BUROS
CENTER FOR TESTING

Psychometric Review of Reading Screeners for the Nebraska Reading Improvement ACT

January 2026 [Amended January 26, 2026]

Prepared by:
Jessica L. Jonson
Tim Moses
Ryan Kettler
Pam Bazis
Emily Fisher

Contact: Jessica L. Jonson, PhD
jjonson@buros.org

Table of Contents

Psychometric Review of Reading Screeners

for the Nebraska Reading Improvement Act (NRIA)

Executive Summary

This report presents conclusions from a review of the appropriateness, technical adequacy, and usability of 11 reading screeners from seven vendors under consideration for statewide approval to meet the requirements of the Nebraska Reading Improvement Act (NRIA). NDE requested that the Buros Center for Testing engage a team of psychometric, screening, and reading experts to review the documentation requested from vendors. In October 2025, that documentation was received and reviewed independently and collectively by the team over the next four months to determine whether evidence for each screener met expectations. This review resulted in ratings for each screener of *met*, *met with weaknesses*, *partially met*, or *more evidence needed*. The full report details NDE's request, the review team, the review process, the review criteria, and a summary of conclusions for each screening measure. A summary of ratings for each screener reviewed is as follows.

The **i-Ready Early Literacy Screener** *met expectations* for appropriateness for NRIA purposes and *partially met* the technical adequacy criteria. Although much of the technical evidence submitted met the criteria, questions remained about norms and the need for additional evidence on classification consistency and fairness.

The **mClass DIBELS 8th Edition** *met expectations with weaknesses* in its appropriateness for NRIA purposes, and *partially met* the technical adequacy criteria. Questions were raised about the use of multiple forms and the weighting of subtests to compute the screening composite score. The review also noted missing or incomplete information on classification accuracy, classification consistency, and fairness.

FastBridge earlyReading, one of three screening measures in the FastBridge Suite, *met* expectations. The measure is intended for use with kindergarten and first-grade students. The evidence raised questions about multiple forms. The screener *partially met* the technical adequacy criteria due to insufficient evidence on classification consistency and fairness, as well as the need for updated norms, classification accuracy, reliability, and validity studies.

FastBridge aReading is recommended for screening 2nd and 3rd-grade students as part of the FastBridge Suite. aReading *partially met* expectations for appropriateness due to concerns about alignment of key reading skills for screening at 2nd grade and about the computer-adaptive algorithm's administration of content. The screener *partially met* the technical adequacy criteria due to insufficient evidence on classification consistency, internal validity, and fairness, as well as the need for updated norms and classification accuracy, reliability, and validity studies.

FastBridge CBMReading is an oral reading fluency measure in the FastBridge Suite that is also used in screening decisions along with earlyReading and aReading for 1st-, 2nd-, and 3rd-grade students. It met expectations for appropriateness, though it had some weaknesses due to questions about the administration of multiple passages. It *partially met* technical adequacy expectations due to a lack of evidence for classification consistency and internal validity as well as the need to update norms, classification accuracy, reliability, validity, and possibly fairness studies.

Amira Reading Mastery (ARM) Universal Screener *partially met* expectations. The coverage and alignment of proposed content were found appropriate. However, if the algorithm governs content exposure and the lack of clarity about how pattern scoring works are concerns. Technical adequacy was also *partially met*. The vendor provided updated psychometric studies but lacked sufficient detail to fully evaluate those studies, and there were concerns about variables and results in others.

Acadience Reading was found to have some weaknesses in its appropriateness expectations, particularly regarding content coverage and alignment. More evidence was needed to demonstrate technical adequacy due to outdated norms, cut score analyses, and classification-accuracy studies, prompting the discontinuation of the review of this measure.

MAP Reading Fluency is rated as *more evidence needed* for both appropriateness and technical adequacy. There was insufficient documentation on how screening outcomes are determined and how the multi-stage adaptive test ensures consistent content delivery. *More evidence* was also *needed* to demonstrate technical adequacy, including more information on the impact of the model-based approach used to calculate norms, external criteria for the classification-accuracy study, internal validity and fairness evidence, and documentation of reliability studies.

MAP Growth was rated as *more evidence needed* for both appropriateness and technical adequacy. We are unclear whether MAP Growth was designed for screening decisions and whether it provides evidence on how MAP Reading Fluency, MAP Growth K-2, and/or MAP Growth 2-5 should be used separately or together to make those decisions. The extent to which key skills for screening at each grade level are assessed by MAP Growth K-2 and MAP Growth 2-5. For technical adequacy, concerns include the appropriateness of the model-based approach for calculating screening norms, classification accuracy, and consistency evidence, as well as the need for additional reliability and validity studies.

STAR Early Literacy and **STAR Reading** were found to require more evidence for appropriateness and technical adequacy. Although the content of the screeners appeared to align with key foundational reading skills, concerns arose about whether enough items assessed each skill and whether the adaptive algorithm accounted for content coverage. Questions about the appropriateness of the unified scale for both STAR measures were also raised. The lack of data collection on norms and outdated classification accuracy studies led to the decision to discontinue the review of the measures.

Usability of each screener was also addressed, but not rated. A summary of usability considerations is addressed for each screener in the report.

The report concludes with some recommendations about what NDE should expect from any vendor of screening measures approved for statewide use.

The members of the review team are available to discuss our review findings with NDE and Nebraska educators as needed.

Psychometric Review of Reading Screeners

for the Nebraska Reading Improvement Act (NRIA)

Overview of NDE Request

The Nebraska Reading Improvement Act (NRIA) (Section 79-2601-79-2607) requires that each school district administer an approved reading assessment three times during the school year to all students in kindergarten through grade three.[1] The legislation requires the State Department of Education (NDE) to make public, before March 1 of the subsequent school year, a list of approved reading assessments and the threshold performance level for each. Any student performing below the threshold is identified as having a reading deficiency for NRIA purposes. The NRIA states that an approved reading assessment will:

> (a) measure progress toward proficiency in the reading skills assessed pursuant to subsection (5) of section 79-760.03 on the statewide assessment of reading for grade three;
>
> (b) be valid and reliable;
>
> (c) be aligned with academic content standards for reading adopted by either the State Board of Education pursuant to section 79-760.01 or the school district administering such assessment pursuant to section 79-760.02;
>
> (d) allow teachers access to results in a reasonable time period as established by the department, not to exceed fifteen contract days; and
>
> (e) be commercially available and compliant with requirements established by the department.

For the 2025-2026 school year, NDE approved eight reading assessments but would like to select a smaller set of screeners for school districts to use for NRIA purposes. Therefore, to assist NDE in making this decision, the Buros Center reviewed the submitted responses and documentation for the following screening measures to evaluate the appropriateness, technical adequacy, and usability of each screener. NDE requested review of the following 11 screeners from seven different vendors:

[1] All students except for any student receiving specialized instruction for limited English proficiency who has been receiving such instruction for less than two years, any student receiving special education services for whom such assessment would conflict with the individualized education plan, and any student receiving services under a plan pursuant to the requirements of section 504 of the federal Rehabilitation Act of 1973, 29 U.S.C. 794, or Title II of the federal Americans with Disabilities Act of 1990, 42 U.S.C. 12131 to 12165, as such acts and sections existed on January 1, 2021, for whom such assessment would conflict with such section 504 or Title II plan.

- Acadience Reading
- STAR Suite (STAR Early Literacy & STAR Reading)
- FastBridge Suite (earlyReading, aReading, CBMreading)
- Amplify mClass DIBELS 8th Edition
- MAP Reading Fluency and Growth
- Amira
- I-Ready Early Literacy Screener

Review Process & Team

To complete this review, the lead investigator for the Buros Center, Dr. Jessica Jonson, assembled a team with expertise in psychometrics, universal screening, and reading. The potential team members were shared with NDE for approval. The review team included the following individuals.

- Dr. Jessica L. Jonson, Project manager and Psychometric reviewer
  Associate Director of Assessment Literacy & Research Professor, Buros Center for Testing, University of Nebraska-Lincoln

- Dr. Tim Moses, Psychometric reviewer
  Research Professor, Buros Center for Testing, University of Nebraska-Lincoln

- Dr. Pam Bazis, Reading assessment reviewer
  Assistant Professor, Special Education and Communication Disorders
  Co-Director, Kit and Dick Schmoker Reading Center, University of Nebraska-Lincoln

- Dr. Emily Fisher, Reading assessment reviewer
  Assistant Professor of Practice, Teaching, Learning, & Teacher Education
  Co-Director, Kit and Dick Schmoker Reading Center, University of Nebraska-Lincoln

- Dr. Ryan J. Kettler, Universal screening & psychometric reviewer
  Senior Research Scientist, F, K,&K Consulting
  Professor and Associate Dean, Rutgers University-New Brunswick

The team undertook four tasks to complete this work. The first was to draft a Request for Information (RFI) for NDE to send to vendors. The review team met with NDE to discuss the draft RFI before finalization. The RFI asked vendors to respond to a series of questions about their screening measures and to provide evidence to support the requested criteria. Given the brief review period, submitters were asked to cite specific documents and page numbers for the requested documentation. Vendors were given two weeks to respond, and all materials were received by October 1, 2025.

For the second task, the team identified criteria that would guide the review. In assembling these criteria, the team reviewed those used by other states that had conducted reviews of reading screeners and added considerations deemed essential to the Nebraska review process. For the third task, the review team members were assigned to review documentation aligned with their areas of expertise. Reviews were conducted individually, with all documentation reviewed by two to three team members, and each individual's conclusions were shared collectively to reach consensus. The final task was to draft a report summarizing the team's findings and provide NDE with ratings indicating whether screeners met, partially met, or required additional evidence to meet the identified criteria.

Report Organization

In this report, we provide an overview of what we reviewed and summarize our findings for each screening measure. For each screening measure, we determined whether the submitted documentation *met expectations, met expectations with weaknesses*, *partially met expectations*, or *required more evidence*. In cases requiring additional evidence, that evidence may exist but was not submitted or was incorrectly referenced in the submitted materials. Within reason, we attempted to locate the documentation that the vendor stated was available, but it could not be found at the provided reference. Therefore, the rating of "more evidence needed" does not imply that a screening measure is ineffective; it simply indicates that we were not provided with sufficient information to conduct a complete review.

Review Criteria

Criteria for the review process are organized into three types of considerations: 1) appropriateness of the screener for purposes of NRIA, 2) technical adequacy of the screener, and 3) usability of the screener. Although the technical adequacy of the screener was the primary focus of the review, the team also considered the screener's appropriateness and usability. Note that, in reviewing the submitted documentation, we sought information or evidence to support all claims that the vendor made about its screening measure.

Appropriateness of the screener

Our review of the appropriateness of a reading screener included confirming that the measure is intended to screen K-3 students for reading difficulties three times a year. It also involved ensuring appropriate coverage and alignment with relevant reading skills and administrative and scoring models that produced scores that accurately reflected the stated skills.

Intended interpretation and use. Screening for reading difficulties was the intended interpretation and use for most screeners, although some measures also served other purposes, such as achievement and growth reporting. However, our review did not evaluate evidence supporting these different uses, as our focus was on screening interpretation and use. Several of these vendors combine a curriculum-based measure (CBM) or performance measures with a proficiency measure for screening. In our review, we identified these different approaches and any potential issues, as detailed in the summaries for each screening measure.

Content coverage and alignment. Ideally, the selected screening measures would align closely with the key reading skills identified in the *2021 NE ELA CCR Standards*. However, because screening aims to identify at-risk students efficiently using a brief skills measure, it is unlikely that all reading skills will be comprehensively assessed, as might be done in an end-of-year summative assessment. Vendors were asked to specify the reading skills assessed and document any alignment studies between the screener's content and the *2021 NE ELA CCR Standards*. Most vendors discussed an alignment process with either the *NE ELA CCR Standards* or the Common Core standards. However, the use of independent panels was not typical, and/or documentation of alignment results was insufficient for us to evaluate content adequacy.

Therefore, the two review team reading experts (Drs. Bazis and Fisher) reviewed content to determine whether it was sufficiently aligned with the *2021 NE ELA CCR Standards*. This task was difficult, given that the 2021 NE ELA CCR Standards typically measure more than one skill and often ask for higher-level or a range of proficiencies within a single standard. This occurs because these standards are meant to guide learning objectives over the course of an entire year and so require more than a screener can provide.

The team decided to supplement the vendor's documentation of the content by reviewing whether a screener assesses key skill areas and selected subskills that align with the *2021 NE ELA CCR Standards*. This alignment approach would ensure consistency with the policy without addressing the less relevant issue of whether the screening content fully met the standards, as in a proficiency test. The findings from Dr. Bazis and Fisher's review are in Appendices A and B. In addition to the skills assessed, there are differences in how students are prompted and permitted to respond. We note in the summaries below how test content is presented to students and whether students are prompted to respond orally. This information is considered important for assessing skills, such as blending sounds, to allow for a more authentic demonstration of proficiency.

Administration and scoring models. For some screeners, the documentation in Appendices A and B appears to support adequate coverage of fundamental reading skills. However, an evaluation of the administration and scoring models raised questions about whether the screener administered sufficient items to represent each skill adequately and about the extent to which each skill was represented in the scores used for the screening decision. Test blueprints or specifications could have provided evidence of coverage, but they were often not provided. In addition, documentation of the CAT test specifications or flows needed to demonstrate that students would be administered items across all appropriate skill content areas was either missing or unclear for some screening measures.

How scores are calculated or generated also affects screening decisions, such as weighing some subtest scores more than others or using scoring algorithms that consider factors beyond item parameters (e.g., difficulty) and whether the student answered items correctly.

A detailed discussion of the different types of administration and scoring models used for the screeners is provided in Appendix C. Questions and concerns raised about those methods are also discussed in the summaries for each screener.

Technical Adequacy of Screener

The technical adequacy of each screener was reviewed in two phases. In the first phase, we focused on the relevance and representativeness of the norms, the appropriateness of the cut scores, and the accuracy of screener-based classification decisions. If screeners were found to at least partially meet these psychometric considerations, we then examined the validity, reliability, fairness, and usability of the measures. We discontinued our review for two measures after this first phase: Acadience and STAR Suite. The reasons for those decisions are detailed in the summary for those measures below.

Relevance and representativeness of norms. We were concerned that the norms for several measures have not been updated since the pandemic, even though vendors that updated their norms after the pandemic observed notable declines in student performance. It is best practice to update norms at least every five years and whenever a measure undergoes substantive revision, such as new content, scoring, or scaling approaches. In most cases, norms were described as nationally representative based on national demographic samples. However, the reporting of demographic information varied, particularly in the extent to which norms included students from diverse geographic locations, racial/ethnic backgrounds, ability levels, and socioeconomic backgrounds.

However, if a screener has updated its norms with a nationally representative sample within the past seven years and has not undergone substantive changes, those norms are appropriate for use if other psychometric evidence supports the reading-screening decisions. In these cases, we recommend that NDE inquire about when the norms will be updated and whether the vendor can provide the calculation of local (statewide or district) norms. For example, the FastBridge Suite indicates that it offers local norms when a large percentage of students are administered the screener. Other vendors may also provide local norms at the district or state level. For most screeners, norms were directly calculated from raw or weighted data; however, for MAP, a model-based approach with a more limited fit of the data was used. We have concerns about using this approach for normative considerations in screening decisions, as discussed in greater detail in the summary of MAP measures below.

Appropriateness of cut scores and classification accuracy. How cut scores were set and validated for accuracy is highly consequential for reading screening decisions. When evaluating the process for setting cut scores, we focused primarily on Tier 1 cut scores (students identified as at some risk of reading difficulties). Several measures also offered cut scores for Tier 2 (students at clear risk for reading difficulties). We examined how and whether cut scores were set for each grade and administration window. Although recent cut-score studies were preferred, what concerned us more were inconsistencies in the timing of cut-score setting, norm updates, and substantive test changes. Wherever these inconsistencies were a concern, they are noted in the summary for each measure. We also reviewed whether classification accuracy indices (ROC AUC, sensitivity, and specificity) were used to adjust the cut scores. These studies use a criterion variable to validate the cut scores used to make screening decisions. Ideally, a criterion that measures reading skills and is independent (external) to the screening measure is best practice. All measures described how cut scores were set and used the prediction of a criterion variable to confirm or adjust those scores. However, not all criterion variables were deemed appropriate. This concern is noted, along with the respective measures, in the summaries below.

When validating cut scores, conducting a separate classification-accuracy study to cross-validate the cut-score results is a best practice because using classification-accuracy indices derived from the same analyses used to set the cut scores can potentially inflate them. In most cases, a second classification-accuracy study was not conducted. We also found that studies reporting classification accuracy indices did not, in some cases, use their own recommended cut scores, possibly to maximize classification accuracy. For reported indices, we had minimum expectations of .80 or higher for sensitivity and Negative Predictive Value (NPV), and .60 or higher for specificity and Positive Predictive Value (PPV). Sensitivity and NPV thresholds were set higher to ensure that struggling readers were not missed (i.e., to minimize false negatives), thereby erring on the side of overidentifying rather than underidentifying. We would note that several vendors indicated they adjusted cut scores to avoid overidentification (increasing specificity rather than sensitivity) out of concern for time and resources. The reported indices for several screeners were poor for some grades and administration windows, or were not provided at all. Out-of-date studies were also a concern.

Given that classification-accuracy studies across screeners differed in criteria, samples, and reported results, it is difficult to say definitively that one screener is more likely to provide more accurate classification than another. However, some vendors conducted more rigorous studies with results that met our minimums. Therefore, we have documented the extent to which studies adhered to best practices and provided supportive evidence for each screener. Conducting a rigorous classification-accuracy study can be challenging, but that does not diminish its importance, particularly when seeking to improve the comparability of reading-screening decisions across the state. We recommend that NDE ask vendors to assist in gathering and analyzing data to conduct classification-accuracy studies for any measures approved for NRIA.

We also requested evidence of classification consistency from vendors. While classification accuracy supports the validity of screening classifications, classification consistency (decision consistency) supports their reliability. In most applications, evidence of classification consistency was either absent or insufficient. We expected this, but it is a valuable indicator of screening decision stability and is particularly important when the screener uses a CAT or has multiple forms. The extent to which classification consistency evidence was provided is addressed in the summaries below. Again, this is a data point NDE may want to ask vendors of selected screeners to either provide or collect.

Evidence of Reliability, Validity, and Fairness. Vendors were asked to submit evidence supporting the reliability, validity, and fairness of scores used to make screening decisions.

To ensure reliability, we evaluated whether all relevant studies were conducted to demonstrate the consistency and precision of reading screening scores. We also assessed the representativeness of the study sample, as reliability results are sample-dependent. Study results supporting the stability of scores across time, forms, and raters were common. Precision data were typically reported only for the full-scale standard error of measurement (SEM), rather than for conditional SEMs (CSEMs) at key cut-score points. Even when conditional SEMs (CSEMs) were reported, they were often presented across scales rather than at key cut scores. Alternate-form reliability is important for non-adaptive screeners with multiple forms (DIBELS 8th Edition, FastBridge earlyReading, FastBridge CBMreading). Although informative, those studies conducted do not provide evidence of whether some forms were more difficult than others. Direct evidence would involve equating forms, classification consistency evidence, or, at a minimum, reporting CSEMs for each form at key cut scores. For any statewide reading screener adopted, the vendor should provide the state with reliability evidence based on Nebraska students, both in the aggregate and specifically at the cut scores used to make screening decisions, and, when applicable, for each form used in the screening decision.

When evaluating validity, we focused on internal and external validity. To assess internal validity, we examined whether vendors demonstrated that the test subskills were related rather than redundant. We also requested evidence of how reading screening scores relate to other scores or variables of similar constructs. Samples, variables, methods/ procedures, and reported results were evaluated to determine whether the evidence was sufficient to support the validity of inferences from reading screening scores. Most vendors provided adequate proof of validity for at least one source. As with reliability, NDE could work with vendors to provide validity evidence using data from Nebraska students to ensure that the assessment of individual skills by reading screeners is being measured as expected and that comparability between decisions made based on reading screening scores and external variables of students' reading skills (scores from other measures or teachers' assessment of student reading skills) exists.

For fairness, we sought studies evaluating potential subgroup differences in reading screening scores to determine whether the test is less effective at identifying reading difficulties for specific subgroups. We preferred psychometric studies over panel reviews, although a rigorous fairness review process is good practice. Vendor-submitted fairness evidence varied, and in a few cases, the results raised possible concerns. The most useful evidence we found was a comparison of reliability and validity results across subgroups to assess comparability. These analyses were provided by some vendors, but the reporting of this evidence was inconsistent. Some vendors provided subgroup comparisons only for reliability, not validity evidence, and others offered spotty evidence across subgroups. If the fairness of reading screening decisions for specific student subgroups is a concern for NDE in districts with diverse student populations, the vendor could provide the state with a comparison of reliability and validity evidence for students in selected subgroups.

Usability of the Reading Screener

The screener's usability addressed practical issues related to the time and cost of administration and training, how data are accessed, reported, or used for instruction, and evidence of the measure's accessibility and the availability of accommodations. We also reviewed the availability of dyslexia screeners for each measure.

Administration and training time and cost considerations. Differences in how measures are administered and scored affect the time and resources required to assess each student. Measures also varied on whether tests had to be administered 1-on-1 with each student or could be administered to a group. One-on-one administration is more time-intensive and expensive but also allows teachers to engage more with students when assessing their skills. Group administration can be more efficient by administering to multiple students at the same time, but also may take more time to administer (up to 40-60 minutes) and require a digital device for each student as well as equipment for playing and recording audio with headphones and microphones that minimize interference of background noise, which in some cases means testing students with sufficient space in between. The differences in test formats and their administration have pros and cons. These differences are essentially matters of preference if the measures accurately reflect students' reading skills. Table 1 outlines the format, time requirements, and devices required to administer each measure.

The cost to administer each measure, whether training is required, and the cost of that training also vary across measures. Table 2 lists the licensing costs, training requirements, and training costs for each measure. What is included in licensing costs across measures appears to be somewhat comparable. To the extent possible, we have shared a summary of the vendor's documentation in Tables 1 and 2, but NDE may want to confirm the licensing, administration, and training costs and requirements. At a minimum, a license to administer the test to students statewide and access to a digital data and reporting platform to track screening information are recommended. We also would recommend training educators to administer 1-on-1 tests to improve consistency in administration and scoring across students, even if the vendor indicates that training is not required.

Access, reporting, and using data. Vendors were asked to provide information on how data can be accessed by school and state personnel, and whether data can be embedded in management systems at different levels. If the vendor provided this information, we reviewed whether access to student data was controlled and in compliance with FERPA privacy requirements. Most vendors provided this information. Whether these considerations were addressed is noted in the summaries for each measure below.

Vendors were also asked whether score results could be aggregated at different levels. All vendors indicated their reports could be aggregated at different levels. We recommend that, when selecting measures for statewide use, NDE consider whether the reports for those measures allow teachers to see errors for individual students and whether reports at the grade, teacher/classroom, individual, and subgroup levels inform data-driven decisions about student instructional needs.

We also consider, to some extent, the instructional resources provided for each measure. However, we recommend caution regarding the extent to which schools and teachers should rely on instructional resources for two reasons. One, the nature of screening is not to provide a comprehensive diagnosis of an individual’s reading strengths and weaknesses. It serves more as an indicator that a student may require further assessment to determine which instructional supports will help the student reach reading proficiency. Screeners, by their nature, are often shorter than proficiency measures and rarely have subscales that inform instruction that are both internally consistent and independent of one another to the degree necessary to represent separate constructs. Some screening measures may provide more data to inform these decisions than others, but assessing the extent to which instructional recommendations led to intended outcomes was beyond the scope of our review. The second reason is that some recommended instructional activities may not align with the school's reading curriculum and, therefore, may not support students in achieving the reading proficiency defined by that curriculum.

Accessibility and Accommodations. Vendors provided information on steps taken to increase the accessibility of their tests for all students and recommended or endorsed various accommodations. We reviewed vendor documentation to determine whether both accessibility and accommodation were considered in the measure's design and in the administration and use recommendations. A summary of the vendor's attention to accessibility and accommodations is provided in the individual summaries below.

Dyslexia screening. Although dyslexia screening is not required under NRIA, it remains an essential indicator for schools to consider when screening students for reading difficulties. We asked vendors to indicate whether their measures screened for dyslexia and to provide evidence supporting that screening. We did not review the psychometric evidence for dyslexia screeners; however, the summaries below indicate whether a dyslexia screener is offered, whether it is embedded in or offered as a supplement, and whether the vendor provided psychometric evidence for the dyslexia screening decisions. If the availability of a dyslexia screener becomes a criterion for selecting measures for statewide approval, a more thorough review of the screener's psychometric evidence should be conducted.

Screening Measure Summaries

For each screener, we address whether the submitted documentation met our review criteria for appropriateness and technical adequacy. Summaries are ordered from highest to lowest ratings, although different measures within a suite are listed sequentially. The order of the summaries should not be interpreted as a ranking, as qualitative considerations beyond our ratings should also inform decisions on which screening measures to approve. For ease of review, each screener summary starts on a new page. Table 3 lists our rating for each screener.

i-Ready Early Literacy Screener

a. Appropriateness of i-Ready Early Literacy Screener - ***Met***

Intended interpretation and use. The stated goal of the Early Literacy and Dyslexia Screener is to identify students with reading deficiencies to provide early supplemental intervention. The Early Literacy and Dyslexia Risk Screener uses a two-step process: first, administering the i-Ready Diagnostic Reading test, followed by a grade- and time-specific benchmark fluency Literacy Task.

Content coverage and alignment. The vendor shared results from a recent alignment between the i-Ready diagnostic and the *2021 NE ELA CCR Standards,* but did not document the panelists or the process. Results indicated good alignment, but did not specify whether alignment between test content and each standard was full, parsed, or related. Documentation of an independent alignment study between i-Ready Diagnostic and CCR Standards in 2017 was also provided. Literacy tasks that contribute to the screener for each grade are letter naming fluency for kindergarten, word recognition for 1st grade in fall, and passage reading fluency for 1st grade in winter and spring, 2nd grade in all administration windows, and 3rd grade in all administration windows. According to Appendix A and B, the content covered by the two assessments encompasses all skills necessary for reading screening across all four grades (K-3). This is in addition to measures of rapid automized naming and spelling. Although content on the i-Ready diagnostic test is not delivered orally, and students select a response, the literacy tasks require an oral response.

Administration and scoring model. An adaptive algorithm selects items for administration of the i-Ready Diagnostic Reading tests. The vendor provided documentation showing that the algorithm is partly based on grade-specific test flows that control the content administered to different test takers (Appendix C), ensuring that the skills outlined in the table are assessed for each student administration. Both the i-Reading Diagnostic Reading score and the Literacy Task score are equally weighted to calculate the Early Literacy Screener composite score used to make the screening decision.

b. Technical Adequacy of i-Ready Early Literacy Screener – ***Partially met***

Norms. i-Ready provided extensive documentation of norms for its Diagnostic test, updated in 2022-23. The study used a demographically representative sample across all grades and all three administration windows. However, norms for the Early Literacy Screener scores used in screening decisions were not found. We assume norms for the Early Literacy Screener scores would be needed, given placement levels are based on the 25th and 50th percentiles of the Diagnostic and "the 25th and 50th percentiles for the operational data available in the i-Ready system for the Diagnostic and Literacy Tasks." (p. 44 Early Literacy and Dyslexia Screener Technical Manual). Except for the norms reported in the Technical Manual for the i-Ready Diagnostic, percentiles are not provided for the Early Literacy Screener composite score, and they are not explicitly referenced in the RFI form. Based on this documentation, we are unclear whether the cut-score decisions for the Early Literacy Screener are based on percentile ranks. If so, a normative study for the Early Literacy Screener would be needed.

Cut scores. Cut scores for the early literacy screener were recently set (2023) using an external criterion (DIBELS 8). Although a separate study of classification accuracy was not conducted, the recommended cut scores were used, but sample demographics were unavailable. The sensitivity and specificity indices for the approaching/on performance level cut (assumed to be the 50th PR) were all above 0.80 across all grade levels and administration windows. Positive predictive value (PPV) and negative predictive value (NPV) were not reported; NPV is particularly important given the desire to avoid false-negative cases.

A classification consistency study for the i-Ready diagnostic was conducted in 2021-22, and the vendor states that a new study with the Early Literacy Screener will be conducted in 2025-26. Results support the consistency of the i-Ready Diagnostic, although it was not clear which administration window the results were calculated for.

Reliability. Multiple reliability estimates were provided for the i-Ready Diagnostic, Literacy Task, and Early Literacy Screener composite. Current IRT-based and test-retest reliability studies for i-Ready Diagnostic reading were conducted with large samples of students. Although the sample demographics for these studies were not shared, subsequent subgroup reliability analyses of IRT-based indicators showed representation across gender, race/ethnicity, economic disadvantage, English Learners (EL), and special education (SPED). Concurrent and delayed alternate-form reliability was provided for the Literacy

Tasks using samples from 2020-21, 2021-22, and 2024-25. Classification consistency for the i-Ready Diagnostic (as mentioned above) was also reported. All reliability results were strong, and SEM and CSEM for the i-Ready Diagnostic were also reported, indicating appropriate precision for the i-Ready Diagnostic score. At a minimum, a classification consistency study for the Early Literacy Screener Composite and, perhaps, other forms of reliability would be preferred. The vendor indicates that a classification consistency study for the Early Literacy Screener composite is planned for 2025-26. NDE should request data from this study (or any other reliability study for the Early Literacy Screener composite) when available.

Validity. The vendor provided evidence of internal validity and validity with external variables. For internal validity, the structure and correlational evidence for the screening tests were provided. Correlations between Literacy Tasks and the i-Ready Diagnostic score were also reported, with moderate-to-strong effects. A reported correlation among Literacy Tasks given in the same grade would have been helpful. To demonstrate validity with external variables, the Early Literacy Screener composite was correlated with DIBELS 8 and i-Ready Diagnostic Mathematics, providing evidence of divergent validity. Correlations with DIBELS 8 were high across all grades and administration windows, and higher than those between the Early Literacy Screener and the i-Ready Diagnostic Mathematics. Additional validity evidence with external variables was provided for i-Ready Diagnostic and Literacy Tasks. The predictive validity of i-Ready Diagnostic reading test scores with the literacy tasks in spring, fall, and winter provided support for the utility of intervention planning for those tasks.

Fairness. DIF analyses of i-Ready Diagnostic Reading items were conducted in 2015 on a large sample of students with representation for gender, ethnicity (Caucasian, African American, and Hispanic), Region (Midwest, South/West, Northeast), EL, SPED, and SES. The percentages of items showing small, moderate, and large DIF were reported for each group, and fewer than 2% of those items showed large DIF. A more recent analysis would have been preferred, but it also depends on how much the item pool has changed since then. Subgroup performance on IRT reliability and alternate-form reliability for Literacy Tasks was assessed in 2023-24 and 2024-25, with comparisons by race, gender, EL, SPED, and economic disadvantage. Reliability coefficients were similar across subgroups, except for a few subtests with small sample sizes. The vendor stated they had documentation comparing classification accuracy across subgroups, but we were unable to locate it in the submitted materials.

Even though much of the psychometric evidence for the i-Ready Early Literacy Screener met expectations, we rated it Partially Met due to a lack of clarity regarding the use of percentile ranks as cut scores and the possible need for a norming study for the screener scores.

c. Usability of i-Ready Early Literacy Screener

Administration time, cost, and training. The i-Ready Diagnostic assessment is group-administered and takes 25-35 minutes at the kindergarten and 1st-grade levels and 40-60 minutes at the 2nd- and 3rd-grade levels. The Literacy Task, which is also used for reading screening, is administered 1-on-1 and takes 1-5 minutes per student, depending on the task. A digital device (e.g., a computer or iPad), as well as headphones and speakers, is required. The estimated licensing fee for the 2026-27 school year is $8.25 per student. Training is required and costs $2,400 per three-hour session, with a maximum of 30 participants. Training is provided to school personnel and school leadership. Included in the licensing cost are online professional development courses that supplement the training, a support website, and collaborative learning extensions for facilitating meetings among colleagues. See Tables 1 and 2 for more information.

Reporting, access, and data. The vendor claims score results can be shared at the individual, classroom, school, and group levels. Example reports were not provided, but detailed descriptions of the available reports indicate that they include the necessary information. Family reports that explain the meaning of scores are available and are easy to interpret. Unique logins can be assigned to different roles for school personnel, and reports can be downloaded in CSV format. The data integration process is reported as FERPA-compliant and does not share student data with any unauthorized parties. Instructional tools are included with the license, and diagnostic reports for each individual identify the relevant instructional tools. Before selecting the i-Ready Early Literacy Screener for statewide use, NDE should request sample reports to ensure they meet teachers' and schools' needs.

Accessibility and accommodations. i-Ready Diagnostic complies with the 508-compliant Web Content Accessibility Guidelines (WCAG) 2.0 Level AA, with limited exceptions in the document. i-Ready indicates a commitment to address those exceptions and to fully meet higher compliance standards. i-Ready provided an extensive list of embedded and non-embedded student supports and accommodations that could be used with both i-Ready Diagnostic and Literacy Tasks.

Dyslexia screening. A dyslexia screener task is included and can be administered as needed. The results of the Early Literacy Screener can be used to identify which students should be further assessed for dyslexia risk factors. Classification accuracy data for the dyslexia screener were provided in the documentation.

mClass with DIBELS 8th Edition

a. Appropriateness of DIBELS 8th Edition - ***Met with weaknesses***

Intended interpretation and use. DIBELS 8th Edition was designed to serve as a universal reading and dyslexia screener, as well as a benchmark and progress-monitoring assessment. The mClass system provides a digital platform for entry, analysis, and reporting of student data. It is also used to administer the Maze subtest.

Content coverage and alignment. The vendor conducted an alignment study of the 2021 NE ELA CCR Standards with independent panelists. Panelists decided whether the alignment was clearly or somewhat consistent, though those decisions for each standard were not documented. A results table from the alignment study suggested coverage of most standards. However, this included the spelling and vocabulary subtests, which are not used in the screening composite score. Reading skills considered essential to assess in a reading screener at each grade level appear to be addressed at appropriate grade levels (Appendices A and B). For the most part, content is presented orally, and students may respond orally.

Administration and scoring models. Screening is based on a composite score that weighs the raw scores from multiple subtests, including letter-naming fluency, phonemic-segmentation fluency, nonsense-word fluency, word-reading fluency, oral-reading fluency, and Maze. Maze is for 2nd grade and higher grades. Oral Reading fluency is first grade and higher. All other subtests are for kindergarten and higher. However, several factors could influence the content delivered and the resulting score.

- DIBELS has continuation and discontinuation rules that are important to review and follow in administration. For example, if a student in an older grade performs poorly on an initial assessment, teachers should be encouraged to follow the recommended sequence for assessing lower-level skills to determine the student's proficiency level. Gaps in phonemic awareness, syllable knowledge, or phonics related to word-attack skills may contribute to lower oral reading accuracy, rate, or comprehension scores.
- NDE should also review gating rules to determine whether those rules align with preferred practice across the state. Dr. Bazis and Fisher noted that the suggestion not to administer nonsense word fluency to a student who scores high on oral reading fluency – words read correctly might miss some students who are highly adept at memorizing words and/or passages but would struggle with the skills needed to read nonsense words. This is a particular concern, given that the oral reading fluency measure relies on a single passage. More information on rule recommendations is available in the DIBELS 8th Edition Composite Score Calculation Guide Supplement (July 2020).

- There are 20 different alternate forms for progress monitoring available for some subtests. However, the screening composite score is based on unequated raw scores, with each subtest weighted before the scores are summed to derive the composite. The vendor acknowledges the need to equate these raw scores across subtest forms; however, this has not been done. The issue with using raw scores without equating is that the difficulty of the administered form may influence the resulting scores and screening decisions. (See Appendix C).
- Indices from validity studies discussed below suggest the weighting for the kindergarten and 1st-grade composite might not be a good fit. We would encourage NDE to review the weights assigned to each subtest score to ensure they align with state guidelines for which skills should be used to screen for reading. How much each subtest score is weighted to derive the composite score can be found in DIBELS 8th Edition Composite Score Calculation Guide Supplement (July 2020).

b. Technical Adequacy of DIBELS 8th Edition – ***Partially met***

Norms. DIBELS 8 Norms were collected post-pandemic in 2021-22 from a large sample of students across all 50 states, although the demographics of that sample were not documented. Norms for the composite and all subtests were reported for all grades and all three administration windows. NDE should request DIBELS 8 demographic information to ensure that the norm sample is representative of Nebraska students.

Cut scores. Cut scores were set using external criterion variables. Although the cut-score study was conducted in 2018-19, the test has not undergone substantial changes since then. A separate classification-accuracy study was not conducted, but cut scores were updated in 2020-21 using 2018-19 data to improve concordance between subtest and composite scores. It was unclear how many students from the 2018-19 sample were included in the cut-score study or the demographics of that group. The sensitivity and specificity indices for the 40th PR were all above the minimums for all grades and administration windows except for K Fall. Given the recent normative update, NDE should request that DIBELS 8 provide updated cut scores based on the new norm sample, or, ideally, classification-accuracy results for Nebraska students.

The vendor cited a classification-accuracy study comparing subgroups on each subtest as evidence of classification consistency, but, as discussed in the review criteria section, classification accuracy and classification consistency provide different types of evidence for the screening decision. Classification consistency is essential for evaluating decision consistency because DIBELS 8 scores and screenings are based on raw, unequated scores from one of several alternate forms that may be administered to different students.

Reliability. Multiple reliability studies were conducted with large samples collected in 2017-18 and 2018-19. Sample demographics were not documented. Studies included test-retest, interrater, concurrent alternate forms, and delayed concurrent forms. However, the only type of reliability provided for the composite used in the screening decision was evidence of delayed concurrent reliability. All other reliability studies involved the subtests. The results for the composite score indicate consistent performance over time. SEMs were reported, but not the CSEM, at the individual cut score level. NDE should request conditional SEMs (CSEMs) or classification consistencies that address consistency across alternate forms for each decision-making cut point at each grade level and administration window, or request the collection of this information based on state-level data.

Validity. Evidence based on validity studies with external variables was strong. Multiple concurrent studies with large samples were documented. Correlations with DIBELS Next and Iowa Assessment composite scores were strong. A confirmatory factor analysis (CFA) was used to support internal validity. The strong correlation among some subtests suggests they may be redundant, though the implications for instructional decisions are greater than those for screening decisions. The reported poor fit indices (RMSEA) from the CFA raise questions about whether the weighted composite score model is appropriate for K and possibly 1st grade. NDE should request more information on the vendor's perspective on these high RMSEAs and/or plans to revise them to address this issue.

Fairness. A logistic regression of each subtest on DIBELS 8 to predict scoring at the 20th percentile or lower on DIBELS Next in kindergarten and on the Iowa Assessment total reading score for 1st-3rd grade was conducted to assess whether there was an interaction with demographic characteristics. They examined effects by gender, race/ethnicity, EL, SPED, and free/reduced lunch. No differences were found, except for the Nonsense Word Fluency-Words Read Correctly (NWF-WRC) subtest score in first grade, which showed greater predictive power for white students than for black students. They explained that this is not a concern because the screening decision is based on a composite of subtests rather than on a single subtest. Fairness studies conducted on the composite that provide evidence supporting the fairness of screening decisions for all subgroups would be an important follow-up to this study.

C. Usability of DIBELS 8th Edition

Administration time, cost, and training. All subtests in the mClass DIBELS 8 Edition are administered 1-on-1 and take 4-6 minutes in total, except for the 3-minute Maze task, which is administered in groups to 2nd- and 3rd-grade students. All subtests can be administered on paper or on a digital device, but scores must be entered into a digital device for reporting. The mCLASS DIBELS 8th Edition licensing fee is $9 per student, but is lowered to $7 per student with a multi-year license. The licensing fee includes access to the reporting platform and instructional tools. Training sessions are available onsite, remote, or as a self-paced online course. Full-day and half-day onsite and remote sessions are available for up to 30 participants. More information about administration time, costs, and training can be found in Tables 1 and 2.

Reporting, access, and data. The vendor claims score results can be shared at the individual, classroom, school, and group levels. Sample reports were provided and found to be useful. Targeted student instructional activities are provided one-on-one and with small groups. Family reports also offer activities. A school district can set role-based access for teachers, administrators, and students, and reports can be downloaded in CSV format. Vendor claims its products comply with Federal and State student privacy laws, including support for district compliance with the Family Educational Rights and Privacy Act (FERPA).

Accessibility and accommodations. Amplify claims to work with external experts on digital accessibility to ensure products are built according to WCAG guidelines. No confirmation of compliance was shared, although this is less important for a screener where students' use of a computer device is limited. DIBELS 8th Edition was designed for use without modification, although it does not mention universal design. Accommodations can be used when necessary to obtain accurate scores or as specified by students’ 504 plans. A list of six accommodations applicable to each subtest is provided.

Dyslexia screening. Separate subtests can be used alongside the DIBELS 8th Edition measures to screen for dyslexia, including RAN and/or spelling. A flag is displayed in the mCLASS system if a student demonstrates risk on the DIBELS 8th Edition composite score and the mCLASS RAN or spelling measure. A white paper was cited, but it did not provide psychometric evidence for dyslexia screening. The response indicated that the University of Oregon has established updated validity evidence for the screener, but no documents were cited.

FastBridge earlyReading (kindergarten & 1st Grade)

a. Appropriateness of FastBridge earlyReading - ***Met with weaknesses***

Intended Interpretation and Use. earlyReading is a curriculum-based screening recommendation for kindergarten and first-grade students, aligned with the "building blocks for foundational literacy" framework. It is one of three measures offered as a more comprehensive screening solution. aReading is offered for 2nd and 3rd grade. CBMreading is included in the earlyReading composite scores starting in 1st grade in the winter.

Content coverage and alignment. The skills blueprint (not items or subtests) was aligned with standards from different states, including the 2014 (not the 2021) NE ELA CCR Standards. Some alignment was found. Nine of the kindergarten standards and three of the first-grade standards were found to align. earlyReading content appears to screen key skills at the kindergarten and 1st-grade levels but may also be needed at grade 2 (Appendices A and B). However, because psychometric evidence for the use of earlyReading with 2nd graders is not provided, we do not recommend its use for screening in 2nd grade. Content is delivered orally, and students can respond orally.

Administration and scoring model. The composite score used for the screening decision is calculated by weighting and summing the raw scores from subtests assessing print concepts, phonemic awareness, phonics, and decoding. Alternate forms exist for each subtest, but there is no documentation that the scores from each form are equated, raising the same concern discussed above for the DIBELS 8 edition. CBMreading scores are included in the earlyReading composite starting in the winter of 1st grade.

b. Technical Adequacy of FastBridge earlyReading - ***Partially met***

Norms. earlyReading norms were collected in 2017-18. Given that other vendors reported declines in students' reading performance after the pandemic, a post-pandemic update is preferred. We found online that norm updates may have occurred more recently in 2023, but this documentation was not submitted. The vendor indicates that local norms are automatically generated when at least 70% of students are tested. NDE may want to inquire about this option and whether norms can be provided at the state, district, and school levels[2].

A large norm sample of K- and 1st-grade students was collected in 2017-18. Although the vendor states that the sample was insufficiently robust to represent the U.S. school population, it appeared demographically representative of gender, race/ethnicity, and

[2] It is possible other vendors may offer the option of local norms but it was not documented in the materials submitted.

free/reduced lunch, with a Midwest representation. The vendor claims the norms are more accurate across the full range of abilities, but no demographic data for SPED, EL, or gifted were documented. Norms were reported for K and 1 for the composite, all subtests, and all three administration windows.

Cut scores. According to NCII, cut scores for earlyReading were set in 2012-13 using spring GRADE (Group Reading and Diagnostic Evaluation) scores as an external criterion. This study would have occurred before CBMreading was included in the earlyReading composite score for 1st-grade students. The study samples were small from two school districts. Race/ethnicity and free/reduced lunch were reported and were somewhat proportional to NE. Gender, urban/rural, or geographic location were not documented. The area under the curve (AUC), sensitivity, specificity, and accuracy from these studies support the use of the 40th percentile as the cutoff.

Vendor referred to NCII for a more recent classification-accuracy study from 2021-22, using STAR Early Literacy as an external criterion for fall and winter cut scores. However, the GRADE study was still reported for Spring. FastBridge is now owned by the same vendor as STAR Early Literacy, but a different vendor developed it, so we considered STAR Early Literacy a valid external criterion. The sample of first-grade students was larger, but the demographics for this more recent study were not reported. Although the 40th PR was used for earlyReading in these studies, the 11th PR for Fall and 17th PR for Winter for STAR Early Literacy were used, which are closer to the Tier 1 (15th PR) than the Tier 2 (40th) cut scores for this criterion. Sensitivity and NPV results were reported for kindergarten and 1st grade. All sensitivities exceeded the minimum, except for the 1st-grade fall, which was moderate. NPV values were more moderate for kindergarten, and all exceeded the minimum for first grade. Specificity and PPV values were at or above the minimum thresholds. Given the issues raised about these studies, NDE should ask the vendor about its plans to update them and the possibility of collecting this evidence at the state level.

The vendor stated in their RFI response that evidence of classification consistency is not applicable, as the screener does not use multiple forms. Although students administered earlyReading receive a single form, the submitted Content Description and Use Guidelines state that “FastBridge earlyReading consists of 13 different evidence-based assessments for screening and monitoring student progress” (p. 36). In the submitted technical manual, alternate-form reliability is reported for the earlyReading measure; however, the vendor claims that multiple forms are not used. Therefore, we disagree with this conclusion, as the potential use of those forms across students could vary.

Reliability. Multiple forms of reliability were provided, including internal consistency in 2017-18, test-retest reliability in 2017-18, alternate forms in 2018-19, and inter-rater reliability. Sample sizes were sufficient, but demographics were not reported for all studies. Internal consistency reported gender and race/ethnicity, but the group was primarily white (70%). Data for alternate forms came from 15 states, which claimed to be representative of race/ethnicity, urban/rural/suburban, and free/reduced lunch, but no demographics were documented. Inter-rater-reported race/ethnicity, SPED, and free/reduced lunch demographics for districts where studies took place. Two of the districts were primarily white. No demographics were reported for the test-retest reliability study. Kindergarten and 1st-grade reliability results supported the composite's internal consistency and score consistency over time, but precision data for the composite were not reported. High correlations were reported for alternate forms for each subtest, but these results do not provide evidence that some forms are more difficult than others. Given insufficient information and the need to update these studies, NDE should ask the vendor when these studies will be conducted and provide the data for Nebraska if the screener is approved for use statewide.

Validity. A 2023-24 intercorrelation study between subtests for each grade and administration window was reported, but no documentation of the sample was provided. The correlations between subtests ranged from moderate to high, suggesting that subtests measure distinct but related skill sets. External validity studies were concurrent and predictive validity studies with GRADE total score and aReading. It is unclear when these occurred. The GRADE study was likely conducted in 2012-13, when the cut-score study with GRADE occurred. This was before CBM reading scores were included in the earlyReading composite for 1st-grade students. Samples were found to be reasonably diverse. GRADE was correlated with subtest scores for kindergarten and 1st grade with correlations that suggested similar constructs. Correlations between aReading with the earlyReading composite for kindergarten and 1st grade provided evidence that the two scores measured similar constructs in all administration windows. The vendor mentioned additional validity evidence in the form of growth studies demonstrating high correlations across scores across all three administration windows, but no documentation of these studies was found. As suggested for reliability studies, NDE should inquire about vendors' plans to update these studies and to provide state-level data if the screener is selected for statewide use.

Fairness. The vendor claims that earlyReading is context-free and therefore subgroup-difference analysis is unnecessary. It is not easy to prove the claim that any educational measure is entirely free of context. However, with sufficient data, the vendor could have provided a comparison of subgroup results from reliability and validity studies to support this conclusion. Evidence that earlyReading is equally precise and accurate across genders and ethnicities would strengthen the context-free argument.

c. Usability of FastBridge Suite

A usability summary for all three FastBridge screening measures (earlyReading, aReading, and CBMreading) is provided here.

Administration and training time and cost considerations. All earlyReading subtests are administered 1-on-1, and each takes up to 11 minutes per student. All subtests can be administered on paper or on a digital device, but scores must be entered into a digital device for reporting. aReading is administered on a computer and takes 10-15 minutes. CBMreading is administered 1-on-1 and takes 6-8 minutes per student. The cost for the entire FastBridge Suite (earlyReading, aReading, and CBMreading) ranges from $7.25 to $8.50 per student. This fee includes the screeners and the reporting system. It also includes asynchronous online training and support resources, with unlimited phone, live chat, and email support, a help article database, a vendor blog, and a resource portal. Training is recommended but not required and is provided as a 60- or 90-minute webinar or on-site. However, given the subjective nature of administration and scoring for earlyReading and CBMreading, we would recommend training. See Tables 1 and 2 for details.

Reporting, access, and data. Sample reports were provided and found to be useful. Family score reports are also available. Intervention reports provide relevant, specific topics and lessons. A school district can establish role-based access for teachers, administrators, and students, but there is no mention of FERPA compliance. A vague statement is made that data extracts can be downloaded and provided to NDE, but the format of those extracts is not mentioned.

Accessibility and accommodations. No mention of universal design or ensuring digital tests are compliant with ADA 508. A list of compatible modifications is included in the documentation, but these features are not built in. Specifically, the vendor addresses the importance of not offering untimed accommodations for timed tests used in screening decisions.

Dyslexia screening. The vendor claims that FastBridge was designed to include screening for students with characteristics of dyslexia. The vendor states that earlyReading effectively screens readers, including those with dyslexia, and that RAN measures are available with CBMreading. Hence, measures to screen for dyslexia appear to be supplemental to the screening measures. Psychometric evidence supporting the screener's use was not cited. However, this evidence could have been missed. It was not easy to locate the cited evidence in the appendices submitted by FastBridge.

FastBridge aReading (2nd & 3rd Grades)

a. Appropriateness of FastBridge aReading – ***Partially met***

Intended interpretation and use. aReading is a computer-adaptive test with multiple-choice or fill-in-the-blank items that produces an overall score recommended for screening 2nd- and 3rd-grade students, along with CBMreading.

Content coverage and alignment. In 2019, an independent organization aligned the aReading item pool with the Common Core and various state standards, including the 2014 (not the 2021) NE CCR ELA Standards. Results suggest aReading aligned well with standards at all grade levels even though it is only recommended for screening 2nd- and 3-rd grade students (Appendices A and B)**.** Content covered by aReading appears appropriate for screening in 2nd and 3rd grade, except that it does not assess phonological awareness, which is essential for screening reading at 2nd grade. earlyReading assesses phonological awareness, but earlyReading has not been validated for use with 2nd-grade students. Reading content is delivered both auditorily and visually, but students do not respond orally. However, screening decisions also include CBMreading which includes oral responses.

Administration and scoring model. How content is managed in the administration of the computer-adapted test (CAT) is not documented, which is a concern. It appears to be based on item difficulty and not content. The concern is that the algorithm could administer different content to different students for screening purposes and affect composite-score screenings, as they may not be based on comparable content across students. More details about adaptive algorithms are provided in Appendix C.

b. Technical Adequacy of FastBridge aReading – ***Partially met***

Norms. aReading norms were gathered in 2017-18. Given that other vendors reported declines in students' reading performance after the pandemic, a post-pandemic update is preferred. However, the vendor indicates that local norms are automatically generated when at least 70% of students are tested. NDE may want to inquire about this option and whether norms can be provided at the state, district, and school levels[3]. The study involved a large sample of 2nd- and 3rd-grade students that appeared demographically representative with respect to gender, race/ethnicity, and free-reduced lunch, with a Midwest representation. The vendor claims the norms were collected from a full range of

[3] It is possible other vendors may offer the option of local norms but it was not documented in the materials submitted.

abilities, but no demographic data for SPED, EL, or gifted were documented. Norms were reported for 2nd- and 3rd-grade students across all three administration windows.

Cut scores. Cut scores for aReading, according to NCII, were established in 2010-11 using the MAP and the Gates-MacGinite Reading Test as external criteria. The study samples were small and limited to two school districts. Gender, race/ethnicity, SPED, and EL demographics were documented and were nearly proportional to Nebraska's student population. AUC, Sensitivity, specificity, and accuracy for 2nd and 3rd grades were reported, although not for each administration window. Sensitivity and specificity values were high to moderate.

Vendor referred to NCII for a classification accuracy study that occurred in 2017-18 and 2018-19 with Georgia state end-of-the-year tests. Demographic information for this sample was not documented. The 40th PR was used for aReading for these studies, but it is unclear which cut was used for the external criterion. Reported indices were moderate to strong. Sensitivity exceeded the minimum for all 2nd-grade administration windows and for the spring administration window for 3rd grade. Third grade fall and winter were moderate. NPVs were all high (above .90) for both grades. NDE should ask when the vendor plans on updating these studies and whether they can provide state-level results.

The vendor states in the RFI that classification consistency evidence does not apply to a computer-adaptive test because there are no alternate forms. This statement does not acknowledge that adaptive assessments administer tests composed of different items to different test takers. That is, essentially administering different forms to other test takers. Therefore, we disagree with this conclusion and believe that evidence of decision consistency is crucial to establishing the stability of the screening decision in addition to the reported marginal reliability. [4].

Reliability. Internal consistency and test-retest reliability were provided for large samples collected in 2018. Internal consistency reported demographic data for groups based on gender, race, and ethnicity. Average SEM and CSEM values at the 5th and 95th PRs are reported. Second- and 3rd-grade reliability results supported the internal consistency of aReading scores and aReading scores over time although representativeness of the sample beyond gender and race/ethnicity is unknown. The reported CSEM provides some insight into precision of scores but CSEM at cut scores was not reported.

Validity. No internal validity evidence was provided. External validity evidence included

[4] Marginal reliability is a measure of internal consistency in IRT-based measures that indicates how consistently items work together in the measurement of a trait.

concurrent and predictive validity studies using three external measures (MAP, GATES, and the Comprehensive Test of Basic Skills). However, these studies were conducted 15 years ago in 2011. Samples were of reasonable size, and demographic data on race/ethnicity, free/reduced lunch, EL, and SPED were provided but a large percentage of white students (70%). Correlations with the three criterion variables were high. As additional evidence of validity, the vendor mentioned growth studies demonstrated high correlations across scores for all three administration windows, but no documentation of these studies was found.

Fairness. A 2019 DIF analysis was conducted for aReading; however, because only 40% of items had sufficient data, the data for 2nd and 3rd grades were combined. The vendor evaluated approximately 200-250 items for each subgroup comparison. The studies addressed potential discrepancies by gender and by race (white/African American, white/Hispanic, and white/Asian). No items in the 2nd and 3rd grades exhibited DIF.

In the 2022-24 school year, grade 2 subgroup analyses were conducted to compare the concurrent validity and classification accuracy results with DIBELS 8 across subgroups defined by gender, race/ethnicity, low income, EL, and SPED. There was a small amount of variation, but nothing of note. Sensitivity indices for DIBELS 8 were above the minimum and were similar across all subgroups and administration windows. It is unclear why these analyses were not conducted for the 3rd grade. Marginal reliability across subgroups was also compared; results were similar, although it was unclear which grades were included in the analysis.

c. Usability of FastBridge aReading

See usability summary for aReading above under earlyReading.

FastBridge CBMreading (1st – 3rd Grades)

a. Appropriateness of FastBridge CBMreading – ***Met with weaknesses***

Intended interpretation and use. CBMreading is a measure of oral reading fluency recommended for students in Grade 1, starting in the winter, as well as in 2nd and 3rd grades. According to the *Using Renaissance FastBridge Assessments for NebraskaREADS Initiative*, for screening purposes, CBMreading scores are included in the earlyReading composite score for 1st grade and as a secondary assessment to aReading scores for 2nd and 3rd grades.

Content coverage and alignment. The skills blueprint for CBMreading was aligned with the 2014 (not 2021) NE CCR ELA Standards, but only reported results for 1st grade. Results for 2nd and 3rd grade were not found. CBMreading is not administered until 1st grade in winter, but assessing oral reading fluency for 1st-grade students in the Fall was considered preferable. An optional CBM-comprehension measure that includes recall (or retelling) is recommended if the student has unexplained poor reading comprehension, which would be useful for screening decisions.

Administration and scoring model. Three reading passages are administered and produce three scores for the number of words read correctly, the number of errors, and accuracy computed as (total words read correctly – errors)/total words read correctly. It appears the total score used in screening is the median of the raw scores from three reading passages. Given that score reporting seems to be based on raw scores, the extent to which different "forms" or passages might be administered to other students and warrant equating procedures for the scores for these alternate passages is unclear.

b. Technical Adequacy of FastBridge CBMreading – ***Partially met***

Norms. CBMreading norms were collected in 2017-18. Given that other vendors reported declines in students' reading performance after the pandemic, a post-pandemic update is preferred. However, the vendor indicates that local norms are automatically generated when at least 70% of students are tested. NDE may want to inquire about this option and whether norms can be provided at the state, district, and school levels[5]. A large sample of 1st-, 2nd-, and 3rd-grade students was collected and appears demographically representative with respect to gender, race/ethnicity, and free-reduced lunch, with a Midwest representation. The vendor claims the norms were collected from a full range of

[5] It is possible other vendors may offer the option of local norms but it was not documented in the materials submitted.

abilities, but no demographic data for SPED, EL, or gifted were documented. Norms were reported for 1st-3rd grade for all three administration windows.

Cut scores. A cut-score study for CBMreading was conducted in 2011-13 using MAP and the Test of Silent Reading Efficiency and Comprehension. The study samples were small and limited to Minnesota school districts. Race/ethnicity, free/reduced lunch, and SPED were reported and aligned with Nebraska rates. Gender and urban/rural were not documented. The sample only included students proficient in English. AUC, Sensitivity, specificity, and accuracy were reported for all three grades at the 20th and 30th PR but not for each administration window. Sensitivity with MAP in the 3rd grade was low.

An updated classification-accuracy study was conducted in 2018-19, using MAP Growth Reading as the end-of-year criterion. The sample included students from seven midwestern states, including NE, but the demographics of that sample were not documented. The 40th PR was used as the cut score for CBMreading, even though the 20th and 30th PR were identified in the cut score study, and the 15th PR was used for the end-of-year score on MAP Growth. Sensitivity and NPV values for the 2nd and 3rd grades exceeded the minimum thresholds. First-grade sensitivities were moderate to low, as expected, given that MAP Growth may not be an appropriate criterion for first-grade reading screening. Specificities exceeded the minimums for all three grades, but PPV values were below the minimum for all grades except the 2nd-grade winter administration, which suggests increased probability of overidentification. Given these issues, NDE should inquire about the vendors plan to update these studies and if providing statewide results is possible.

The vendor stated in their RFI response that evidence of classification consistency is not applicable, as the screener does not use multiple forms. However, the RFI and the submitted Technical Manual state that CBM-reading students read three passages for screening and receive the median of their scores across the three passages. But we are unclear whether a single form with the same passages is administered to all students, or whether three of the 20 unique passages are used. If the latter is the case, evidence of score equivalence across passages or equating of scores should be provided. In addition, the reporting of alternate-forms reliability appears to contradict claim that multiple formas are note used. Therefore, we disagree with this conclusion, as the potential use of those forms across students could vary.

Reliability. Test-retest and alternate forms reliability conducted in 2018-19 were provided for CBMreading. Large samples for both studies and demographics for alternate-form reliability were reported, including variables such as gender, race, and ethnicity. Reliability results for 1st-3rd grades supported the consistency and precision of the CBM reading score. However, as mentioned, alternate-forms reliability does not provide direct evidence that the passages are of equal difficulty. The vendor should indicate when they plan to update these studies and provide state-level results if the screener is approved for use statewide.

Validity. No internal validity evidence was provided. External validity evidence included concurrent and predictive validity studies using three external measures (AIMSweb, DIBELS Next, MAP) as well as aReading. It is unclear when the studies were conducted, and the demographics for the samples were not documented. Correlations with external variables, as well as aReading, were high, providing evidence that the tests measure similar constructs. The vendor mentioned growth studies with high correlations across scores for all three administration windows as evidence of validity, but no documentation of these studies was found. The vendor should indicate when they plan to update these studies and provide state-level results if the screener is approved for use statewide.

Fairness. The concurrent and predictive (classification) validity results for CBMreading with aReading as the criterion were compared across subgroups by gender and race/ethnicity for each grade. It is unclear when this study occurred. Demographics for 1st – 3rd grade were reported and proportional. Although there were some differences in the correlations across subgroup, they were not considered meaningful. Depending on when this study was conducted, a updated fairness study may need to be conducted.

C. Usability of FastBridge CBMreading

See usability summary for CBMreading above under earlyReading.

Amira Reading Mastery (ARM) Universal Screener

a. Appropriateness of Amira Reading Mastery (ARM) Universal Screener – ***Partially met***

Intended interpretation and use. Amira's assessment was designed to serve as a universal screener and progress monitor for early literacy skills.

Content coverage and alignment. Amira had a panel of four experts align the *2021 NE ELA CCR Standards* to skills statements and representative items. Vendor claims the panelists had no conflicts and states that the panel found 100% alignment between the skills statements and the standards, but this alignment was at a very broad level. There appears to be good coverage of the skills needed to screen across all grade levels (See Appendices and B). However, there are concerns regarding adaptive algorithms and pattern scoring as discussed below. The vendor indicates that multiple modes are used to deliver content (oral, audio, written, and visual) and that students may respond orally or through selection. AI speech recognition software is used to score students' oral responses.

Administration and scoring models. We identified several possible concerns regarding how the screener is administered and scored.

First, we did not find clear documentation of how the adaptive algorithm selects items for administration. A statement from the vendor suggests that item difficulty and discrimination characteristics are used. Still, it is unclear whether additional aspects of the algorithm govern content exposure, ensuring that all intended constructs are represented in the items administered. However, the reporting of subscores across specific skill domains suggests that content representativeness may be preserved to some extent. The lack of clarity regarding whether the adaptive algorithm accounts for content representativeness prompted a rating of "partially met," as students may not receive the same level of content coverage during test administration. This absence could affect the reliability and validity of reported subscores and possibly composite scores and composite score screenings, as they may not be based on comparable content across students.

Another issue that prompted the partially met rating is the scoring pattern Amira uses, which weights some items more than others. This scoring approach adds another layer of complexity, making it difficult to know whether composite scores are comparable across students. The specific pattern-scoring method is EAP scoring for the 2PL IRT model, where scores are produced not only from students' patterns of correct/incorrect responses on the items they take, but also from a prespecified prior distribution that weights their test performance. This pattern scoring raises at least two additional questions about the limitations of comparability. First, there are questions about the limitations of cross-grade comparisons, because overall scores are weighted to "grade-specific prior means that reflect typical ability levels for each grade" (p. 57), where different means appear to be used for grades K-1 (the unified base scale of Amira's vertical scale, p. 58), grade 2, and grade 3. If a grade 2 student performed as well on the same items as a grade 1 or K student, their performance would be weighted to a (presumably) higher prior ability mean than that of the grade 1 or K student, and they would likely receive a higher score. This potential issue seems to undermine cross-grade comparability even with Amira's established cross-grade vertical scale.

A third question concerns subscores, which the Technical Manual states are computed using students' "overall theta (ability) estimate" as the prior mean (p. 57). To the extent that two students take the same items and have the same performance on a subscore but differ in their performance on the rest of their tests, their identical subscore performance would be weighted to lower and higher prior distributions, and the subscores would vary. In other words, subscores partially reflect total test performance, so interpreting and comparing students' subscores requires awareness of each student's total test performance.

A more detailed discussion of these concerns can be found in Appendix C. We would encourage NDE to ask the vendor the questions posed in Appendix C regarding how Amira's algorithm addresses content and pattern scoring in the administration and scoring of the screener.

b. Technical Adequacy of Amira Reading Mastery (ARM) Universal Screener – ***Partially met***

Norms. Amira norms were gathered for all grades and all three administration windows in 2024-25. A second sample was collected in 2025 to verify those norms. The initial norm sample was large and appears to include representation from all 50 states, but other demographic characteristics were not reported. The vendor states they stratified districts by geographic region (Northeast, Midwest, South, and West) and community types (urban, suburban, rural). Then, within districts, they stratified by socioeconomic background and possibly race/ethnicity. Those demographics were then weighted to match national demographics. Weighting the sample data to improve the representativeness of the norms is appropriate, but ideally, sharing data summarizing the differences between the original, unweighted samples, the nationally representative targets, and the final weighted samples across the chosen background variables would have been preferable. There was no mention of gender representation, EL, gifted and talented, or SPED. The vendor states that norms are reported as percentile ranks by grade and across all three administrative windows, but no documentation of the norm tables was provided.

Cut scores. Cut scores for the early literacy screener were recently set using the 2023-24 sample and MAP Growth as the end-of-year criterion. The vendor states that they used a large, nationally representative study sample, but the sample demographics were not documented. In addition, a separate classification-accuracy study was not conducted. Sensitivity and specificity indices for the 30th PR on Amira and the 20th PR on MAP Growth were above the minimum for all grades and administration windows, except for kindergarten Fall, which was just below the minimum. However, given concerns with MAP Growth discussed below, we question whether it is an appropriate criterion for kindergarten and 1st grade (see MAP Growth summary below). Positive predictive value (PPV) and negative predictive value (NPV) were not reported; NPV is particularly important given the desire to avoid false-negative cases. NDE should request that Amira provide a separate classification-accuracy result with a different criterion to cross-validate the cut scores set using MAP Growth if it is selected for state-wide use.

A test-retest study for the fall window in 2025 was presented as evidence of classification consistency. However, no sample data or information on how the data were collected was shared, making it difficult to evaluate the study's quality based on the documentation provided. The percentage agreement for screening cut scores in the fall administration window for all four grades was high. Analyzing classification consistency for the winter and spring administration windows for all grades would strengthen evidence for the screener.

Reliability. Multiple forms of reliability were provided, including internal consistency (year unknown), test-retest reliability (2022-23), alternate-forms (2022-23), and inter-rater reliability with AI scoring (year unknown). Demographic information for the large sample was not reported, but a subsequent subgroup analysis on fairness indicates some diversity in the sample. Reliability results for all studies were strong, although only the SEM was reported. CSEM would provide more substantial evidence of precision at appropriate cut scores.

Validity. The vendor provided internal validity and validity with external variables. For internal validity, correlations among subscales were reported for student samples from unspecified years. Some of those intercorrelations were very high, suggesting possible redundancy. However, redundancy affects instructional planning more than screening. The CAT algorithm was described. Although evaluations of the fit of the chosen 2PL IRT model are generally presented alongside scale transformation and evaluation procedures, no specific results from these procedures are provided, aside from some DIF results and some means and standard deviations of SEMs by grade. In addition, the extent to which the difficulty of the item pool matches the lower and higher abilities of typical test takers is not described. Other vendors compared difficulty and proficiency distributions (i-Ready

Diagnostic) or compared ranges of item difficulty and student ability (FastBridge aReading). The vendor conducted multiple concurrent and predictive validity studies in 2022-24 with a large, diverse sample, but no demographic information was shared. The criterion variables used in the concurrent study were MAP for all grades and i-Ready Diagnostic Reading for 3rd grade only. It is unclear whether MAP Growth or MAP Reading Fluency was used, as the description refers to the MAP Reading Assessment. MAP Growth is likely more appropriate for 2nd and 3rd grade, while MAP Reading Fluency is more appropriate for K and 1st grade. Correlations were high. Predictive validity evidence between fall Amira scores and spring MAP and i-Ready Diagnostic was also provided. Predictive correlations were also high. Predictive validity between the fall Amira subtests in kindergarten and 1st grade and the MAP subtests (appearing to be from MAP Reading Fluency) was also provided, and supported similar constructs.

Fairness. DIF and subgroup comparison studies on reliability and classification accuracy were reported for kindergarten through 2nd grade. It is unclear why 3rd grade was excluded from these analyses. A DIF study was conducted in 2020 comparing multiple subgroups (gender, Black/White, Hispanic/White) within each grade. A summary table listed the number of items exhibiting DIF. The vendor reports removing DIF items from the calibration set. Six items exhibited DIF in the white/black comparison. Several flagged items involved content that is foundational and unlikely to be biased (e.g., letter-sound fluency, initial-sound segmentation, initial-sound deletion task). Similar results were observed among 1st-grade Hispanic students. The vendor did not provide documentation addressing possible reasons for these differences, leaving us to question whether they resulted from AI scoring or other factors.

Subgroup analyses of studies on reliability, predictive validity, and classification accuracy were conducted. Although the variety and breadth of subgroups compared were good (gender, EL, SWD, race/ethnicity, home language, and pre-K education), reporting of results was inconsistent. It was unclear why this was the case. For example, internal consistency was reported by gender for each grade level, whereas no other results were reported at the grade level. Differences in classification accuracy results were only reported for students with different language backgrounds and pre-kindergarten education, and reliability result differences were reported only for gender

c. Usability of Amira Reading Mastery (ARM) Universal Screener

Administration and training time and cost considerations. The Amira assessment is group-administered, and the administration time varies by grade. Kindergarten takes 14-17 minutes; 1st grade takes 20-25 minutes; 2nd grade takes 16-20 minutes; and 3rd grade takes 13-17 minutes. A digital device (e.g., a computer or iPad), along with headphones, speakers, and a microphone, is required to administer the test to each student. The licensing fee is $5 per student and includes a screener, progress-monitoring assessments, reporting and dashboard tools, and onboarding support. Training is not required, as the system administers and scores the assessment; however, a no-cost 45-minute self-paced online training covers setup, administration, data interpretation, and instructional decision-making. On-site coaching, live workshops, or certification programs are optional and available for purchase, but the vendor does not consider them necessary for successful implementation. See Tables 1 and 2 for more information.

Reporting, access, and data. The vendor claims score results can be shared at the individual, classroom, school, and group levels. Example reports were provided with different views regarding process, skills, and student summary. Parent reports are available, and audio recordings can be shared with families. Role-based controls govern access, and a secure login is required after permission is granted. Data can be downloaded in CSV format, and a note indicates the availability of custom extracts. Access to data is reported as FERPA-compliant for only the designated purpose and personnel.

Accessibility and accommodations. Amira meets the 508-compliant Web Content Accessibility Guidelines (WCAG) 2.1 Level AA. They provide a document listing conformance level with those criteria, and all relevant criteria were supported except for adjustable timing, which was partially supported. The vendor also claims the use of Universal Design for Learning principles. A listing of available accommodations was also provided in the User manual.

Dyslexia screening. A dyslexia screener is embedded in the measure, and evidence of classification accuracy for the dyslexia screener was provided.

Acadience Reading

a. Appropriateness of Acadience Reading – ***Met with weaknesses***

Intended interpretation and use. Acadience Reading is a universal screening and progress monitoring assessment that measures the acquisition of early literacy skills from kindergarten through sixth grade.

Content coverage and alignment. The vendor provided a table summarizing the alignment of the screener with the *2021 NE ELA CCR Standards*, but there was no documentation of the alignment process. The vendor claims that its screener directly aligns with most standards, but does not document whether each standard was directly or indirectly aligned. Appendices A and B suggest that key screening skills may be inconsistently assessed across grade levels. No vocabulary is included in the screener, but the use of multiple passages to assess Oral Reading Fluency is a strength. Content is presented orally and visually, and students respond orally.

Administration and scoring model. Screening is based on a composite score computed by summing the raw scores of designated subtests for each grade. These include First Sound Fluency, Letter Naming Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, Oral Reading Fluency, Retell, and Maze. Although randomized ordering and multiple passages are used, there is a possibility that raw scores from those variations may not be comparable (See Appendix C). Although Word Use Frequency-Revised is mentioned several times in the documentation as a vocabulary measure, it is described as experimental. It does not contribute to the composite screening score.

b. Technical Adequacy of Acadience Reading – ***More evidence needed***

Norms. A large sample of norms was collected between 2015 and 2019. Although the samples were demographically representative with respect to gender, race/ethnicity, free/reduced lunch, and school location (urban, suburban, and rural), the data were collected 6-10 years ago, before the pandemic. Given that other vendors found a downward shift in student reading performance after the pandemic, a post-pandemic update is preferred. The vendor recommends using local norms, but doesn’t indicate that it provides test users with support for gathering and generating them. Norms are reported for each relevant subtest and composite for each grade and all three administration windows.

Cut scores. A 2009-2010 predictive cut-score setting study was conducted using the 40th percentile of the PR on GRADE (Group Reading and Diagnostic Evaluation) as the external criterion. Vendor claims that data were collected from a sizable number of students in kindergarten through 6th grade across 13 schools in five districts in the north central Midwest and Pacific West; however, the student sample was predominantly white (94%). Further, only a third of those students also completed the GRADE test, so the study sample size for K-3 grades is unclear. The vendor claims that the study included students with disabilities (SWD) and ELs, but the demographic data for those groups were not reported. Receiver operator characteristics (ROC) (e.g., Area under the curve, sensitivity, and specificity) were used to adjust cut scores. The vendor reports cut scores for all grades, administration windows, the composite, and the subtests.

No separate, up-to-date classification-accuracy study was reported. Classification indices were reported only for the 2009-10 cut-score setting study with GRADE as the external criterion. Although the screener appears not to have changed much over the last 15 years, given shifts in student demographics and performance since that time, the vendor should update its classification-accuracy data and norms. Classification accuracy indices for the subtests and composites were reported for each grade level but not for each administration window. Sensitivity indices were very low for the kindergarten composite and moderate for the composite score in grades 1-3. NCII reported sensitivity, specificity, NPV, and PPV for all three administration windows across grades 1-3 and only for the fall and winter administration windows in kindergarten. NPV, PPV, and specificity exceeded minimum expectations, but sensitivity did not, except for 2nd-grade in the fall. Low sensitivity was reported for kindergarten in fall and winter, 1st grade in fall, and 3rd grade in the fall, winter, and spring. All other sensitivity indices were below the minimum and considered moderate.

Reliability, Validity, and Fairness. The team discontinued the technical adequacy review of the Acadience Reading measure due to outdated norms and classification-accuracy studies, as well as poor classification accuracy results.

C. Usability of Acadience Reading

Given the discontinuation of the Acadience Reading review due to concerns about norms and classification accuracy, the documentation for usability was not reviewed. Information on usability can be found in the vendor’s response if needed. Information about the time, cost, and administration of the measure can be found in Tables 1 and 2.

MAP Reading Fluency

a. Appropriateness of MAP Reading Fluency – **More evidence needed**

Intended interpretation and use. MAP Reading Fluency is described as assessing foundational reading skills and oral reading fluency, but the vendor did not explicitly mention screening as a purpose in their RFI response.

Content coverage and alignment. No independent review of alignment between test content and Common Core or NE CCR ELA Standards was reported. However, the vendor documents a rigorous development process to ensure alignment with the Common Core. A review of key reading skills for screening appears to be addressed at all grade levels (Appendices A and B). However, some of those skills are assessed indirectly rather than directly. That is, items that require students to apply prerequisite skills that are not directly measured but are necessary for correctly responding. For example, students might be asked to substitute phonemes in a word, even though substitution is not an indicator; however, deletion is an assessed skill required for substitution. The vendor states that test content is presented orally and visually, and students can respond orally through selection.

Administration and scoring model. MAP Reading Fluency produces a flagged screening outcome based on the extent to which three domain scores (Phonological awareness, Phonics & Word Recognition, and Language Comprehension) and the Sentence Reading fluency task score predict the 10th PR's performance on MAP Growth. A dyslexia subtest score is also mentioned, but it is unclear how it contributed to the flagged outcome. MAP Reading Fluency offers two administration options. Either a teacher selects which test to administer (Foundation Reading Skills or Oral Reading Fluency), or a sentence-reading fluency measure routes students to Foundational Skills or Oral Reading based on a raw score rather than an ability estimate. There is some concern that this routing may produce inequity among students if some have more background knowledge of the content than others, which can affect students' scores in accuracy, rate, and comprehension. Better background knowledge can positively impact students' scores, while a lack of background knowledge can negatively affect them. Given this concern, fairness studies comparing how the sentence reading fluency task routes students from different demographic groups are warranted.

It is also unclear from the documentation how MAP Fluency's adaptive functioning works. MAP Fluency uses a multi-stage administration, which may help ensure a more consistent delivery of content across these three domains. However, data and vendor statements suggest that differences in content delivery may occur among students within the same grade. For more details on this concern, refer to Appendix C.

b. Technical Adequacy of MAP Reading Fluency – ***More evidence needed***

Norms. The most recent norming study for MAP Reading Fluency was conducted in the 2018-19 school year. MAP Growth norms were updated in 2022-24, and the data indicate substantial shifts in performance following the pandemic; therefore, it is likely that MAP Reading Fluency norms also need updating. The number of students in the norm sample for each grade level declined significantly in 2nd and 3rd grades, as MAP Reading Fluency is no longer recommended for students once they demonstrate fluent reading. A large sample for 2018-2019 norms was collected; however, the vendor cautions that these norms should be interpreted only as user norms, not as national norms. Demographic data on gender and race/ethnicity were provided, but not for SPED, ELL, free/reduced lunch, or geographic location.

MAP uses a model-based approach to calculate norms, assuming normality in achievement and growth. Where other screeners calculate percentiles directly from their normative sample data or from a weighted version that improves national representativeness. The MAP approach produces percentiles from a model rather than from the observed data, with a more limited fit. Accuracy evaluations of the fit of normal distributions and of percentiles derived from normal distributions usually indicate better fit in the middle of the distribution and worse fit at the highest and lowest scores.[6] Given that the RIT score at the 10th percentile on MAP Growth is used to flag outcomes for screening with MAP Reading Fluency, more information is needed about the accuracy of the model-based norms for low cut scores. Norms are reported for K-3 and all three administration windows, but only for the domain scores.

Cut scores. Cut scores for identifying reading difficulties using MAP Reading Fluency are set by predicting, based on three MAP Reading Fluency domain scores and the Sentence Reading Fluency score, which students are likely to perform at or below the 10th PR on MAP Growth at the end of the year. The predictive model was fine-tuned using ROC analyses on MAP Growth data to maximize classification accuracy for each grade and administration window. RAN scores were not included in this predictive model, apparently due in part to "insufficient data" (p. 82), yet they are described as supplementing the model. Details about this predictive model and the cut scores set for the Foundational Skills and Sentence Reading Fluency scores were not documented, so we were unable to evaluate how these scores are combined to produce the flagged outcome or the extent to which the RAN scores influence the screening decision.

[6] See the Q-Q plots in the following report, https://www.nwea.org/uploads/2021/06/2022-English-MRF-Foundational-Skills-Norms-Report-2022-07-19.pdf

A separate classification-accuracy study was not conducted. The vendor reported indices only after adjusting the predictive model for end-of-year performance on MAP Growth, which is not an ideal criterion because it is not an external criterion for validating cut-score decisions. Reported sensitivity and specificity for each grade and administration window exceeded the minimum, except for the sensitivity for kindergarten and 2nd grade in the fall administration window. NPV and PPV were not reported. NPV is particularly important given the desire to avoid false-negative cases

Classification consistency results were generally described in the RFI form. However, the referenced technical report included test-retest reliability rather than actual consistency results.

Reliability. Marginal and test-retest reliability were assessed in 2020-21 using large, reasonably representative samples. Analyses were conducted at the Foundational Skills domain level, but the scoring rules remain unclear, particularly for screening implications. Although the means and standard deviations of SEMs were provided by grade and season, we did not know the cut scores used for each foundational skills domain and for sentence reading fluency in making the flagged-outcome decision. Hence, we were unable to evaluate the precision of the scores at those cut-score points, which is critical given our concerns about the model-based approach to norms.

Validity. No relevant evidence of internal validity was provided, which is particularly important given the use of a multivariate model to flag outcomes based on three foundational skill domains and sentence-reading fluency. To assess validity based on relationships with other variables during 2017-2020, they conducted studies correlating MAP Reading Fluency with MAP Growth. They conducted analyses of classification accuracy comparing DIBELS Next with Sentence Reading Fluency and Scaled Words Correct Per Minute (SWCPM). MAP Growth is not an appropriate criterion, given that the two tests are designed to align closely and that there are concerns about using it for reading screening. Although the results from the DIBELS Next study were supportive, they provided evidence only for a portion of the content covered on MAP Reading Fluency. The vendor also reported human-machine agreement studies with positive results, but these applied only to the oral reading fluency portion of the screener.

Fairness. Fairness and bias considerations are undertaken in item development and review. The process is rigorous, and any flagged items are revised or rejected. The only reported comparative or statistical analysis was a subgroup difference in human-machine agreement studies. Still, it was not reported at the grade level, so it provided very little information for evaluating the fairness of this component or the screening measure.

c. Usability of MAP Reading Fluency

Administration and training time and cost considerations. MAP Reading Fluency is administered in groups and takes 20-30 minutes. Each student needs a computing device to take the test, along with a headset with a microphone. They suggest testing in smaller groups of 8-10 students to reduce background noise. The annual licensing fee for MAP Suite (both MAP Reading Fluency and MAP Growth) is $13 per student and includes access to both measures, account management, support services, and standard reporting. Training is not required, but optional professional learning is available to improve implementation and data literacy. MAP Suite is also available with the Coach add-on, an AI tool that uses MAP Reading Fluency results to provide personalized, one-on-one reading tutoring. MAP Suite with Coach is available for $20 per student.

Reporting, access, and data. The vendor claims score results can be shared at the individual, classroom, school, and group levels. Example reports were provided, including multiple data types: screener outcomes, domain scores, and performance levels. A report specifically for parents is not generated, but the vendor indicates that a visual indicator report could be shared with families. The vendor did not respond to questions about data access and privacy, as well as embedding data into local data management systems.

Accessibility and accommodations. NWEA shared documents demonstrating its commitment to using Universal Design, as outlined in the Council of Chief State School Officers (CCSSO) Accessibility Manual, to improve the accessibility of its measures. A list of available embedded and non-embedded accommodations was also provided. There was no mention of meeting 508-compliant Web Content Accessibility Guidelines (WCAG), which is common for screeners that are heavily web-based

Dyslexia screening. A dyslexia screener is available and appears to assess appropriate skills. Student performance is flagged in the report as suggesting risk factors for dyslexia. However, as discussed above, it is unclear whether or how the RAN score contributed to the flagged reading screening outcome. The vendor indicated that a predictive model is used to determine flagging, but no psychometric evidence supporting the dyslexia screener flagging results was referenced.

MAP Growth

a. Appropriateness of MAP Growth – ***More evidence needed***

Intended interpretation and use. MAP Growth consists of two measures (K-2 and 2-5) that measure different types of reading skills. The vendor describes MAP Growth as a longitudinal, adaptive assessment that provides precise composite scores for tracking reading achievement and growth over time. Screening is not mentioned, although reading screening reports were cited in the submission based on MAP Growth K-2. The test guidance does not clarify when to use MAP Reading Fluency and/or MAP Growth K-2, as the two measures assess similar skills but differ in administration format. The vendor's RFI response was also confusing, as it referenced a MAP Growth Early Learning Benchmark that was not included in any of the submitted documentation. It was unclear whether the MAP Growth Early Learning Benchmark is a new measure or a combination of MAP Reading Fluency and MAP Growth. Although we believe it is the latter, more information is needed to clarify these statements, along with evidence demonstrating that using both MAP Reading Fluency and MAP Growth K-2 (or 2-5 if recommended) is better for reading screening decisions than using either alone.

Content coverage and alignment.

The vendor recommends using MAP Growth K-2 with kindergarten and 1st-grade students, and with 2nd-grade students with a MAP Growth 2-5 RIT score of 170 or lower. MAP Growth K-2 claims to assess content for students in early grades who are pre-, emergent, or beginning readers, while MAP Growth 2-5 should be administered once students demonstrate fluent reading with comprehension. If MAP Growth 2-5 is intended for students who demonstrate fluent reading with comprehension, it is measuring skills that develop after foundational reading skills have already been acquired. The vendor mentioned several alignment studies that may have involved NE CCR ELA Standards. However, the cited documentation is unclear about what information was gathered. In addition, there was no documentation for the process or results. Appendices A and B suggest a lack of alignment with key skills in some areas and some grades. Test content consists of audio, written, and visual materials, and students respond by selecting the correct option.

Administration and scoring model. The MAP Growth tests are item-level adaptive, such that students who answer items correctly (or incorrectly) receive more difficult (or easier) items. The adaptations are based not only on item difficulty, but also on a content-balancing procedure that selects items from the most underrepresented content area. Test scoring is implemented in Rasch Unit Scales, where student abilities are estimated from their responses to test items modeled with Rasch IRT models, and these abilities are converted to RIT scale scores ranging from 100 to 350. Appendix C provides greater detail on how adaptations of MAP Growth and MAP Reading Fluency differ.

b. Technical Adequacy of MAP Growth – ***More evidence needed***

Norms. A large, demographically representative sample of K-3 was collected in 2022-2024 to produce norms. The rationale for updating norms was to align with the transition to a new, enhanced item-selection algorithm, changes in U.S. demographics, and post-pandemic shifts in student performance. Demographics in terms of urban/rural, race/ethnicity, and geographic location were shared, but for the entire sample, not for each grade or administration window. They claim that EL (Spanish-language assessment) and SPED (students receiving accommodations) were included in the sample, but they provide no summary statistics. Norms are reported as means, standard deviations, and percentile ranks for each grade and term, but with little interpretive guidance.

MAP uses a model-based approach to calculate achievement norms, computing the mean and variance and then constructing the distribution on which percentiles are calculated (p. 19 of the 2025 MAP Growth Norms Technical Manual). Where other screeners calculate percentiles directly from their normative sample data or from a weighted version that improves national representativeness, the MAP approach produces percentiles from a model rather than from the observed data, with a more limited fit. Accuracy evaluations of the fit of the constructed (and most likely normal) distributions and of percentiles derived from normal distributions usually indicate better fit in the middle of the distribution and worse fit at the highest and lowest scores.[7] Given that the 35th PR for MAP Growth and the 10th percentile rank for MAP Growth are used for flagging or screening by MAP Reading Fluency, more information is needed on the accuracy of the model-based norms for low cut scores. We looked for this reliability evidence, but it was not provided.

Another need for more evidence in the norms reports for MAP Growth is which MAP Growth Reading test the grade 2 students actually took: the K-2 test or the 2-5 test. The MAP Growth Norms Technical Manual (2025) reports its norms only by overall grade (K, 1, 2, etc.), without identifying grade 2 students taking the K-2 and/or 2-5 Reading tests. Given the content differences between MAP Growth K-2 and 2-5 (discussed above), MAP Growth norms should have disaggregated grade 2 students for each Reading test, at least for some of its norm reporting. This was another factor that led to the “More evidence needed” rating for the Technical Adequacy of MAP Growth.

[7] See the Q-Q plots in the following report, https://www.nwea.org/uploads/2021/06/2022-English-MRF-Foundational-Skills-Norms-Report-2022-07-19.pdf

Cut scores. Cut scores were set using a large sample of Grade 3 students from five states, with state summative assessments as the external criterion. MAP Growth data were collected from students in grades K, 1, 2, and 3, so the comparison with this criterion was longitudinal for kindergarten through 2nd grade. We had several concerns about this study.

- Data was collected between 2014 and 2018, before the 2025 normative update, suggesting this data from this study would not apply to those updated norms. Norms were updated in 2025 due to substantive changes to the test and changes in the normative results (see above). A 2025 memo reported updating the cut score from the 30th to the 35th PR using the same 2021 sample, but we were unable to evaluate the fit because classification accuracy indices for this new cut score were not reported
- Because the criterion variable was summative test results from five states, the vendor conducted linking studies to create a comparable criterion variable. However, the linking studies found substantial variability in cut scores across states, so more information is needed to evaluate the appropriateness of this linking procedure and whether it produced a comparable criterion variable across students who took different state tests.
- The reported correlations between linked scores and MAP Growth were low (all below .70) for K and 1st Fall. We assume a MAP Growth K-2 score was used in this study. The resulting cut score was a RIT score at the 30th PR, which was recommended as a universal cut. Sensitivity and NPV were very low in kindergarten and 1st grade in fall, but exceeded the minimum in 1st grade in winter and spring, and in 2nd and 3rd grade. These predictions of state assessment performance for Grade 3 students from their kindergarten-2nd-grade MAP Growth scores would be expected to have reduced precision, as evidenced by the low correlations for kindergarten and 1st-grade students mentioned above. In addition, because a MAP Growth K-2 score was likely used for kindergarten and 1st-grade students in this study, the results raise concerns about the classification accuracy of MAP Growth K-2, at least with respect to the criterion of state test performance in third grade.
- A second sample from the 2021 study, with scores from Indiana's state summative test, was used to conduct a follow-up classification-accuracy study to confirm the cut score from the first study; however, similar poor results for kindergarten and moderate results for the 1st-grade Fall administration were obtained.

The vendor states in the RFI form that a 2022-2024 classification accuracy study was conducted using Amira as an external assessment tool. This evidence would have been helpful given the 2025 update to the norms and the concerning results reported in previous studies for Kindergarten and 1st grade. Still, no documentation of this study was submitted for our evaluation. In their RFI response, the vendor reports AUC values of 0.70 or higher, but did not report other classification accuracy metrics. The vendor also did not indicate what grades or administration windows were analyzed.

A 2025 study was cited, but it included only a 3rd-grade sample. It used a standard setting to set proficiency cut scores, with no ROC analysis or classification accuracy indices reported or used. A 2025 NE linking study was also cited, but it included only a 3rd-grade sample. It assessed the classification accuracy of MAP growth reading RIT scores in predicting NSCAS proficiency decisions, not screening decisions.

Although the RFI form cites multiple reports as evidence of classification consistency, none of those cited documents provided relevant documentation of classification-consistency studies.

Reliability. Marginal and test-retest reliability studies with alternate forms were cited. Demographic information was not provided, and results exceeded minimum expectations, except for kindergarten on the test-retest with alternate forms. However, these studies were conducted in 2016-17, and given the substantive changes that prompted the 2025 normative update, new reliability studies should be performed. These new studies should be accompanied by SEM tables that provide the CSEM around the recommended cut scores.

Validity. An internal validity study was cited, but no information on the methods, variables, or results was provided. It also appeared that this was conducted only in the 4th and 7th grades. A 3rd-grade NSCAS linking study was provided as validity evidence based on relationships with other variables. The correlation between the two was high, but given that the tests share similar measurement models, we did not find this evidence particularly convincing. No validity evidence was provided for kindergarten through 2nd-grade students; as such, none was available for MAP Growth K-2. As additional evidence of validity, they cite a Virginia study, but no information on methods, variables, or results was provided.

Fairness. DIF studies were conducted in 2016-17 on a random sample of the item pool and in 2011 on the entire item pool. Studies were performed only for the 1st, 2nd, and 3rd grades, not for kindergarten. Sample information and results were not reported by grade, so we were unable to evaluate the results for each grade. DIF analyses were conducted for females and across different races/ethnicities. Results indicated that 2-5% of the items showed large DIF, except for the Asian group. Any items with moderate or large DIF were flagged and reviewed for bias. An update of this study would be appropriate, including comparisons across subgroups on reliability, validity, and classification accuracy.

c. Usability of MAP Growth

Administration and training time and cost considerations. MAP Growth is group-administered, and administration time is 20-40 minutes for younger students and 40-50 minutes for older students (Grades 2 & 3). A digital device (e.g., a computer or iPad) with headphones or speakers is required to administer the test to each student. The annual licensing fee for MAP Suite (both MAP Reading Fluency and MAP Growth) is $13 per student and includes access to both measures, account management, support services, and standard reporting. Training is required to administer, interpret, and use RIT scores. See Tables 1 and 2 for more information.

Reporting, access, and data. The vendor claims score results can be aggregated at individual, classroom, school, district, and group levels. Example reports were provided for school, class, and district levels. However, most reports were student growth and achievement reports, not screening reports. Role-based controls govern access, and data can be downloaded in CSV format. The vendor also states that it has policies in place to protect personal information in accordance with FERPA and other applicable state and federal laws.

Accessibility and accommodations. NWEA shared documents demonstrating its commitment to using Universal Design, as outlined in the Council of Chief State School Officers (CCSSO) Accessibility Manual, to improve the accessibility of its measures. A list of available embedded and non-embedded accommodations was also mentioned. There was no mention of meeting 508-compliant Web Content Accessibility Guidelines (WCAG), which is common for screeners that are heavily web-based

Dyslexia screening. A dyslexia screener is not included in MAP Growth.

STAR Early Literacy & STAR Reading

a. Appropriateness of STAR Early Literacy & STAR Reading – ***More evidence needed***

Intended interpretation and use. STAR Early Literacy is designed to screen pre-K through 3rd-grade students who are beginning readers and do not yet read independently. However, it is typically used in grades K-1. STAR Reading assesses reading achievement and screens for reading difficulties in students typically in grades 2-3.

Content coverage and alignment. STAR Early Literacy assesses skills in three blueprint Domains: Word Knowledge and Skills, Comprehension Strategies and Constructing Meaning, and Numbers and Operations. Star Reading assesses reading skills on five blueprint Domains: Word Knowledge and Skills, Comprehension Strategies and Constructing Meaning, Analyzing Literacy Text, Understanding Author's Craft, and Analyzing Argument and Evaluating Text. The vendor aligned content with the *2021 NE ELA CCR Standards* and provided a summary table of alignment results. Those results suggest extensive alignment with NE Standards at all grades. However, the panel was not independent, and there was no documentation of the process or of whether alignment with each standard was full, direct, partial, or indirect. Both measures were found to cover most key reading skills, although phonemic awareness is not assessed at grade 2, and phonics are not assessed at grade 3. If this is the case, that could be problematic. In addition, both measures administer a limited number of items (22 literacy items for STAR Early Literacy & 34 items for STAR Reading) yet claim to assess 10 skills. The use of limited items raises questions about whether sufficient coverage is provided to evaluate each skill used in the screening decision.

Administration and scoring model. Another notable concern that led to the rating of *More evidence needed* is that the CAT administration, scoring, and score reporting practices for STAR Early Literacy and STAR Reading raise questions about how test content is addressed. In general, assessment practices that yield scores that mean the same thing and can be reported on the same scale (i.e., Unified Scales) require that the assessments reflect duplicate content. This is clearly not the case for content assessed by each measure, despite some overlaps across assessed domains, as noted above. The Star Early Literacy assessment assesses three Blueprint Domains, and the Star Reading assessment assesses five Blueprint Domains. In addition, the item-level adaptations for Early Literacy and Reading are described as reflecting only item difficulty and not content, raising the possibility that content domains with harder vs. easier items would be overrepresented among higher vs. lower ability students. This could affect composite score screenings, as they may not be based on comparable content across students. For more details, see Appendix C.

b. Technical Adequacy of STAR Early Literacy & STAR Reading – ***More evidence needed***

Norms. A large sample of kindergarten through 3rd grade was used to produce norms in 2022-23. Normative data were collected only in the fall and spring administrations, not in the winter administration. Demographics were reported for the fall and spring administrations, but not by grade. Normative data were collected from all 50 states. Demographic characteristics were reported for school size, district SES, urban/rural, gender, and race/ethnicity, and were representative of the Nebraska student population. The vendor claims to include SPED and EL, but no demographic data were provided. Differential weighting was used to approximate population norms. Normative data were based on scores from students who took STAR Early Literacy, STAR Reading, or both, but no information was provided on the sample equivalence of these groups. Norms were reported for the Unified scale instead of separately for scores from STAR Early Literacy or STAR Reading. Further evidence is needed to substantiate the validity of the unified-scale norm scores, given that the two tests measure different skill sets and no winter data were collected.

Cut scores. Cut scores of the 40th, 25th, and 10th percentiles were set for STAR Early Literacy and STAR Reading based on an expert's recommendations. Normative data were used to establish benchmarks for fall and spring; however, winter was interpolated because no normative data were collected during the winter administration window. Classification indices or ROC analyses were not provided to support these proposed cut scores.

A STAR Early Literacy classification-accuracy study was conducted in 2021-22, using FastBridge Early Reading as the external criterion for kindergarten and 1st grade. FastBridge is now owned by the same vendor as STAR Early Literacy, but it was developed by a different vendor using its own distinct test-development procedures. Therefore, it is considered a valid external criterion. However, this study occurred before the 2022-23 norming update, and it is unclear whether the Unified scale was used at that time. Evidence of classification accuracy was considered particularly important for STAR Early Literacy and STAR Reading, given that interpolation was used to determine cut scores for the Winter administration. Vendor documentation stated that cut scores for STAR Early Literacy and FastBridge Early Reading were at the 20th percentile. However, this cut score does not align with the 40th-, 25th-, and 10th-percentile cut scores recommended by the expert and the state recommendation from the vendor (i.e., that any student who scores below the 40th PR on the STAR Early Literacy assessment take STAR CBM Reading for additional insights into their risk for reading difficulties). These cut scores also do not align with the documentation submitted by FastBridge Early Reading with STAR Early Literacy, which reported using the 11th and 17th PR for STAR Early Literacy. These studies were reportedly conducted in the same year, but the reported sample statistics and indices differed. Sensitivities reported in the STAR Early Literacy documentation for kindergarten were below the minimum for fall and winter administration, but low for spring. First-grade sensitivities were below the minimum for fall and spring administrations but met the minimum for winter administration. Although the winter sensitivities met the minimums, these data were collected before the normative update and the setting of winter cut scores via interpolation. NPV and PPV were not reported. All specificities met the minimum.

A second 2024 study reported classification accuracy for only the 1st-grade spring administration of STAR Early Literacy with California Wonders. The sample comprised a single school district in California, and no demographic information was provided. The 25th PR on STAR Early Literacy was used, although this does not align with the recommended 40th PR for screening for reading difficulties. Sensitivity and specificity were reported and exceeded minimum expectations. NPV and PPV were not reported

STAR Reading classification accuracy evidence for 2nd and 3rd grade was based on data collected between 2011 and 2023. Details on samples or criterion variables used in each study were not provided. Results were averaged across studies; however, the unified scale and normative update were introduced during this period, so the comparability of the studies remains unclear. The average sensitivity and specificity indices reported for each exceeded minimum expectations, but were not reported separately for each administration window. Grade 2 was based on a single study, and the year it was conducted is unclear. NCII documents studies conducted in 2012-13, 2017, and 2018; however, all of these occurred before the unified scale and norm update.

Decision accuracy and decision consistency studies conducted for STAR Early Literacy and STAR Reading were conducted in 2018-19, before renorming in 2022-24, and perhaps before the use of the unified scale. The sample used was not clearly documented. Results were reported for the 10th-, 25th-, and 40th-percentile cut scores; they were strong, but results were not reported for each administration window. In addition, CSEMs near cut scores are needed to confirm that the method used to assess decision accuracy meets the assumptions.

Reliability, Validity, and Fairness. The team discontinued its review of the technical adequacy of the STAR Suite measures (reliability, validity, and fairness) for the following reasons:

- The STAR measures did not feature evidence of content validity in the form of expert review or independent alignment review. Test developers indicated which standards the test addressed; this type of review is not independent and does not address alignment in the other direction (i.e., the percentage of grade-level standards the test covers). Early Literacy reports a large number of scores based on only 27 items, raising concerns about independence among the scores.
- Lack of documentation that the scale and scores of the unified scale are equivalent for the two measures, particularly since they measure different dimensions of content. There is insufficient evidence that a student would receive the same screening outcome regardless of which measure is administered.
- They do have recent norms, but students in the sample may have taken STAR Early Literacy, STAR Reading, or both, and there is no data on how many students were in each group, nor is there any information on whether the norms were similar if you compared students from those three groups.
- In addition, they did not collect normative data in the winter, instead using interpolation to set winter cut scores. Classification accuracy studies could have demonstrated that winter cut scores correctly identify students but were conducted before the 2022-2024 norming study and may have occurred before the unified scale was implemented.
- STAR Early Literacy provided a large number of scores (it was unclear how many; perhaps 10 or more) based on as few as 27 total items. Generating numerous scores from so few items raises concerns about the internal consistency and independence of the scores produced. Clarification is needed in this area, particularly if separate scores are intended to yield separate inferences (e.g., different interventions).

c. Usability of *STAR Early Literacy & STAR Reading*

Given the discontinuation of the STAR Suite review due to concerns about outdated norms and classification-accuracy studies, the criteria for usability were not revisited. Information on usability is available in the vendor’s response if needed.

Conclusion: Considerations and Recommendations

The information in the report is provided to assist NDE in deciding which screening measures to approve for the 2026-2027 school year. No screening measure met all expectations, and there are various considerations beyond the documentation and evidence reviewed for this report. The goal of the review was to evaluate each measure individually rather than to compare them. We hope this report will provide NDE with sufficient information to develop a short list of evidence-based, useful screening measures that align with Nebraska's reading priorities. To the extent that different NDE school districts might select different measures, additional evaluations comparing those selected screening measures may be warranted. Below, we have outlined recommendations NDE may want to ask each vendor to consider when making its final decision.

1. Indicate the exact scoring procedures used to make screening identifications at each grade level. Include a graphic, such as a flow chart, if necessary. If multiple tests are used, it should be clear how the scores are combined to make a single determination.
2. A timetable for the development of Nebraska-specific norms, internal consistency, reliability, and validity evidence if the measure were adopted. A measure that is adopted should eventually have enough participants from Nebraska to (1) compare Nebraska norms to national norms, (2) calculate reliability for all scores used and for the entire population and prominent subgroups, and (3) gather evidence of internal structure validity (e.g., factor analysis) for Nebraska.
3. If individualized programming is provided based on the results of the screener, evidence should include (1) how scores connect to individualized interventions, (2) technical quality of each of the scores used (e.g., coefficient alpha), and (3) independence of each score used from each other score used (e.g., a correlation matrix indicating scores that are not too highly correlated).
4. A plan for collecting accessibility, usability, utility, and satisfaction information from Nebraska districts and schools using each measure. Examples of sources of such data might include interviews, surveys, and focus groups.

As with large-scale proficiency assessments, maintenance of screener performance for NDE's purposes should be viewed as an ongoing process; thus, each measure's performance should be reviewed at least once per year across the domains of usability, reliability, validity, and fairness. Any changes made to approved measures, including modifications to the cut score, should be reviewed and addressed with requests for vendor evaluation(s) as needed. Although newly adopted measures may not have data on Nebraska student populations, after implementation for a year or more, they should be able to provide technical reports on performance specifically within the State. NDE is advised to develop a standardized process for obtaining and evaluating such information, as well as for communicating the results and implications back to vendors. Such a procedure will ensure the State's screening measures consistently perform at or above expectations and in the best interest of students, teachers, and schools.

Tables

Table 1. Screener administration format, time, and required materials

Screener	Format	Administration Time (minutes)	Student Materials
i-Ready Early Literacy	Group (Diagnostic)	25-35 (K & 1) 40-60 (2 &3)	Computer or tablet Headphones
	1-on-1 (Literacy Task(s))	1-5 per student	
Amplify DIBELS 8 mCLASS	1-on-1 (except Maze)	4-6 per student	Computer or tablet (Paper optional)
	Group (Maze only)	4	
FastBridge earlyReading	1-on-1	11 per student	Computer or tablet
FastBridge aReading	Group	10-15	Computer or tablet Headphones
FastBridge CBMreading	1-on-1	6-8 per student	Computer or tablet
Amira	Group	14-17 (K) 20-25 (1) 16-20 (2) 13-17 (3)	Computer or tablet Microphone
Acadience Learning	1-on-1 (except Maze)	8 per student	Computer or tablet (Paper optional)
	Group (Maze only)	3	
MAP Reading Fluency	Group	20-30	Computer or tablet Headphones Microphone

MAP Growth	Group	20-40 (K & 1) 40-50 (2 & 3)	Computer or tablet Headphones
STAR Early Literacy	1-on-1	10-15	Computer or tablet Headphones
STAR Reading	Group	20-30	Computer or tablet

Table 2. Screener licensing costs, training requirements, and training costs

Screener	Annual Licensing Cost Per Student	What license included in addition to the screener	Training	Training Time (hours)	Online PD costs	Onsite PD costs
i-Ready Early Literacy (i-Ready Diagnostic & Literacy Tasks)	$8.25	Report access Supplemental online PD Support Learning extensions	Required	3	Not mentioned	$2,400 (up to 30)
Amplify DIBELS 8 mCLASS	$9.00 single yr $7.00 multiple yrs	Report platform Instructional tools	Required	4-8	$750-$1,500 (up to 30) or $49 individual asynchronous	$3,200 (up to 30)
FastBridge Suite (earlyReading, aReading, CBMreading)	$7.25-$8.50	Report access Asynchronous online training Supports Resource bank How to webinars	Recommended	1-1.5 (Online) 6 (Onsite)	$275-$300	$2,500 (# participants not mentioned)
Amira	$5.00	Reporting dashboards Data management tools Support PD (Amira Academy)	Not Required	.75	No cost (Getting started optional)	N/A
Acadience Learning Online Reading	$7.45	Reporting dashboard	Required	Not reported	Online/live (Not reported)	Not reported
MAP Suite (Reading Fluency & Growth)	$13.00 ($20 for Coach)	Standard reporting Account management Support services	Required for MAP Growth	1-3	$630 - $1,260 (up to 30)	$2,100 (up to 35)
STAR Suite (Early Literacy & Reading)	$7.25-$8.50	Report access Online resources Support	Recommended	1-1.5	$275-$300	$2,500(?)

NOTE: All table information should be confirmed with the vendor.

Table 3 – Summary of screener ratings

		Appropriateness of the screener		Technical Adequacy of the screener	
Screener	**Grades**	**Rating**	**Partial (P) or Insufficient (I) evidence**[8]	**Rating**	**Partial (P) or Insufficient (I) evidence**
i-Ready Early Literacy Screener	K-3	Met expectations	N/A	Partially Met	Norms (P) Classification consistency (P) Fairness (P)
mCLASS DIBELS 8th Edition	K-3	Met with weaknesses	Administration/Scoring Model (P)	Partially Met	Norms (P) Classification accuracy (P) Classification consistency (I) Reliability (P) Validity (P) Fairness (P)
FastBridge earlyReading	K & 1	Met with weaknesses	Administration/Scoring Model (P)	Partially Met	Norms (P) Classification accuracy (P) Classification consistency (I) Validity (P) Fairness (I)
FastBridge aReading	2 & 3	Partially Met	Content coverage/alignment (P) Administration/Scoring model (I)	Partially Met	Norms (P) Classification accuracy (P) Classification consistency (I) Reliability (P) Validity (I) Fairness (I)
FastBridge CBMreading	1-3	Met with weaknesses	Content coverage/alignment (P) Administration/Scoring model (P)	Partially Met	Norms (P) Classification accuracy (P) Classification consistency (I) Validity (P)

Amira Reading Mastery Universal Screener	K-3	Partially Met	Administration/Scoring model (I)	Partially Met	Norms (P) Classification accuracy (P) Classification consistency (P) Reliability (P) Validity (P) Fairness (P)
Acadience Learning	K-3	Met with weaknesses	Content coverage/alignment (P)	More evidence needed**	Norms (I) Classification accuracy (I) Classification consistency (I)
MAP Reading Fluency	K & 1	More evidence needed	Content coverage/alignment (P) Administration/Scoring Model (I)	More evidence needed	Norms (I) Classification accuracy (I) Classification consistency (I) Reliability (P) Validity (I) Fairness (I)
MAP Growth	2 & 3	More evidence needed	Intended interpretation and Use (I) Content coverage/alignment (I) Administration/Scoring Model (I)	More evidence needed	Norms (I) Classification accuracy (I) Classification consistency (I) Reliability (I) Validity (I) Fairness (P)
STAR Early Literacy	K & 1	More evidence needed	Content coverage & alignment (P) Administration/Scoring model (I)	More evidence needed**	Norms (I) Classification accuracy (I) Classification consistency (P)
STAR Reading	2 & 3	More evidence needed	Content coverage/alignment (P) Administration/Scoring model (I)	More evidence needed**	Norms (I) Classification accuracy (I) Classification consistency (P)

**Reliability, validity, and fairness evidence for this measure was not reviewed due to insufficient evidence for content, norms, and/or cut scores.

Appendices

Appendix A. Key Reading Skill Area by grade

Note: Grade areas listed represent the grades that those key skills areas are measured within that assessment

Screener	Concepts of Print	Phonemic Awareness	Phonics	Letter Recognition	Word Recognition*	Comprehension	Vocabulary
Acadience	--	K,1	K-2	K-2	--	1-3	--
Amira	--	K-3	K-3	K-2	K-3	K-3	K-3
Amplify DIBELS 8th	--	K, 1	K-3	K-1	K-3	2, 3	--
FastBridge earlyReading	K	K, 1	K, 1	K-1	K-1	--	--
FastBridge CBM	--	--	--	--	1-3	--	--
FastBridge aReading	K,1	K,1	K-3	K-3	K-3	1-3	K-3
i-Ready Early Literacy Screener	--	K-3	K-3	K-3	K-3	K-3	K-3
MAP Growth K-2	K-2	K-2	K-2	K-2	K-2	--	--
MAP Growth 2-5	--	--	--	--		K-3	K-3
MAP Reading Fluency	K-3	K-3	K-3	K-3	K-3	K-3	K-3
Star Suite	K	K-1	K-2	K-2	K-3	K-3	K-3

*Word Recognition includes real-word reading and oral reading fluency accuracy subtests

Appendix B. Selected sub-skills by grade

Note: Grade areas listed represent the grades that list subskills are measured within that assessment

Concepts of Print Subskills Measured											
	Acadience	Amira	Amplify DIBELS 8th	FastBridge earlyReading	FastBridge CBM	FastBridge aReading	i-Ready Early Lit Screener	MAP Reading Fluency	MAP Growth K-2	MAP Growth 2-5	STAR Suite
Orienting/ Directionality	--	--	--	K	--	K,1	--	K-1	K-2	--	--
Differentiating units of text	--	--	--	K	--	K,1	--	K-1	--	--	K
Differentiating organizing features of text	--	--	--	K	--	K,1	--	K-1	--	--	--

NOTE: Differentiating units of text= differentiating letters, words, sentences, and paragraphs (units). Differentiating organizing features of text= differentiating spaces, indentations, and punctuation (organizing features of text).

Phonological Awareness Subskills Measured											
	Acadience	Amira	Amplify DIBELS 8th	FastBridge earlyReading	FastBridge CBM	FastBridge aReading	i-Ready Early Lit Screener	MAP Reading Fluency	MAP Growth K-2	MAP Growth 2-5	STAR Suite
Task											
Blending	--	K-3	--	K-1	--	K-1	K-3	K-3	K-2	--	K- 1
Segmenting	K-1	K-3	K-1	K-1	--	K-1	K-3	K-3	K-2	--	K-1
Manipulation	--	K-3	--	--	--	--	K-3	K-3	K-2	--	K-1
Unit											
Syllable	--	K-3	--	K-1	--	K-1	K-3	K-3	K-2	--	K-1
Onset Rime	K-1	K-3	--	K-1	--	K-1	K	K-3	K-2	--	K-1
Phoneme	K-1	K-3	K-1	K-1	--	K-1	K-3	K-3	K-2	--	K-1
Fluency											
Accuracy	K-1	K-3	K-1	K-1	--	K-1	K-3	K-3	K-2	--	K-1
Speed	K-1	K-3	K-1	--	--	--	K-3	K-3	--	--	K-1

Phonics Subskills Measured

	Acadience	Amira	Amplify DIBELS 8th	FastBridge earlyReading	FastBridge CBM	FastBridge aReading	i-Ready Early Lit Screener	MAP Reading Fluency	MAP Growth K-2	MAP Growth 2-5	STAR Suite
Skills											
Graph to Phoneme	K-2	K-3	K-3	K-1	--	K-3	K-3	K-3	K-2	--	K- 2
Unit											
Grapheme	--	K-3	--	K-1	--	K-3	K-1	K-3	K-2	--	K-2
Word	K-2	K-3	K-3	K-1	--	K-3	K-3	K-3	K-2	--	K-3
Fluency											
Accuracy	K-2	K-3	K-3	K-1	--	K-3	K-3	K-3	K-2	--	K-3
Speed	K-2	K-3	K-3	K-1	--	--	K-3	K-3	--	--	K-3

NOTE: Graph to phoneme = Grapheme to Phoneme.

Letter Identification and Word Reading Subskills Measured

	Acadience	Amira	Amplify DIBELS 8th	FastBridge earlyReading	FastBridge CBM	FastBridge aReading	i-Ready Early Lit Screener	MAP Reading Fluency	MAP Growth K-2	MAP Growth 2-5	STAR Suite
Skills											
Letter Naming	K-1	K-2	K-1	K-1	--	K-3	K-3	K-3	K-2	--	K- 1
Word Reading	--	K-3	K-3	K-1	--	K-3	K-3	K-3	K-2		K-3
Passage Reading	1-3	K-3	1-3	--	1-3	K-3	K-3	K-3	--	--	1-3
Fluency											
Accuracy	K-3	K-3	K-3	K-1	1-3	K-3	K-3	K-3	--	--	K-3
Speed	K-3	K-3	K-3	K-1	1-3	--	K-3	K-3	--	--	K-3

Selected Comprehension Subskills Measured											
	Acadience	Amira	Amplify DIBELS 8th	FastBridge earlyReading	FastBridge CBM	FastBridge aReading	i-Ready Early Lit Screener	MAP Reading Fluency	MAP Growth K-2	MAP Growth 2-5	STAR Suite
Skills											
Monitoring Understanding	3	K-3	2-3	--	--	1-3	K-3	K-3	--	K-3	K-3
Retelling	1-3	K-3	--	--	--	--	K-3	--	--	--	--
Answer Questions	--	K-3	--	--	--	1-3	K-3	K-3	--	K-3	--
Unit											
Reading	1-3	1-3	2-3	--	--	1-3	K-3	K-3	--	K-3	1-3
Listening	--	K-2	--	--	--	--	--	K-3	--	K-2	K-2
Fluency											
Accuracy	3	1-3	2-3	--	--	1-3	--	K-3	--	--	K-3
Speed	3	1-3	2-3	--	--	--	--	--	--	--	1-3

Selected Vocabulary Skills Measured											
	Acadience	Amira	Amplify DIBELS 8th	FastBridge earlyReading	FastBridge CBM	FastBridge aReading	i-Ready Early Lit Screener	MAP Reading Fluency	MAP Growth K-2	MAP Growth 2-5	STAR Suite
Skills											
Expressive	--	--	--	--	--	--	--	--	--	--	K-2
Receptive	--	K-3	--	--	--	K-3	K-3	K-3	--	--	K-3

Appendix C. Administration and Scoring Approaches in the Reading Screener Assessments

The reviews conducted in support of the Nebraska Reading Improvement Act (NRIA) for the Nebraska Department of Education (NDE) are comprehensive considerations of the content, norms, classifications, psychometrics, reporting, cost and other aspects of reading screening assessments used by Nebraska school districts. This section focuses on the administration and scoring approaches of the reading screening assessments. It supplements the other sections by addressing questions about how the intended content domains of the assessments are reflected in their administration and scoring. First, major types of administration and scoring approaches are summarized, along with the questions they prompt. Then, the administration and scoring practices implemented in the considered assessments are described. Tables C1-C3 summarize major aspects of the administration and scoring approaches of each reading screening assessment, along with some outstanding questions prompted about the assessment procedures.

General Descriptions of Administration and Scoring Approaches

Two major approaches to test administration in large-scale assessment are non-adaptive and adaptive approaches. Non-adaptive assessments administer test forms assembled to meet uniformly defined content and difficulty targets. Computer-adaptive tests (CATs) adapt to test takers' performance by selecting subsequent test items that are "optimal" with respect to difficulty, and that may vary in content. As summarized in the review of reading screeners by Truckenmiller et al. (2025) for Connecticut, "curriculum-based measurement" (CBM) approaches contrast with CAT in how well CBM assessments cover different content and in the scoring approaches they use (e.g., Item Response Theory, etc.). The scoring approaches for CATs can be especially complex.

Non-adaptive administration approaches tend to be most closely aligned with the intended content of the assessment. That is, fixed, non-adaptive test forms can be developed that directly meet content targets for all test takers. The test items from these content areas can be scored in ways that preserve content assembly targets (i.e., in number correct scoring, the contributions of item scores to reported scores tend to be direct reflections of the forms' assembly and the number of items). Some questions to consider with non-adaptive administration and scoring approaches are how form differences are addressed in scores and, specifically, whether score equating procedures are used to ensure that scores do not reflect easier or more difficult items on some forms vs. others. These questions are relevant for directly designed alternate test forms and for assessments that use random selection methods to vary the orders, presentations, and selections of words and/or passages.

Adaptive testing through CATs is based on the notion that when a test is tailored to each student's ability (i.e., is harder for more able students and easier for less able students), the test and scores will ultimately be more efficient and supportive of high reliability with relatively short testing time (Lord, 1980). The adaptations might be implemented after the administration and scoring of individual items (Item-level), or after discrete sets of items (Multi-Stage Adaptive). Achieving increased efficiency almost always involves the use of item-level Item Response Theory (IRT) measurement and scoring models. The IRT models used in practice are almost always unidimensional and do not account for different subcontent domains. The scoring approaches are typically pre-equating approaches that assume items' IRT scoring parameters hold even as those test items are administered in different orders and positions for different test takers. The assumptions of these practices have been described as strong and as raising questions about the ultimate accuracy of test takers' scores (Kolen & Brennan, 2014; Moses, 2025; Moses & Dorans, 2025). These suggest that questions can be raised about CAT implementations, specifically about how content presentations are managed, and also about what, if any, procedures are used to ensure that the IRT model used to support item scoring is accurate.

The reading screening assessments considered in these reviews involve unique non-adaptive and adaptive administration and scoring approaches. The assessments primarily characterized as non-adaptive are first presented and summarized in Table C1, including DIBELS, Acadience, and the FastBridge CBM and earlyReading assessments. Next, the adaptive CAT assessments are described and summarized (Table C2), including FastBridge aReading, Star Reading, STAR Early Literacy, and Amira ISIP. Finally, assessments that combine non-adaptive and CAT implementations are reviewed and summarized (Table C3), including MAP Fluency, MAP Growth, and i-Ready. A final section summarizes some implications of these reviews, focusing on the extent to which test content is maintained in the assessment and scoring approaches, and the unanswered questions they raise.

DIBELS 8th Edition

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) consists of a set of curriculum-based measures for assessing reading skills (University of Oregon, 2021). For the 8th edition, DIBELS 8th, there are six subtests to assess component skills, including Letter Naming Fluency (LNF), Phonemic Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), Word Reading Fluency (WRF), Oral Reading Fluency (ORF), and Maze. Maze is a group-administered measure intended for grades 2 and higher, ORF is intended for grades 1 and higher, and the other measures are standardized, individually administered measures intended for grades K and higher. For several of these measures, 20 alternate progress monitoring forms were developed. Truckenmiller et al. (2025) distinguish DIBELS from CAT assessments, although the submitted RFI form indicated that DIBELS was adaptive (possibly its Maze measure).

The University of Oregon (2021) report describes screening and risk classifications in terms of unequated raw scores for LNF, NWF, WRF, ORF, and Composites obtained as weighted sums of raw scores. Cut scores for these measures by grade and time of year are provided in Appendix E. Interestingly, the report describes the screenings as based on raw scores, even while acknowledging their weaknesses in interpretation and the need for equating for form effects (pp. 84-85). The use of raw scores for screening recommendations, despite acknowledged weaknesses, raises the most important questions about the interpretation of DIBELS 8 scores and the extent to which they reflect the specific progress-monitoring form.

Acadience Reading

Acadience Reading is a set of measures used to assess early literacy and reading skills for students from grades K-6 (Good, Kaminski, Dewey, Wallin, Powell-Smith & Latimer, 2013-2019). The measures include First Sound Fluency, Letter Naming Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, Oral Reading Fluency, Retell, Maze, and Word Use Fluency. The administration of these measures is not adaptive, but involves some variation in the presentation of assessment stimuli, including randomized orders and word selection. Some "equating" procedures are described for the different passages that could be administered in Oral Reading Fluency. However, descriptions of these "equating" procedures covered readability analyses, field testing, and passage selection, but not actual evaluations of scores or the need for score adjustments from passages found to exhibit higher or lower difficulty. Composite scores are obtained by summing a combination of measures corresponding to particular grades and times of year (Table 2.6, Good et al., 2013-2019). The approaches and prompted questions for Acadience Reading are summarized in Table C1, focusing on the extent to which raw scores from the measures and the Composites are comparable across different passages and across different orders for word presentation within each measure.

FastBridge

The FastBridge Assessment Suite includes a large selection of brief assessments designed for use as universal screeners, progress monitoring, or supplemental diagnostics (Renaissance, 2025). The assessments recommended for screening include the earlyReading Composite, CBMreading, and aReading. These assessments are recommended for universal screening in specific grades, where the grades differ by documentation (i.e., Table 1 of Renaissance Learning, 2022 vs. Table 1 of Renaissance Learning, 2024).

- FastBridge earlyReading assesses unified and component skills for grades K-1 in Concepts of Print, Phonemic Awareness, Phonics, and Decoding. Subtests for these domains are administered individually to students. Composite scores are produced by weighting and summing students' raw test scores using subtest score weights to maximize predictive validity and provide comparability of scores across seasons. CBMreading appears to be included in the composite score calculations for Winter and Spring seasons of grade 1 (Table 13, Renaissance Learning, 2022). For earlyReading, alternate forms were developed for progress monitoring for each skill. Score equating for these alternate forms was not described, and they are included as questions in Table C1.
- The FastBridge CBMreading is a curriculum-based measure of oral reading fluency. It uses a standard administration protocol where a teacher or other qualified educator administers reading passages to individual students (three for screening). Three scores are provided: the number of words read correctly, the number of errors, and accuracy, computed as (total words read correctly – errors)/total words read correctly. Although score reporting appears to be based on raw scores, the extent to which different "forms" or passages might be administered to different students and warrant equating procedures for these alternate passages is unclear (though alternate-forms reliabilities for students' scores across their three passages are provided).
- The FastBridge aReading assessment is a CAT that assesses particular skills and content areas for different grades, ranging from K-12 (Table 3, Renaissance, 2025). Details about the CAT implementation are not provided, particularly of the IRT models used for the assessment items and scores (though one mention of Rasch difficulties is provided). Important, but unprovided details of the CAT implementation would cover how the IRT parameters of the items are linked to a common scale, and periodically evaluated for potential drift and exposure effects. The basis of the adaptation is described as reflecting test length and precision criteria, as well as the difficulty ranges of the content areas for each grade. Statements such as those quoted below prompt some questions about aReading and how its content is administered to students of different abilities in each grade (Table C2).
    - "*The computer adaptive testing (CAT) algorithm used for administering FastBridge's aMath and aReading assessments was designed to maximize efficiency without compromising the reliability of overall ability estimation. For that purpose, the algorithm selects items that are best aligned to the student's ability estimate based on the student's performance on the items already completed. For example, if the student's ability estimate after*

> *completing 10 aMath items was 210, the algorithm searched the item bank for the 10 items best matched to that ability and randomly selected one for administration. That process continues until the student completes up to 30 items or has attained a sufficiently precise score estimate (SEM <=.20). In addition, two item constrictions are employed for item selection on the adaptive assessment. First, an item cannot be readministered to a student if it was administered during the last screening period. Second, item selection for aMath only is constrained to items that assess standards no higher than two grades above the student's enrolled grade level."* (p. 581 of Renaissance, 2025).

Star Reading and Early Literacy

The Star Suite includes three assessments: Reading, Early Literacy, and CBM (curriculum-based measurement, Renaissance, 2023a, 2023b, 2025). Star Early Literacy is designed for pre-K through 3rd-grade students who are beginning readers, typically in grades K-1, who do not yet read independently or need early literacy skills assessed in three Blueprint Domains: Word Knowledge and Skills, Comprehension Strategies and Constructing Meaning, and Numbers and Operations. Star Reading assesses achievement for students typically in grades 2-3 on reading skills on five Blueprint Domains: Word Knowledge and Skills, Comprehension Strategies and Constructing Meaning, Analyzing Literacy Text, Understanding Author's Craft, and Analyzing Argument and Evaluating Text. This summary focuses on the Reading and Early Literacy assessments, as Star CBM is recommended for assessments beyond initial screening.

Star Early Literacy and Reading are item-adaptive CAT assessments, where the adaptations are based on selecting items with difficulty levels that reflect students' performance on previously administered items. Both assessment models score their test items using the Rasch IRT model, in which items are modeled only in terms of their difficulty, and where individual item scores make equal contributions to students' ability estimates and scores. Using IRT modeling procedures, students' scores on the Early Literacy and Reading assessments are produced to reflect a Unified Scale ranging from 0-1400, with Early Literacy scores ranging from 200-1100, and Reading scores ranging from 600-1400. The Unified Scale implies that the Early Literacy and Reading scores mean the same thing, apparently leading to the reporting of scores and norms for a combined group of Early Literacy and Reading students, with screening decisions for both assessments based on norms/percentile ranks below 40 in the combined group.

The CAT administration, scoring, and score reporting practices for Star Early Literacy and Reading raise questions about how test content is addressed. In general, assessment practices that support scores that mean the same thing and can be reported on the same score (i.e., Unified Scales) require that those assessments reflect the same content. This is clearly not the case for a Star Early Literacy assessment that assesses three Blueprint Domains and a Star Reading assessment that assesses five Blueprint Domains (despite some overlap in these domains). Questions about the comparability of scores could be addressed in disaggregated norms reports, which might allow comparisons of the distributions of Early Literacy and Reading scores for students in the same grades who are similar in other characteristics. The item-level adaptations for the Early Literacy and Reading are described as reflecting only item difficulty, not content, raising additional questions about whether content domains with harder vs. easier items would be overrepresented among higher vs. lower ability students. Addressing these questions would require more information about the assessment item pools, specifically the item difficulty ranges in each domain. These issues raise the possibility that reading screening decisions reflect different content domains for different students. Finally, the CAT implementations in the Star assessments warrant evaluations of IRT parameter stability, and reviews of manuals describe IRT parameter linking procedures using common “anchor” items, large item banks, and random item selection rules that reduce overexposure, but no description of item drift evaluations.

Amira ISIP

Aimira ISIP (Istation’s Indicators of Progress, 2025-2026) is an assessment system used by Amira to measure early literacy development, serve as a universal screener, provide benchmarking, and support progress monitoring. It provides teachers with data to monitor student progress and differentiate instruction, while also informing students of their strengths and weaknesses through personalized activities. The constructs measured include Phonological and Phonemic Awareness, Alphabetic Knowledge, Phonics/Decoding, Oral Reading Fluency, Vocabulary, Spelling/Encoding, Reading Comprehension, Oral Language, Rapid Automatized Naming, and Visual Attention. Student screening for reading difficulties is based on the Amira ISIP Reading Mastery ARM composite score.

The Amira assessment uses an adaptive approach described as unlike either CAT or CBM. Details on the adaptive algorithm are unclear, though page 60 of the Technical Manual indicates that calibrated items in the item pool are “*categorized by their difficulty and discrimination. These calibrated items form the foundation for Amira ISIP's CAT algorithm, which dynamically adjusts test difficulty in real-time based on a student's performance to provide an optimal challenge and accurately measure their abilities*.” This statement indicates that item difficulty and discrimination characteristics are part of the basis of Amira's adaptive testing. Whether additional aspects of the algorithm are introduced to control for content exposure, and especially to represent all of its intended constructs, is unclear. However, the reporting of subscores on specific skill domains suggests that content representativeness might be preserved in some way. The Technical Manual describes linking transformation procedures for establishing and maintaining consistent IRT-based vertical scales across grades using common anchor items and Stocking-Lord transformation procedures. Multiple calibration studies were conducted to assess the stability of linking functions.

One unique aspect of Amira ISIP is its use of Expected A Posteriori (EAP) scoring based on the 2-parameter logistic IRT model. An implication is that items and content domains with higher IRT discrimination values will make greater contributions to ability estimates and reported scores than those with lower discrimination values. Another implication is that students' scores are based not only on their pattern of correct/incorrect responses to all the items they take, but also on their grade, where grades K-1 are an established Base Scale, and higher grades are weighted to different (presumably higher) means. This feature implies that scores partially reflect students' grades. Consider a somewhat unrealistic situation where students from grades K, 1, and 2 were administered the same test and items and performed the same way in terms of their correct and incorrect item responses. To the extent that grade 2 students and scales have a higher prior mean than grades K and 1, grade 2 students would receive higher scores than grades K-1 students for the same test performance. This oversimplified example illustrates that IRT EAP scores are known to reflect not only test performance, but also the group of the test takers (Kolen, 2006; Moses, 2025). In addition to overall ability estimates, Amira ISIP computes subscores for specific skill domains, with the overall grade-based ability estimate serving as the prior mean for the subscores. These features indicate that although Amira ISIP is described as having a vertical scale, comparability across grades is questionable. Additional comparability questions can be raised about CAT adaptations that may involve only item difficulty and discrimination, not test content, and the potential for differentially weighted items and item content for higher vs. lower ability students who take items with different IRT discrimination values. Some questions about these issues are listed in Table C2.

MAP Reading Fluency

MAP Fluency is an adaptive online assessment that supports students on their path to reading comprehension by assessing and improving both oral reading fluency and foundational reading skills (NWEA, 2024). It is designed for students who do not yet read with solid fluency and understanding, and adapts to accommodate pre-readers, early readers, and independent readers in pre-K through Grade 5, helping all students read with comprehension. MAP Fluency is part of a Suite that also includes MAP Growth, an assessment designed for higher grades.

The MAP Fluency assessment is described as consisting of six test forms that a teacher can select from. In the Adaptive Oral Reading form, students are assessed on Sentence Reading Fluency and routed to either Oral Reading Fluency when they score above the Sentence Reading Fluency threshold, or to Foundational Skills when they score below it. The Foundational Skills measure is described as multi-stage adaptive (p. 62), clearly in terms of the initial Adaptive Oral Reading and Sentence Reading Fluency assessments, but possibly also in terms of the three Foundational Skills domains, Phonological Awareness, Phonics & Word Recognition, and Language Comprehension (Table 2.1, NWEA, 2024). Sentence Reading Fluency is described as a measure of the Phonics & Word Recognition domain that does not contribute to that domain score (Table 5.1). The three Foundational Skills domains are modeled as separate Rasch IRT scales and maintained through calibration procedures and IRT parameter drift evaluations. The scale of Sentence Reading Fluency is unclear, specifically whether it is based on the Phonics & Word Recognition IRT scales. The Dyslexia Screener is based on a multivariate predictive model involving Sentence Reading Fluency and the three Foundational Skills measures. In addition, a Rapid Automatized Naming (RAN) measure supplements the screening.

The CAT administration and scoring implementation for MAP Fluency may reflect some adaptive and non-adaptive aspects. This implies that the MAP Fluency setup potentially offers greater control over content exposure for different test takers than other item-level adaptive approaches summarized in Table C2. The separate modeling of the Foundational Skills domains, the use of multi-stage adaptive testing rather than item-level CATs, and the incorporation of separate domain scores into a multivariate predictive model for screening may result in more consistent administration of test content. However, content differences across the sub-measures of the Foundational Skills domains do occur among students in the same grade (Appendix B), possibly due to MST test assembly and administration features that are not fully described. Details are needed to understand better the MST setup for MAP Fluency, including whether adaptations within the three assessed Foundational Skills domains result in different content for different students. Questions about these issues are listed in Table C3.

MAP Growth Reading

MAP Growth assessments are interim adaptive tests that measure a student's academic achievement and growth in Reading, Language Usage, Mathematics, and Science (NWEA, 2019). The intended uses of MAP Growth scores are to monitor student achievement and growth over time, from kindergarten to high school, to plan instruction, to compare students within normed groups, to make universal screening and placement decisions, to predict student performance on external measures of academic achievement, to evaluate programs and conduct school improvement planning, to summarize scores for district- or school-level resource allocation, and to combine RIT (Rasch Unit) scores with other information to make educational decisions. For Reading, MAP Growth includes K-2 assessments for use as screening tests and skills checklists, and 2-12 assessments that are adaptive Growth and Screening tests. As summarized in the main report, the validity and cut-score evidence for MAP Growth are primarily based on students' performance on state assessments administered in grades 3 and up.

The MAP Growth tests are item-level adaptive, such that students who answer items correctly (or incorrectly) receive more difficult (or easier) items. The adaptations are based not only on item difficulty, but also on a content-balancing procedure that selects items from the most underrepresented content area. Test scoring is implemented in Rasch Unit Scales, where student abilities are estimated from their responses to test items modeled with Rasch IRT models, and these abilities are converted to RIT scale scores ranging from 100 to 350. Classifications are made by linking MAP results to state assessments and performance levels for students who took those tests in the Spring of a school year (with Fall and Winter classifications established using MAP Growth tables). IRT modeling for MAP Growth is similar to that for MAP Fluency (NWEA, 2024) in its use of the Rasch IRT model and in the reviews for potential item parameter drift. The evaluations of the adaptations described in the technical manuals for the two MAP assessments differ somewhat. MAP Growth describes content-balancing procedures as part of the adaptive implementation. However, it does not present the resulting numbers of students exposed to different sub-content areas, and MAP Fluency describes separate models for its three Foundational Skills domains and reports the numbers of students exposed to different sub-domains of the Foundational Skills (Appendix B, NWEA, 2024).

i-Ready Diagnostic and Literacy Task

The i-Ready Diagnostic (Curriculum Associates, 2025a) is an adaptive assessment designed to evaluate students' proficiency in the College- and Career Readiness Standards (CCRS) for reading and mathematics in kindergarten through Grade 12 (K–12). The Early Literacy and Dyslexia Risk Screener (Curriculum Associates, 2025b) uses a two-step process of first administering the Early Literacy Screener with the Diagnostic Overall assessment and a grade- and time-specific benchmark fluency Literacy Task, and then administering the Dyslexia Risk Screener with a rapid automatized naming task and a pseudoword decoding task. These assessments can be administered up to three times a year (fall, winter, spring).

Each assessment used in the i-Ready screenings follows a specific administration and scoring approach.

- The i-Ready Diagnostic is an item-level CAT that uses an item selection algorithm that maximizes measurement precision while also ensuring that students are assessed on appropriate content. From Curriculum Associates (2025a), the CAT algorithm uses a b-matching algorithm that selects students' test items as subsets based on each student's grade level, the content domain, grade-level caps and cap exceptions, whether the test is a diagnostic or growth monitoring assessment, and which items the student has been previously administered. Item content and content domain coverage are determined based on a test flow that specifies the test's content ordering (Reading or Math) and the specific grades (K and 1, 2, and 3-8). Testing ends when the student has completed the predetermined number of required operational items. Item modeling and scoring is based on the Rasch model. The Technical Manual describes procedures for calibrations, evaluations of parameter stability for common anchor items, and other evaluations to assess potential drift and scale stability (Curriculum Associates, 2025a).
- The i-Ready Literacy Tasks include Letter Naming Fluency for grade K, Word Recognition for grade 1 (fall), and Passage Reading Fluency for grades 1 (winter and spring), 2 and 3. Letter Naming Fluency assesses the skill of quickly and accurately naming letter names aloud in a timed task that produces raw scores of the number of letters correctly named in one minute. Word Recognition Fluency measures students' automatic word recognition for grade appropriate, high-frequency words, and reports scores as the number of correct responses. Passage Reading Fluency presents a grade-appropriate literary passage and an informational passage. Students read these aloud, and the administering teacher marks which words the student reads incorrectly.

Early Literacy screenings are based on a composite score that converts the Diagnostic overall scale score and the fluency measure into standard deviation units, and weights and sums them. From page 43 of Curriculum Associates (2025b), both components are equally weighted in the composite score. This scoring results in compensatory screenings that allow for higher performance on the Diagnostic or Literacy Task components to compensate for lower performance on the Literacy Task or Diagnostic component.

The use of adaptive and non-adaptive assessments for the i-Ready Early Literacy and Dyslexia Risk screenings and the added test flow controls for content presentations in the Diagnostic CAT imply that the i-Ready screening assessments are partly adaptive and partly non-adaptive. The i-Ready implementations appear to yield particularly strong control over the content administered to different test takers. Questions might consider the non-adaptive assessments included in the Early Literacy and Dyslexia Risk screenings, as these are based on unequated raw scores. Questions about i-Ready and the extent to which its screening results reflect variation in "forms," or alternate presentations of the letters, words, pseudo words, and stimuli in the Fluency Tasks, RAN, and PWD-F tasks are listed in Table C3.

Implications

This section was produced based on the view that the administration and scoring approaches used in the reading screening assessments are important aspects to be considered in their own right when evaluating these assessments. Similar to Truckenmiller et al. (2025), the reviews in this section show differences between non-adaptive assessments (sometimes referred to as CBM assessments) and CATs in terms of content coverage and scoring approaches. The reading screening assessments varied in administration and scoring, with the fully item-level adaptive assessments most different in how they manage content coverage. That is, for assessments with item-level CAT administration procedures (Table C2), selections of test takers' test items based only on "optimal" difficulty and possibly discrimination parameters from unidimensional IRT models make it very difficult for external evaluators to determine how content is managed for higher vs. lower ability test takers. There would seem to be greater potential for reading screening decisions to be based on different content for different test takers with most item-level CAT implementations.

Reading screening assessments administered using adaptive and non-adaptive procedures appear to yield clearer, more direct control over examinee content than other item-level CAT implementations. This suggests that the i-Ready Diagnostic and, possibly, MAP Fluency are assessments with greater content control (Table C3). The i-Ready Diagnostic controls content through implementing its CAT in a test flow, where item selections are based not only on items' IRT parameters, but also on content ordering. MAP Fluency is a stage-based, multi-stage adaptive implementation that presumably results in greater content control. However, details about this were more difficult to determine, especially whether and how multi-stage procedures were implemented across the three Foundational Skills domains and whether these may or may not have minimized the content differences described (e.g., Appendix B, NWEA, 2024).

The other reading screening assessments that used non-adaptive approaches also appear to yield clearer content coverage for test takers. These administration approaches include the development and administration of alternate forms, such as for performance monitoring, randomly varied presentations of letters, words, word orderings, passages, objects, or other stimuli in their tasks (i.e., the Table C1 reading screening assessments and the non-Diagnostic assessments used with i-Ready). Some challenges with these non-adaptive assessments are that they raise questions about the extent to which score variation across alternate forms, item orders, etc., influences screening decisions. Score equating procedures could be useful for addressing these (as was sometimes acknowledged; University of Oregon, 2021).

References

Amira Learning. (2025-2026). Amira ISIP Assess Technical Guide.

Curriculum Associates. (2025a). i-Ready diagnostic and growth monitoring assessments.

Curriculum Associates. (2025b). i-Ready Early Literacy and Dyslexia Risk Screener.

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. (2013-2019). Acadience® Reading K-6 Technical Manual.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 155–186). Praeger.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking (3rd ed.). Springer Publishing.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Erlbaum.

Moses, T. (2025). Scaling, equating, and linking. In L. L. Cook & M. J. Pitoniak (Eds.), Educational measurement (5th ed., pp. 735–820).

Moses, T., & Dorans, N. J. (2025). Test score linking and equating. Reference Module in Social Sciences. Elsevier. https://doi.org/10.1016/B978-0-443-26629-4.00064-2.

NWEA (2024). English MAP Reading Fluency Technical Report. Portland, OR: Author.

NWEA. (2019). MAP® Growth™ technical report. Portland, OR: Author.

Renaissance Learning. (2022). FastBridge Assessments Content Description & Use Guidelines.

Renaissance. (2023a). Star Early Literacy Appendices.

Renaissance Learning. (2023b). Star CBM Reading Appendices.

Renaissance Learning. (2024). Using Renaissance FastBridge Assessments for NebraskaREADS Initiative.

Renaissance Learning. (2025). Star Reading Appendices.

Renaissance. (2025). FastBridge Appendices.

Truckenmiller, A., Coyne, M., Valentine, IK., Moura, P. & Sarmiento, C. (2025). Independent researcher review of commercial reading screening assessment suites May 2025. Prepared for Connecticut State Department of Education, Performance Office.

University of Oregon (2021). DIBELS® 8th Edition: Administration and Scoring Guide.

Table C1. Summary of the Administration and Scoring Approaches of the Non-Adaptive Reading Screening Assessments.

Assessment	Screening Score(s)	How Scores are Produced	Administration Approach	How Content is Managed in the Administration Approach	Prompted Questions
DIBELS 8th Edition	Raw scores of Subtests and Weighted and Summed Subtest Scores that form Composite Scores.	Unequated raw scores	One from a set of 20 progress monitoring forms	Relies on content management procedures when assembling the progress monitoring forms	• How much variation is there in the scores of different progress monitoring forms? • Is the score variation from different progress monitoring forms large enough to warrant score equating?
Acadience Reading	Reading Composite Score	Composites are summed from other content scores, which are sometimes weighted (Appendix B of the Assessment Manual)	Forms with random presentations of letters, words, and passages for Oral Reading Fluency (ORF)	Content is presumably managed through a consistent administration of measures appropriate for students of particular grades	• How comparable are the scores when different words, orderings, and passages are administered? • Is equating warranted for scores obtained from different presentations of words, orderings, and/or passages?
FastBridge earlyReading	earlyReading- raw subtest scores reported separately and also weighted and summed into a Composite	Presumably unequated raw scores	Progress monitoring forms are individually administered	earlyReading- through subtests;	• Are equating procedures used with the scores from the progress monitoring forms?
FastBridge CBMreading	Scores on students' passages (three passage scores for screening)	Accuracy scores are computed as (total words read correctly – errors)/total words read correctly	One-on-one administration by a teacher or qualified educator	Unclear how passages are selected for administration to students	• To what extent are equating procedures needed for the scores of alternate passages?

Table C2. Summary of the Administration and Scoring Approaches of the Adaptive Reading Screening Assessments.

Assessment	Screening Score(s)	How Scores are Produced	Administration Approach	How Content is Managed in the Administration Approach	Prompted Questions
FastBridge aReading	aReading	Presumably, Rasch scale scores	Item-Level Adaptive	Unclear how content is managed in the CAT administration, as test adaptations appear to be based on item difficulty and not on test content	• How different is the content for forms produced by the CAT algorithm for the screening of different students, especially higher vs. lower ability students? • How are the IRT parameters for aReading estimated, linked to a common scale, and evaluated to ensure negligible exposure and drift effects?
Star: Reading and Early Literacy	Percentile Rank for percentage of students equal to or below 40, presumably based on Unified Scale Scores for both assessments.	Unified Scale Scores are linear conversions of ability estimates from the Rasch IRT Model.	Item-Level Adaptive	Content management is not described, as CAT is based only on item difficulty.	• To what extent are Unified Scale Scores comparable, given that Early Literacy and Reading reflect different content domains? • Does a cut score based on the 40th percentile apply the same way to both Star assessments? • If content is not managed in the CATs, are different students screened based on different content? • How are the IRT parameters evaluated to ensure negligible exposure and drift effects?
Amira ISIP	ARM composite score	2PL EAP scoring, with prior mean abilities for grades K-1, 2, and 3	Item-level adaptive	Unclear- not directly stated	• How are content and subscore domains preserved in the CAT implementation, specifically for higher vs. lower ability students within a given grade? • Given EAP scores are based on the 2PL model, do items from some content domains tend to make greater contributions to scores than others?

Table C3. Summary of the Administration and Scoring Approaches of the Reading Screening Assessments with Adaptive and Non-Adaptive Characteristics.

Assessment	Screening Score(s)	How Scores are Produced	Administration Approach	How Content is Managed in the Administration Approach	Prompted Questions
MAP Fluency	The three Foundational Skills domain scores and Sentence Reading Fluency are used in a multivariate dyslexia screening set. RAN scores also supplement this model.	Separate IRT Rasch models are used for the three Foundational Skills domains.	Multi-stage adaptive (MST) in terms of the Sentence Reading Fluency Routing, and possibly within the Foundational Skills assessment	Unclear due to a lack of details on the MST design and implementation	• Does the multi-stage testing implementation involve only a routing based on Sentence Reading Fluency, or are test adaptations also occurring within the Foundational Skills domains? • Does multi-stage adaptation within Foundational Skills domains account for the within-grade differences in sample sizes for the sub-measure differences in Appendix B?
MAP Growth Reading	Scaled Rasch Unit (RIT) score (ability estimate)	Rasch model proficiencies are linearly scaled to a 100-350 score range	Item-adaptive based partly on item difficulty and student performance, and also on a content-balancing procedure	The content balancing procedure is used to make adaptive item selections that reflect underrepresented content areas	• How well does the content-balancing work to ensure that all students are exposed to sub-domains in the Reading Blueprint?
i-Ready Early Literacy Screener	An Early Literacy Composite is a weighted sum of the Diagnostic overall scale score and a grade-appropriate Fluency Task.	Diagnostic scores are scaled Rasch model proficiencies. The Fluency Task produces raw scores.	Diagnostic uses an adaptive approach that selects items by b-matching, based on student grade, domain, item exposure, and other characteristics, and item difficulty	Diagnostic item selections are partly based on test flows that control and order item content within the assessment	• To what extent does variation in the "forms" or alternate presentations of the Fluency Task affect Composite scores and screening decisions?