reading Fluency	Intensive intervention	Not assessed	Acceptable	Acceptable*	Moderate	Weak	Weak
mCLASS	Intensive intervention	Not assessed	Acceptable*	Not assessed	Strong	Weak	Acceptable

MindPlay Universal Screener	MindPlay Education	Not assessed	Acceptable	Not assessed	Moderate	**	**
Predictive Assessment of Reading	PAR Technical Manual	Not assessed	Acceptable	Acceptable	Strong	Acceptable	Acceptable
STAR Assessments	Star Assessment	Not assessed	Acceptable	Acceptable	Moderate	Acceptable	Acceptable

Note. Numerical ratings below represent median coefficient/alpha ratings. Cells showing the highest rating in each category are highlighted. *Marginal reliability was used as a metric of internal consistency, or alternate form or delayed alternate form reliability was used as a metric of retest reliability. Acceptable level of Interrater, Test-Retest, and Internal Consistency was identified as Acceptable (> 0.7); Low (< 0.7). Ratings of Criterion Validity: Strong (> 0.7), Moderate (> 0.5, < 0.7), Weak (< 0.5). Ratings of Sensitivity and Specificity: Acceptable (> 0.8), Weak (< 0.8). **Sensitivity and specificity were not tested for this screener.

It is important to note that several screeners did not assess two or more of the metrics examined in our review (see *Battelle Early Academic Survey, Exact Path, FastBridge, mCLASS*, and *MindPlay Universal Screener*). Of all of these, *MindPlay Universal Screener* provided the least evidence with information for only two out of six psychometric indices. Two of the sixteen tools (*Battelle Early Academic Survey* and *MindPlay Universal Screener*) did not publish information on sensitivity or specificity.

All screeners reporting indices of reliability performed within acceptable levels, and all screeners reporting criterion validity had either moderate or strong ratings. Regarding criterion validity, it is worth noting that multiple screeners used *MAP Growth* and *MAP* as their measure of comparison, which have been shown to be valid tools and as a result are acceptable criterion measures. *MAP Growth and MAP* are not the same tool as *MAP Reading Fluency* included in our review. Of the fourteen screeners reporting sensitivity, six received a determination of weak (*Acadience, Classworks, Easy CBM, iSTEEP, MAP Reading Fluency*, and *mCLASS*). Only two screeners were determined to have weak specificity (*ISIP Reading* and *MAP Reading Fluency*).

Inconsistencies in reporting were apparent for two screeners. Two screeners appear to have been developed and normed at a narrower grade range than their report to GaDOE suggests. According to the NCII report, *Classworks* was normed on 2nd-8th grade. However, their reporting to GaDOE indicated that their screener is appropriate for K–10th grade. Similarly, *EasyCBM* was reportedly normed on 3rd-5th grade, but their reporting to GaDOE indicated that their screener is appropriate for K–8th grade. Caution is suggested in using tools where publishers may have used reduced rigor in evaluating and reporting.

Relative screener rankings are provided in Table 4. It is important to note that these rankings only compare the screeners approved by the SBOE. For example, a ‘weak’ designation indicates a tool’s relative psychometric standing to the other screeners on the approved list. It does not provide a comparison to all literacy screeners available on the market, including those submitted to GaDOE that were not approved for use.

Table 4: Relative Screener Rankings

Ranking	Screener
Strong (8)	*aimswebPlus* *Amira* *Classworks Reading Universal Screener* *Exact Path Diagnostic Assessment* *i-Ready Assessment for Reading* *ISIP Reading with RAN and ORF* *Predictive Assessment of Reading* *Star Assessments*
Moderate (5)	*Acadience Reading K–6* *FastBridge aReading* *iSTEEP* *MAP Reading Fluency* *mCLASS*
Weak (3)	*EasyCBM* *Battelle Early Academic Survey* *MindPlay Universal Screener*

The maximum possible score for any screener in our ranking was 17; the highest score achieved was 15.87 (*Exact Path Diagnostic Assessment*), and the lowest score received was 3 (*MindPlay Universal Screener*). Screeners in the strong category had at least 14 points in our ranking. The moderate category belongs to screeners that received between 10 and 14 points. Only three screeners received a 0 for psychometric breadth. Four screeners received 0 points for reliability due to only having tested one type of reliability. No screener received less than 2 points for criterion validity. The lowest specificity, averaged across K–3, for any screener that tested it and provided data was 0.71. Cutoffs for rankings of criterion validity and specificity can be found in the notes under Table 3.

Discussion

The authors conducted an independent review of universal literacy screeners approved by the SBOE to meet the screener requirements in HB 538. The review included a detailed summary of each tool's primary features, domains assessed, and evidence of psychometric strength as indicated by metrics of reliability, criterion validity, sensitivity, and specificity. The purpose of this review was to provide a supplement to the SBOE's approved list of screeners to aid LEAs in making an informed choice as to the most appropriate screener for the students they serve. Overall, our findings indicate that for K–3, most of the screeners assess all relevant early literacy domains as specified by GaDOE with acceptable levels of reliability and criterion validity where reported. The available evidence supporting each screener, along with the absence of psychometric evidence for some tools, allows us to discern which tools are supported by the strongest evidence of reliability and validity when identifying students at risk for reading difficulties. Given the information available, the eight tools with the strongest psychometric properties on the SBOE list of approved screeners are *aimswebPlus, Amira, Classworks Reading Universal Screener, Exact Path Diagnostic Assessment, i-Ready Assessment for Reading, ISIP Reading with RAN and ORF, Predictive Assessment of Reading*, and *Star Assessments*. Five tools, *Acadience Reading K–6,*

FastBridge aReading, iSTEEP, MAP Reading Fluency, and *mCLASS* were ranked as having moderate psychometric strength. In contrast, three tools cluster as having weaker psychometric profiles. These include *EasyCBM, Battelle Early Academic Survey*, and *MindPlay Universal Screener.* In consideration of these global groupings, a few issues should be taken into account. These are discussed in detail below.

Given that sensitivity and specificity should be interpreted together when determining the overall usefulness of a diagnostic test (Shreffler & Huecker, 2023), we identified the strongest screeners as those that demonstrate both acceptable sensitivity and specificity with an emphasis on sensitivity. Sensitivity is prioritized because it ensures accurate identification of children who need access to early reading interventions. Six screeners in our review demonstrated weak sensitivity. Tools with low sensitivity will fail to identify a higher percentage of children that are in need of additional instructional support. For the purposes of our review, we reported the average sensitivity of the tool across all grades assessed. The consequence is that the average can mask variability in sensitivity at different grade levels. While a screener might have strong sensitivity at specific grades, the weak sensitivity of a tool at any grade level should be considered a key factor in decision-making when selecting screeners.

One of the primary challenges of evaluating these screeners is inconsistency in available information. This inconsistency is found in both lack of information and discrepancies in reporting. Two screeners (*iSTEEP and MindPlay Universal Screener*) for example, did not provide grade-specific metrics of reliability and validity. Relatedly, two screeners (*Battelle Early Academic Survey* and *MindPlay Universal Screener*) did not provide evidence of sensitivity or specificity. There is an inherent problem with comparing tools lacking information to those that provided information that is less than compelling. Similarly, some inconsistency was noted with regard to NCII reporting on the grade levels the test was developed for versus what grades the publisher indicated the tool could be used. In both cases, it is important to consider that some publishers of tools are less rigorous in the evaluation of their screeners. For the purpose of this review, tools that presented thorough and consistent data were viewed more favorably than those that did not.

Another consideration is the matter of tools having variable performance at different grade levels. Nine out of the sixteen tools do not have convincing evidence for either reliability or validity for kindergarten. While this should be a concern of school districts, it is not surprising for screening tools to perform differently when administered across a multi-year age span. As children develop, their skills change at a rapid pace and certain screener items, or domains are likely to be more or less relevant given a child's developmental level. In literacy development, children in kindergarten present with a highly variable set of skills even within normal expectations. Additionally, kindergarten students undergo rapid acquisition of new skills within the school year. Thus, psychometric strength is more likely to be unstable at the early grades than the upper grades. Our review gave greater weight to tools demonstrating the greatest breadth of strong performance across grades.

The limited nature of this review is important to note, as it was conducted to provide a broad-based synopsis of the psychometric quality of early literacy screeners approved by the SBOE. This review was completed by the authors at the request of the Georgia Council on Literacy to respond to a specific need. Thus, it was conducted as robustly and thoroughly as was feasible within a relatively brief timeline (i.e., about two months). While the review includes ample detail, it was not conducted with the specificity and rigor that would be expected of a full-scale psychometric evaluation. As a result, some nuance and detail were beyond the scope of this project. For example, the review did not conduct an analysis of standardization populations, nor did it

include examination of the full scale of psychometric indices. Relatedly, we utilized two major sources (GaDOE RFI and NCII) to compile this review. Outside these two sources, there were a handful of additional publications used to gather information. Thus, there may be sources regarding these screeners that were not consulted for this review.

Finally, our review was conducted following a review by the GaDOE of a broader set of screeners submitted for consideration of approval. Thus, it is important for LEAs to consider that these tools represent a select set that are likely to be superior to other tools on the market. Thus, our rankings should be considered within this context as **relative only to one another** and not an absolute ranking of overall superiority or weakness.

Conclusion

This review was conducted to enable LEAs to determine which screeners are best suited for the students they serve. Our review demonstrates that GaDOE has selected a number of tools with acceptable psychometric properties enabling statewide implementation of meaningful screening of K–3 students as required by HB538. With proper utilization of these screeners, schools can accurately and consistently identify students in need of additional support. It is recommended that LEAs consider psychometric strength as delineated herein a critical factor when selecting an early literacy screener.

Key Takeaways for LEAs

- This review identified eight screeners (see Table 4) from the SBOE's approved list that present with superior psychometric features relative to the remaining eight screeners.
- This review was completed following a review by the GaDOE of a broader set of screeners submitted for consideration. Our rankings of strong, moderate, or weak should be considered within this context and as relative only to one another and not an absolute ranking of screener acceptability.
- The relative rankings provided for the sixteen screeners included in this review were derived from an examination of all screener characteristics and psychometric features available to us. We ranked screeners based on a weighted combination of factors (e.g., completeness of psychometric testing, robustness across grades, adequate sensitivity).
- This review was conducted to provide a broad-based synopsis of the psychometric quality of early literacy screeners approved by the SBOE and was prepared within a very limited time frame. While this review includes ample detail, it was not conducted with the specificity and rigor that would be expected of a full-scale psychometric evaluation.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. https://www.aera.net/Portals/38/1999%20Standards_revised.pdf

Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology*, *104*(1), 166–181. https://doi.org/10.1037/a0025323

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*(2), 101–192. https://doi.org/10.1016/j.amjmed.2005.10.036

Council on Children with Disabilities, Section on Developmental Behavioral Pediatrics, Bright Futures Steering Committee, & Medical Home Initiatives for Children with Special Needs Project Advisory Committee. (2006). Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics, 118*(1), 405–420. https://doi.org/10.1542/peds.2006-1231

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

McDaniel, C. E., & Ziniel, S. I. (2023). A psychometrics primer: The basics all hospitalists should know. *Hospital Pediatrics*, *13*(3), e63–e68. http://doi.org/10.1542/hpeds.2022-006951

Moodie, S., Daneri, M. P., Goldhagen, S., Halle, T., Green, K., & LaMonte, L. (2014). *Early childhood developmental screening: A compendium of measures for children ages birth to five* (OPRE Report 2014-11). United States, Administration for Children and Families, Office of Planning, Research and Evaluation. http://www.acf.hhs.gov/sites/default/files/opre/compendium_2013_508_compliant_final_2_5_2014.pdf

National Center for Intensive Intervention. (NCII). (n.d.). *Academic screening tools chart rating rubrics.* https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_2020-06-30.pdf

National Center for Intensive Intervention. (NCII). (n.d.). *Classification accuracy.* https://intensiveintervention.org/sites/default/files/Classification_Accuracy_508.pdf

Ornstein, A. C., (1993). Norm-referenced and criterion-referenced tests: an overview. *NASSP Bulletin*, *77*(555), 28–39. https://doi.org/10.1177/019263659307755505

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45–50. https://doi.org/10.4103/0301-4738.37595

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge/Taylor & Francis Group.

Sattler, J. M. (2020). *A primer on statistics and psychometrics. Assessment of children: Cognitive foundations and applications*. Jerome M. Sattler, Publisher, Inc.

Shreffler, J., & Huecker, M. R. (2023). *Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/books/NBK557491/

Swift, A., Heale, R., & Twycross, A. (2020). What are sensitivity and specificity? *Evidence-Based Nursing*, *23*(1), 2–4. https://doi.org/10.1136/ebnurs-2019-103225

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2,* 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5*, Article 307. https://doi.org/10.3389/fpubh.2017.00307

White, R. F., Braun, J. M., Kopylev, L., Segal, D., Sibrizzi, C. A., Lindahl, A. J., Hartman, P. A., & Bucher, J. R. (2022). *NIEHS report on evaluating features and application of neurodevelopmental tests in epidemiological studies.* National Institute of Environmental Health Sciences. https://www.ncbi.nlm.nih.gov/books/NBK581902/

Yang, L., Li, C., Li, X., Zhai, M., An, Q., Zhang, Y., Zhao, J., & Weng, X. (2022). Prevalence of developmental dyslexia in primary school children: A systematic review and meta-analysis. *Brain Sciences*, *12*(2), Article 240. https://doi.org/10.3390/brainsci12020240

Received: December 18, 2023 | Accepted: February 22, 2024 | Published: May 15, 2024