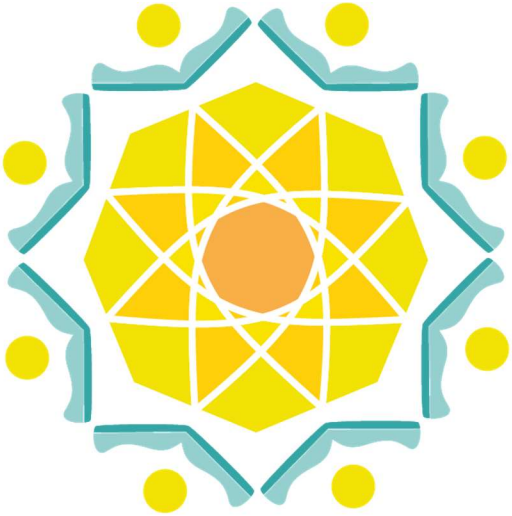
The Sandra Dunagan Deal Center for Early Language and Literacy



SANDRA DUNAGAN DEAL
CENTER FOR
EARLY LANGUAGE
AND LITERACY
AT GEORGIA COLLEGE
& STATE UNIVERSITY

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

Dr. Lindee Morgan, Joseph Wenke, and Dr. Kristina Dandy

2023

Introduction and Context

In 2023 the Georgia Legislature passed the Georgia Early Literacy Act (HB 538). HB 538 represents a sweeping reform effort to improve the quality of early reading instruction in the state. This legislation requires that schools screen children in kindergarten through third grade three times each school year. Related to this requirement, HB 538 requires that the State Board of Education shall approve a list of universal reading screeners which 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 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, a list of proposed screeners was prepared and submitted to the State Board of Education (SBOE). The list of screeners was approved by the SBOE on July 19, 2023, and can be found here.

Purpose of Review

To provide Local Education Agencies (LEAs) with additional context regarding the tools included in this list, the Sandra Dunagan Deal Center for Early Language and Literacy (Deal Center) conducted an independent review to clarify several psychometric properties of each approved screener. That is, our review provides an elucidation of available information regarding each tool's reliability, validity, sensitivity, and specificity. Definitions for each of these metrics are provided below and in Table 1. The purpose of this review is to provide a supplement to the SBOE's approved list so that LEAs can assess the relative psychometric strength of each screener as they select the most appropriate screener for the students they serve.

Table 1. Definitions of psychometric constructs.

Reliability: an index of whether the screener will produce consistent results across time, raters, and items
Validity: an index of how accurately and precisely a screener measures what it purports to measure
Sensitivity: an index of a screener's accuracy in identifying students at-risk
Specificity: an index of a screener's accuracy in ruling out students **not** at-risk

Psychometric Review Process

The Deal Center sought to review all sixteen screeners approved by the SBOE. The review used data published by independent, expert review when available, as well as 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. Eleven of the sixteen screeners were reviewed by the National Center for Intensive Intervention's (NCII) Academic Screening Technical Review Committee (TRC). The TRC is comprised of professors and program experts with background in measurement and research methodology in academic screening. In addition, the NCII included committee members with expertise on 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. For each of these screeners, 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.

Metrics Evaluated

When evaluating the psychometric strength of each screener 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 communication of meaningful information through precise psychological measurements (Sattler, 2020). We identified statistical tests that were performed in evaluating each screener and reported the strength of evidence that each statistical test provided. Together, these metrics provide insight as to 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 predefined population) using grade level norms. 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 the tool's performance.

Reliability is an index of whether the screener will produce consistent results over time, despite extraneous variables, including when the screener is administered, who administers the screener, 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, guessing, testing situation variance, and sample size (Sattler, 2020). For a psychometric test to be considered reliable, it must be consistent across raters, across time, and across items (White et al., 2022). This review focused on metrics of interrater reliability, test-retest reliability, and internal consistency. 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 table. **Interrater reliability** provides an indication as to whether the test will yield similar scores when evaluated by two or more raters. Interrater reliability is used to show objectivity of the assessment and can be demonstrated by percentage agreement, kappa, intraclass correlation coefficient, or product-moment correlation coefficient (Sattler, 2020). Interrater agreement was only reported on screeners evaluated by NCII. Thus, our determination of acceptable levels of interrater agreement corresponds with that deemed by NCII. **Test-retest reliability** demonstrates

that an assessment yields stable results when administered to a group at two time points. 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 and Ziniel, 2023). **Internal consistency** shows that the items in a test are measuring the same construct or concept (Tavakol and Dennick, 2011). 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 and 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.

Validity is an indicator that a tool accurately and precisely measures what it purports to measure. **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. For the current review, we include metrics of criterion validity: concurrent and predictive. **Concurrent validity** shows 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. Both are evaluated by a correlation between the two measures. 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 and Ziniel, 2023).

Sensitivity and specificity are another form of validity that indicate a measure's capacity to correctly identify which students are at-risk and which students are not. **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. **Specificity** demonstrates the tool's capacity to accurately rule out students who are not at risk for a specific condition (i.e., true negatives). 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 (n.d.). No screener can correctly identify at-risk students 100% of the time. Sensitivity and specificity are crucial to show that a screener is identifying the students who need extra assistance without overburdening the screening and response system by flagging children who are not truly at risk for reading difficulties. Sensitivity and specificity are both reported with a range from 0 to 1, with 1 indicating perfect measurement. Acceptable sensitivity and specificity are dependent 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 what is 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). NCII gives high ratings to sensitivities of 0.7 or greater and to specificities of 0.8 or greater (n.d.). We modified this rating scale for the current review, citing sensitivity and specificity ratings of .8 and above as acceptable. This was done to highlight the importance of accurately identifying reading difficulties in K-3 children. Given the interdependence of these measures within the context they are assessed, interpretation of specific scores should be done with consideration for the purpose of the assessments.

Results

The results of our review are presented in Tables 2-5. Table 2 gives an alphabetical listing of each screener, its publisher, and the grades for which the tool is intended. Table 2 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 judgement on evidence was used. For screeners

not evaluated by NCII, cutoff points for reliability and validity, indicated in the metrics evaluated section above, were used.

Two of the screeners (iSTEEP and MindPlay) did not provide grade-specific metrics of reliability and validity. One notable finding is that nine out 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 the supporting evidence each screener has can be found in Table 4. It is 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 test population was beyond the scope of this review. Generally, this information can be found on publisher's websites or in screener technical manuals. The table also states whether a screener requires administrator/teacher training or technology to administer with most tools requiring both.

Table 2. 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	2nd-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 3 (a & b) lists the domains assessed by each screener. Table 3 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 3(a). *Domains assessed.*

	Predictive Assessment of Reading	Acadience Reading K-6	aimswebPlus	Amira Screener	Battelle Early Academic Survey	Classworks Reading Universal Screener	EasyCBM for Reading	Exact Path Diagnostic Assessment
Foundational Literacy Skills								

Phonological Awareness	K-3	K,1	K,1	K-3	K,1,2	K,1,2	K,1	K,1
Phonemic Awareness	K-3	K,1	K,1	K-3	K,1,2	K,1,2	K,1	K,1
Phonics	K-3	K-3	K,1	K-3	K,1,2	K-3	K,1	K-3
Fluency	K-3	1,2,3	K-3	K-3	K,1,2	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	1,2,3	2,3	K-3	Not assessed	1,2,3	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,1,2	K-3	K-3	K-3
Intersection of Reading and Writing	K-3	K,1	1,2,3	Not assessed	K,1,2	1,2,3	3	K-3
Characteristics of Dyslexia								
Sound Symbol Recognition	Not assessed	K,1,2	K,1	K-3	K,1,2	K,1,2	K,1	K
Alphabet Knowledge	Not assessed	1,2	K-3	K-3	K,1,2	K-3	K	K-3
Decoding Skills	Not assessed	K-3	K,1	K-3	K,1,2	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,1,2	Not assessed	K,1	K-3
Accuracy of Word Reading	Not assessed	1,2,3	K-3	K-3	K,1,2	K-3	1,2,3	K-3
Sight Word Reading Efficiency Skills	Not assessed	1,2,3	K-3	K-3	K,1,2	K,1,2	K,1	K,1

Table 3(b). Domains assessed.

	FastBridge aReading	i-Ready Assessment for Reading	ISIP Reading with RAN and ORF	iSTEEP	MAP Reading Fluency	mCLASS	MindPlay Universal Screener	Star Assessments
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	1,2,3	K-3	K-3	K-3	K-3	K-3	K-3
Oral Language	K,1	1,2,3	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	1,2,3	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 4 summarizes the strength of each psychometric index evaluated as well as the source of information for these metrics. NCII was the first source of information used for our evaluation. For screeners not evaluated by NCII, the primary source of information was publisher report submitted to GaDOE. Additional information from publishers' websites, journal articles, and technical manuals was also used. Specifically, Table 4 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 4 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 3. The metrics in Table 4 were analyzed against specific cut points to represent varying levels of reliability, criterion validity, sensitivity, and specificity. A key is provided in the table that indicates these cut points, from low to acceptable, weak to strong, and weak to acceptable.

Table 4. 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 Screener	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 Assessments	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 a number of 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*). Of all of these, *MindPlay* 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*) did not publish information on sensitivity or specificity.

All screeners reporting indices of reliability performed within acceptable levels and all screeners reporting on 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 measures reporting sensitivity, six received a determination of weak (*Acadience*, *Classworks*, *Easy CBM*, *iSTEEP*, *MAP Reading Fluency,* and *mCLASS*). Only two screeners received a determination of weak specificity (*ISIP Reading* and *MAP Reading Fluency*).

It is important to mention inconsistencies in reporting for two screeners. That is, 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 the use of tools where publishers may have used reduced rigor in evaluating and reporting.

Given the information available to us and in examination of each of the aforementioned 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 such that these tools could be considered relative to one another. Relative screener rankings are provided in Table 5. It is important to note that these rankings provide a comparison **only** among the screeners approved by the SBOE. For example, a 'weak' designation indicates a tool's relative standing to the other screeners on the approved list and 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 5. Relative screener rankings.

Strong	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	Acadience Reading K-6, FastBridge aReading, iSTEEP, MAP Reading Fluency, mCLASS
Weak	EasyCBM, Battelle Early Academic Survey, MindPlay

Discussion

The Deal Center conducted an independent review of universal literacy screeners approved by the SBOE as meeting the screener requirements listed 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 the most reliable and valid measures to identify students at risk for reading difficulties with the

greatest precision. 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*. In consideration of these global groupings, a few issues should be taken into account. These are discussed in detail below.

Sensitivity and specificity should be interpreted together when determining the overall usefulness of a diagnostic test (Shreffler & Huecker, 2023). As such, we identified the strongest screeners are those that demonstrate both acceptable sensitivity and specificity, but prioritized sensitivity over specificity. Sensitivity is prioritized because it ensures accurate identification of children who need access to early reading interventions and conserves valuable school resources by accurately identifying children who do not need these 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 of this is that the average can mask variability in sensitivity at different grade levels. While a screener might have strong sensitivity at certain grades, weak sensitivity of a tool at any grade level should be considered as a key factor in decision-making when selecting screeners.

One of the primary challenges of evaluating these screeners is inconsistency in information available. This inconsistency is found in both lack of information and discrepancies in reporting. Two screeners (*iSTEEP* and *MindPlay*) for example, did not provide grade-specific metrics of reliability and validity. Relatedly, two screeners 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 weighting to tools demonstrating the greatest breadth of strong performance across grades.

The limited nature of this review is important to note. This review 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 Deal Center 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 SBOE 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 5) from the SBOE's approved list that present with superior psychometric features as compared to the remaining nine screeners.
- This review was completed following a review by the SBOE 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

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), 10.1037/a0025323. https://doi.org/10.1037/a0025323

McDaniel, C. E., & Ziniel, S. I. (2023). A psychometrics primer: The basics all hospitalists should know. *Hospital Pediatrics*, *13*(3), e63–e68. doi: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

NCII. (n.d.). C*lassification Accuracy*. https://intensiveintervention.org/sites/default/files/Classification_Accuracy_508.pdf

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

Sattler, J. M. (2020). A primer on statistics and psychometrics. In *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. *StatPearls* [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557491/

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

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), 240. https://doi.org/10.3390/brainsci12020240