Technical Report 2604-IAR

Interim Assessments in Reading:

Application of easyCBM®

Gerald Tindal, PhD

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Note: This technical report was supported in part by Riverside Insight, the exclusive distributor of easyCBM®. The report does not reflect any endorsement by any of these organizations.

Copyright© 2026. Behavioral Research and Teaching. All rights reserved. This publication or parts thereof, may not be used or reproduced in any manner without written permission.

APA Reference: Tindal, G. (2026). *Interim Assessments in Reading: Application of easyCBM® (Technical Report 2604-IAR)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

The University of Oregon is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation. This document is available in alternative formats upon request.

Center for Assessment Gateways with Criteria and Indicators for Interim Assessment[1]

The review tool organizes evidence about interim assessments into three sequential "gateways" which represent a major aspect of technical and instructional quality that must be addressed before an assessment can be judged appropriate for use. The same gateways-criteria-indicators are provided for both English Language Arts (ELA) (Center for Assessment & EdReports.org, 2023a) and Mathematics (Center for Assessment & EdReports.org, 2023b). The intent is identical across content areas: (a) confirm strong alignment to college- and career-ready standards and equitable access, (b) confirm that scores and claims have sufficient technical quality for their intended interpretations, and (c) confirm that score reports and guidance enable correct, responsible use by educators and other stakeholders. Within each gateway are criteria (broad requirements) and indicators (specific, observable evidence statements). Indicators are rated, and criteria carry point totals or are "claim-dependent," meaning points apply only when the publisher claims to provide a particular type of result (e.g., predictions, sub scores, growth). Together, the gateways function like an argument: The assessment must first measure the right constructs (Gateway 1), then produce defensible results (Gateway 2), and finally communicate those results in ways that support sound decisions (Gateway 3)

Gateway 1: Alignment, Fairness, and Accessibility

Gateway 1 evaluates whether the interim assessment is built to measure the intended standards and whether the assessment experience is fair and accessible. In both math and ELA, this gateway looks for a clearly articulated test design that targets the breadth and depth of the standards, emphasizes the most important content, and reduces construct-irrelevant barriers so that all students have an equitable opportunity to demonstrate what they know.

Gateway 2: Technical Quality

Gateway 2 evaluates whether the assessment's results are technically strong enough to support the interpretations and uses claimed by the vendor. This gateway is structured around four possible types of results: overall achievement, predictions of future performance on an external criterion, sub scores for strengths and needs, and progress/growth over time. The framework is intentionally "claim-dependent": If a program does not claim to provide predictions, sub scores, or growth measures, those criteria are treated as not claimed and are not scored; if the program does claim them, it must provide evidence that the corresponding results are designed appropriately, reliable/precise, valid for the intended interpretations, and suitable for intended uses.

Gateway 3: Score Reports and Interpretive Guides

Gateway 3 evaluates whether score reports and supporting materials help users interpret results accurately and use them responsibly. Even technically strong scores can be misused if reports are unclear, omit uncertainty information, or lack guidance connecting results to action. Accordingly, this gateway checks for audience-appropriate design, communication of error, and actionable guidance. As in Gateway 2, criteria related to predictions, sub scores, and progress are claim-dependent and scored only when those results are provided.

Conclusion: An important consideration is that Interim Assessments have become relevant with the shift from summative testing programs initially used as part of *No Child Left Behind (NCLB)* to a broader conception of accountability programs that, nevertheless, follow the guidelines used with peer review. Importantly, because of their use in large-scale assessment programs, the same standards described below, apply to them.

Note: At the end of this document, the sources used in this Technical Report are described. The first includes a reference to the original technical reports providing the evidence for the claims. The second source includes the specific criteria and indicators. Finally, each section of this document provides a summary of support for the claims.

---

[1] Center for Assessment, & EdReports.org. (2023a). *Review Criteria Interim Assessment English Language Arts Grades 3-8 (v1.0, Final 5/2023).* Dover, NH: EdReports.org and Center for Assessment

Center for Assessment, & EdReports.org. (2023b). *Review Criteria: Interim Assessment Mathematics Grades 3-8 (v1.0, Final 5/2023).* Dover, NH: EdReports.org and Center for Assessment

Conclusions Supporting Claims for Criterion 1.1: Test Development Alignment

Assessment design specifications align to the expectations of college- and career-ready (CCR) standards.

1.1.a Assessment design specifications provide clear expectations and detailed guidance to support the development of high-quality, CCR standards-aligned materials.

Assessment rationale explains the design of the assessment, the benefits of the assessment, and a research foundation grounding the assessment process. The report provides a comprehensive assessment rationale grounded in both the National Reading Panel (NRP, 2000) and the Common Core State Standards (CCSS, 2010). The design is explained through a vertically articulated, K–8 blueprint in which foundational skills (letter names, letter sounds, phonemic awareness, phonics) support fluency, vocabulary, and comprehension across grade levels. Benefits of the CBM system—including efficiency, sensitivity to incremental growth, suitability for universal screening and progress monitoring, and use of Item Response Theory to support alternate-form consistency—are explicitly stated and research-supported.

Item development documentation is sufficiently robust to support writing and reviewing items that measure CCR standards. Item development procedures are specified in detail for each measure type. Letter Names, Letter Sounds, and Phoneme Segmenting reflected random sampling from the universe of items. Fluency word lists and passages were constructed following explicit story grammar specifications (introduction, rising action, climax, resolution). Writers received structured templates, formatting requirements, and guidance on controlling cognitive complexity—documentation sufficient to support consistent, standards-aligned item production. Vocabulary writers were recruited for ELA expertise, trained on Oregon State Vocabulary Standards, and directed to produce items spanning a range of difficulty. Comprehension item writers were trained in Universal Design for Assessment, item development conventions, and CCSS before drafting passages and questions.

Across all item types, assessment design specifications provide clear scoring information and/or rubrics to evaluate students' levels of understanding with respect to CCR standards being measured. Scoring is clearly defined for each measure type. Timed fluency measures score each response as 0 (incorrect) or 1 (correct), yielding a rate-correct-per-minute score. Vocabulary and comprehension items use a single-keyed multiple-choice format scored 0 or 1. Proficient Reading assigns 1 point per correct response (12 points in Grade 2; 20 points in Grades 3–8). The blueprint additionally clarifies that sub-test scores by text type should not be interpreted as independent domain mastery scores, an interpretive constraint consistent with the validated construct structure of the measure.

Passage selection documentation details a process for the review and selection of texts based on the expectations of CCR standards. For Passage Reading Fluency, original narrative fiction was created rather than drawn from existing sources. Writers followed explicit guidelines specifying story grammar structure, grade-appropriate plot and characters, and readability constraints to ensure form comparability. For comprehension, literary, informational, and read-to-perform-a-task texts were developed by educators and aligned to CCSS Reading anchor standards. All texts were reviewed by trained reviewers for grade appropriateness, linguistic accessibility, and standards alignment prior to piloting, constituting a systematic, documented text development and selection process.

Item development documentation includes a description of processes used to ensure items are content-accurate and without technical or editorial flaws. A multi-stage review process is documented across all measure types. Vocabulary and comprehension items were reviewed for alignment to standards, readability, linguistic accessibility, bias and sensitivity, formatting, and keying accuracy before piloting. For comprehension, 3–5 trained reviewers screened each item, with revisions stored in the item database. Fluency passages were reviewed for grade appropriateness, discourse structure, and construct relevance.

Universal Design for Assessment principles were applied during online item development to reduce construct-irrelevant variance, reflecting a systematic approach to content accuracy and editorial quality.

1.1.b Test blueprints and/or assessment design specifications reflect an appropriate distribution of content and related score points, item types, and cognitive demand within test events.

The expected distribution of test content within and across ELA domains is defined within the test blueprint and/or assessment design specifications and reflects the emphasis established in CCR standards. The blueprint explicitly maps content across ELA domains by grade band. Kindergarten and Grade 1 address NRP foundational components: alphabetic knowledge, phonics, and phonemic awareness through Letter Names, Letter Sounds, Phoneme Segmenting, and Word Reading Fluency. Grade 2 introduces vocabulary and passage reading fluency. Grades 3–8 integrate Passage Reading Fluency, Vocabulary, and both Basic and Proficient Reading Comprehension. Table 4 provides a Standards Crosswalk linking each easyCBM® task to its NRP component and CCSS anchor standard, ensuring content coverage across the K–8 ELA domain progression with a developmental arc reflecting CCR standards.

The expected distribution of score points among ELA domains and subdomains is defined within the test blueprint and/or assessment design specifications and reflects the emphasis suggested in CCR standards. Score distribution is addressed through the blueprint's specification of item counts and weights by subdomain. Proficient Reading Comprehension assigns score points across literal (approximately 8–10 items), inferential (approximately 6–8 items), and evaluative (approximately 4–6 items for Grades 3–8) comprehension categories, mirroring the progressively demanding cognitive emphasis of CCSS Reading anchor standards. Basic Reading uses 25 items across five text-type subtests. Fluency, Vocabulary, and Comprehension each generate distinct scores, creating a multi-component profile consistent with the NRP's framework for reading development.

The expected type and range of item types is reflected in the test blueprint and/or assessment design specifications and are appropriate to address the expectations of CCR standards. Multiple item formats are used across the assessment system, each suited to its construct. Early literacy tasks use individually administered, timed oral production. Fluency tasks use continuous oral reading scored by an examiner. Vocabulary and comprehension measures use computer-delivered multiple-choice items with a stem and three options (one correct, two plausible distractors). Comprehension additionally includes a graphic-based read-to-perform-a-task format. This range of item types across the K–8 system reflects the varied cognitive and linguistic demands of CCR standards at each grade level and across distinct reading constructs.

The suggested ranges of cognitive demand are reflected in the test blueprint and/or assessment design specifications and are sufficient to measure the depth of CCR standards. Cognitive demand is most explicitly specified in the comprehension measures. Proficient Reading forms are sequenced from literal to inferential to evaluative comprehension, with specifications requiring a range of difficulty within each type. Grade 2 forms specify 7 literal and 5 inferential items; Grades 3–8 add evaluative items. For vocabulary, the item bank spans from very low to grade-level difficulty with distractors differentiated by meaning. Fluency tasks address automaticity as an intermediate cognitive skill bridging decoding and comprehension. Together, these specifications reflect the depth-of-knowledge demands embedded in CCR standards across the K–8 reading continuum.

Abstract

This document summarizes item development and key findings across a set of Behavioral Research and Teaching technical reports supporting easyCBM® reading-related measures (and selected supporting work). Development followed explicit test specifications, multi-stage review for alignment and accessibility, pilot testing with standardized administration, and psychometric evaluation using methods matched to measure structure. Rasch calibration was commonly used for early literacy and vocabulary, while passage reading fluency relied on classical form comparisons and comprehension measures used IRT testlet modeling to address passage effects. Across reports, most items and forms demonstrated acceptable fit and targeting, and multiple comparable alternate forms were assembled for benchmarking and progress monitoring. Spanish literacy findings underscore the importance of language-specific constructs, with syllable-level measures showing strong relations to fluency. Results in this summary are illustrative of those reported in each technical report; readers should consult the original documents for full analytic detail and context. for K-8 instructional decision making.

The Development of easyCBM

easyCBM® was created by the Behavioral Research and Teaching (BRT) research unit in the College of Education at the University of Oregon. Development began with a grant from the federal Office of Special Education Programs in 2006, bolstered by subsequent grants from the Institute of Education Sciences (IES). In the spring of 2011, the University of Oregon partnered with Riverside Insights to expand easyCBM® to support the needs of school- and district-wide implementations. Because of the dynamic nature of the system, information derived from easyCBM® reflects the most current research and practice for schools.

Basic Tenets of Curriculum Based Measurement

easyCBM® assessments are Curriculum Based Measures (CBMs), which are standardized measures that sample from a year's worth of curriculum to assess the degree to which students have mastered the skills and knowledge deemed critical at each grade level. They are also known as 'general outcome measures.' Curriculum Based Measurement (CBM) has a long research history, beginning with Stanley Deno and colleagues at the University of Minnesota. CBM was originally created to assist special education teachers in developing individual education plans and monitoring student progress. The use of these measures soon expanded to include general education, as they provide reliable and valid assessments of student progress in reading and mathematics (Tindal, 2013)[1]. In particular, the measures can be used for universal screening (benchmark testing) and progress monitoring, as they are sensitive to small, incremental changes in performance and are efficient to administer and score. Curriculum Based Measures such as Oral Reading Fluency correlate highly with state standardized reading tests and reading comprehension.

The measures that are part of the easyCBM® system are often referred to as 'next-generation CBMs,' as an advanced form of statistics, Item Response Theory (IRT), was used during development to increase the consistency of the alternate forms of each measure and to increase the sensitivity of the measures to monitor growth. At each grade level, alternate forms of each measure are designed to be of equivalent difficulty, so as teachers monitor student progress over time, changes in score reflect changes in student skill rather than variations in the difficulty level of the test forms.

[1] Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education (International Scholarly Research Network)*, Volume 2013, Article ID 958530, 29 pages. http://dx.doi.org/10.1155/2013/958530

easyCBM® K–8 Reading Blueprint using CCSS and National Reading Panel

The easyCBM® Reading Assessment System (K–8) is vertically articulated and grounded in the National Reading Panel (2000)[2] and the Common Core State Standards (2010)[3]. Foundational skills (K–1) measure letter names, letter sounds, phonemic awareness and phonics; Grades 2–8 integrate word and passage reading fluency, vocabulary, and comprehension. Passage Reading Fluency supports automaticity; Vocabulary and Proficient Reading assess literal, inferential, and evaluative comprehension. Alternate-form design, story-grammar consistency, and Rasch scaling support reliability and validity. The blueprint demonstrates construct coherence across grades and alignment to MTSS screening frameworks. This technical blueprint documents the easyCBM® Reading assessment system (Grades K–8) blueprint using the National Reading Panel (NRP, 2000) framework and the Common Core State Standards (CCSS). It integrates task descriptions from the easyCBM® User Manual[4] and a benchmark Student Reading Booklet and articulates a comprehensive construct representation of reading across grade levels.

The National Reading Panel (NRP, 2000) identified five essential components of reading: (1) Phonemic Awareness, (2) Phonics, (3) Fluency, (4) Vocabulary, and (5) Comprehension. These components represent a developmental progression in which early decoding skills support automaticity, which in turn supports comprehension and higher-order reasoning. The blueprint primarily used the NRP for Foundational skills to address phonemic awareness, phonics, and fluency in Grades Kindergarten, 1 and 2, while from Grade 3 through 8, Passage Reading Fluency was deployed with Vocabulary and both Basic and Proficient Reading Comprehension.

In Grades 3 and beyond, the Common Core State Standards (CCSS) for English Language Arts organize Reading Literature (RL) / Reading Informational Text (RI) anchor standards (K–12). RL/RI standards address comprehension processes: key ideas, craft and structure, integration of knowledge, and range of reading.

Foundation for Reading Measurement

We have divided our blueprint into Early Reading Foundational Skills in Grades Kindergarten through Grade 2 and then Comprehensive Reading Skills in Grades 3 through 8 expressed through Word and Passage Fluency, Vocabulary, and finally Basic and Proficient Reading Comprehension.

Early Reading Foundational Skills Blueprint

Letter Names (LN) is a timed, individually administered measure for Kindergarten and Grade 1. It consists of both upper- and lower-case alphabet letters presented in an order based on empirical evidence of difficulty. The student is asked to name the letters presented by row. The student's score is the number of letter names identified correctly in one minute.

The English language contains 26 letters and with both upper and lower case considered, the total sample is 52 letters (though some letters look the same and differ only in size. The blueprint is based on each form presenting a random sample of letters in both upper (capitalized) and lower case with 10 letters per row and 10 rows. Students name letters from left to right and continue for 1 minute. Letters named incorrectly are scored 0 and letters named correctly are scored 1. At the end of one-minute, the test is stopped and scored for the number of letters named correctly and incorrectly with the difference providing the rate correct per minute.

---

[2] National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature*. Washington, DC: National Institutes of Health.

[3] National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards*.

[4] *easyCBM® District User Manual* (2025). University of Oregon, Behavioral Research & Teaching.

Letter Sounds (LS) is a timed, individually administered measure for Kindergarten and Grade 1. It consists of upper- and lower-case letters and letter pairs presented in an order based on empirical evidence of difficulty. The student is asked to produce the sound made by the letter or letter pair. The student's score is the number of letter sounds produced correctly in one minute.

The English language contains 26 letters but 44 sounds: 19 consonants, 7 digraphs, 5 'r-controlled' sounds, 5 long vowels, 5 short vowels, 2 'oo' sounds, 2 diphthongs. The blueprint is based on each form presenting a random sample of letter sounds with 10 per row and 10 rows. Students sound the letter(s) presented from left to right and continue for 1 minute. Letters sounded incorrectly are scored 0 and letters scored correctly are scored 1. At the end of one-minute, the test is stopped and scored for the number of letters sounded correctly and incorrectly with the difference providing the rate correct per minute.

Phoneme Segmenting (PS) is a timed, individually administered measure for Kindergarten and Grade 1. The student is asked to identify the individual phonemes in each word that is orally presented by the assessor. In contrast to letter sounds, this task is presented by having the administrator sound out a word and the students. For example, the administrator is directed to state: “I am going to say a word, and you will give me the sounds you hear in that word. If I say *cap*, you will say /c/ /a/ /p/. If I say *it*, you will say /i/ /t/. If I say *top*, you will say /t/ /o/ /p/. Let’s try.” The student's score is the number of correct phonemes or segments identified in one minute.

Reading Fluency Measure Blueprint

Word Reading Fluency (WRF) is a timed, individually administered measure for Kindergarten to Grade 3 which assesses the fluency of words read in isolation. The student is given a list of words presented in order of increasing difficulty and is asked to read the words row by row. The student's score is the number of words read correctly in one minute.

Words for Word Reading Fluency were selected from a variety of sources, including Dolch word lists, online grade-level word lists, and Fry's 'instant 1000 words.' They include words with both regular and irregular sound patterns and in a variety of lengths. The words were piloted in a large multi-grade study in 2006; the difficulty of each word was then calculated, and test forms were constructed to be of equivalent difficulty within each given grade.

Passage Reading Fluency (PRF) is a timed, individually administered measure for Grades 1 to 8 which assesses the fluency of words read in context. Each passage consists of a brief work of original narrative fiction with content and passage complexity suitable for a given grade level. The student's score is the number of words read correctly in one minute.

For passage construction, the passages maintain a consistent discourse profile, so performance differences are not driven by shifts in text type or structure, and that forms maintain stable reading demands and comparable overall difficulty. These measures were created to follow the same story grammar across passages (introduction, rising action, climax, resolution) to reduce variation in difficulty across the different forms. The plots and characters were selected to be appropriate for the grade levels.

For both fluency measures (words and passages), performance is scored with incorrectly pronounced words (substitutions or omissions) as 0 points and correctly pronounced words as 1 point. The total is the number of words correct, incorrect, and words correct per minute. Within the broader literacy construct progression, fluency is an intermediate outcome that draws on earlier alphabetic-principle and phonological-awareness skills and serves as a strong predictor of early comprehension. Fluency measures are therefore intended to capture a student’s ability to decode and recognize words rapidly and accurately in connected text (passage reading fluency) and in word lists (word reading fluency), with sensitivity to small changes over short instructional intervals (about 1–2 weeks).

Vocabulary Measure Blueprint

Vocabulary is an untimed assessment that measures vocabulary proficiency appropriate for a student's grade level, designed for students in Grades 2 to 8. The student's score is the number of correct responses. The words included in the Vocabulary measure were selected from a variety of content materials and were extensively field-tested. The bank of items represents a wide range of difficulty aligned to grade level content standards.

The Vocabulary CBM blueprint (Grades 2–8) is built around a large, standards-referenced item bank used to assemble benchmark and progress-monitoring forms. Item creation begins with explicit targets: writers use the Oregon State Standards for Vocabulary, and each item is tagged to a specific sub-domain within a standard in the item database. Writers are recruited for English language arts expertise and experience with struggling learners (including special education), and they are asked to produce items spanning very low to grade-level performance ranges.

Item format is tightly specified for computer delivery. Each item presents a single sentence with the target word highlighted, followed by three multiple-choice options. Writers were trained to reduce construct-irrelevant complexity while preserving standards alignment: use simple, direct language, avoid idioms and unnecessary clauses, focus attention on one idea, and keep grammar parallel between the stem and each option. Options are designed to be similar in length and complexity, mutually exclusive, and differentiated primarily by meaning. Writers use grade-appropriate vocabulary and contexts (supported by reference lists) and craft distractors that are plausible and relevant to the sentence.

Procedures support alternate-form comparability. Items were authored in structured templates, reviewed for clarity, alignment, formatting, and keying accuracy, and then piloted online using a design that combines anchor items with randomly selected items from the bank. Online delivery presents one item per screen, captures responses automatically, and randomizes option order to reduce cheating. Item statistics from piloting (difficulty and appropriateness) are then used to assemble multiple comparable benchmark and progress-monitoring forms from the calibrated bank.

Basic Comprehension Measure Blueprint

Basic Reading is an untimed assessment that measures student comprehension utilizing a variety of text, designed for students in Grades 3 to 8. The Basic Reading measures include Read to Perform a Task, Informational Text, and Short Literary Text passages. Each of the passages is about 200 words with 5-10 questions presented after each passage. The Basic Reading Comprehension blueprint uses short sets of text-dependent questions linked explicitly to grade-level standards. Item writers (educators with grade-level expertise and experience with struggling readers) were trained in Universal Design for Assessment, item development conventions, and the content standards. Writers created literary stories and informational passages for Literary and Informational Text subtests and designed graphics for “Read to Perform a Task” subtests; a graphic artist finalized visuals for computer delivery. For each prompt (text or graphic), writers produced five selected-response questions, each aligned to a specific grade-level standard. Every question includes a stem plus three options (one correct answer and two plausible distractors), supporting efficient scoring and reduced construct-irrelevant variance.

Students complete a total of 25 items organized as five subtests (each subtest = one passage/graphic plus five questions). Materials were developed for online delivery, stored in a secure database, and subjected to multi-reviewer quality checks (3–5 trained reviewers) for bias, grade appropriateness, and adherence to Universal Design principles, with revisions made directly in the database. The blueprint anticipates large item pools: approximately 900 items per grade (Grades 3–8) are available for empirical testing.

Common items across test takers enable concurrent scaling of the full grade-level pool even though each student sees only a small sample. The computer system randomly assigns subtests to reduce class effects and broaden sampling. Overall, the blueprint yields a standards-mapped, computer-scored comprehension system with consistent prompt-question structure, systematic review, and piloting designed to calibrate a large bank into usable forms.

Proficient Comprehension Measure Blueprint

Proficient Reading is an untimed assessment that measures student comprehension of written text, designed for students in Grades 2 to 8. Students read an original narrative fiction story and answer multiple-choice questions that are text dependent. Passages contain approximately 700 words in Grade 2 to around 1,300-1,500 words in Grades 3-8. Multiple choice questions at the end of the passage assess students' literal (approximately 8-10), inferential (approximately 6-8), and (in Grades 3 to 8) evaluative comprehension of text (approximately 4-6). The Proficient Reading Comprehension is a group-administered, passage-based measure for Grades 2–8. Each item includes a stem plus three answer choices (one correct, two plausible distractors) and is scored 1 point per correct response (12 points in Grade 2; 20 points in Grades 3–8). The selected-response format was chosen to resemble classroom reading tasks while improving reliability, ease of use, and cost-effectiveness for open-ended formats.

Passage construction was controlled through written story specifications to increase comparability and minimize construct-irrelevant variance across forms. Story writers received guidance on length and key narrative/discourse features—characters, setting, and plot—so forms are similar in structure and demand. Stories were drafted by trained educators/graduate students and screened against criteria before piloting, with quality control embedded in both story and item development.

Question design followed a deliberate comprehension blueprint. Across grades, items assess literal and inferential comprehension; evaluative comprehension is added for Grades 3–8. In Grade 2, the specified distribution is 7 literal and 5 inferential items. In Grades 3 through 8, Specifications require a range of difficulty within each comprehension type and guide sequencing so forms progress from easier literal items toward the most challenging items, ending with the highest comprehension level assessed in that grade. The result is a set of comparable narrative passages paired with systematically balanced question types, difficulty, and ordering across alternate forms.

Summary of Grade and Season Tasks for Benchmarks

- Kindergarten students take LN, LS, and PS (fall), LS, PS, WRF (winter and spring)
- Grade 1 students take LS, PS, and WRF (fall) and LS, WRF, and PRF (winter and spring)
- Grade 2-8 students take PRF, Vocabulary, and Proficient Reading Comprehension (fall, winter, and spring)

Grade Level Summaries of Blueprints

Kindergarten Blueprint

Kindergarten students complete Letter Sounds (LS), Phoneme Segmenting (PS), and Word Reading Fluency (WRF) in winter and spring. Fall additionally includes Letter Names (LN). These tasks directly measure NRP foundational components.

Table 1. Tasks for Kindergarten Blueprint

Task	Construct (NRP)	CCSS Alignment	Administration	Score
Letter Names (LN)	Alphabetic Knowledge	RF.K.1d	1-minute timed	Letters correct/min
Letter Sounds (LS)	Phonics	RF.K.3a–c	1-minute timed	Sounds correct/min
Phoneme Segmenting (PS)	Phonemic Awareness	RF.K.2d	1-minute timed	Phonemes correct/min
Word Reading Fluency (WRF)	Fluency/Decoding	RF.K.3–4	1-minute timed word list	Words correct/min

Kindergarten measures focus exclusively on foundational skills. No comprehension task is administered. The blueprint emphasizes phonemic awareness and alphabetic principle as precursors to decoding fluency.

Grade 1 Blueprint

Grade 1 students transition from foundational skills toward connected-text fluency. Winter tasks include Letter Sounds (LS), Word Reading Fluency (WRF), and Passage Reading Fluency (PRF).

Table 2. Tasks for Grade 1 Blueprint

Task	Construct (NRP)	CCSS Alignment	Administration	Score
Letter Sounds (LS)	Phonics	RF.1.3	1-minute timed	Sounds correct/min
Word Reading Fluency (WRF)	Fluency	RF.1.4a–c	1-minute timed word list	Words correct/min
Passage Reading Fluency (PRF)	Fluency	RF.1.4	1-minute oral passage	Words correct/min

PRF introduces connected-text reading with consistent story grammar structure. This supports development of automaticity and begins linking fluency with comprehension.

Grades 2–8 Blueprint

Beginning in Grade 2, the blueprint integrates three major tasks each benchmark period: (1) Passage Reading Fluency (PRF), (2) Vocabulary, and (3) Proficient Reading.

Table 3. Tasks for Grades 2–8 Blueprint

Task	Construct (NRP)	CCSS Alignment	Structure	Score
Passage Reading Fluency (PRF)	Fluency	RF.2–5.4	1-minute oral passage	Words correct/min
Vocabulary	Vocabulary	RL/RI.4; L.4–6	Untimed; ~20–25 items	Number correct
Proficient Reading	Comprehension	RL/RI.1–10	2 passages × 20 items (40)	Number correct

Each Proficient Reading form contains two grade-level passages, each followed by 20 multiple-choice items (40 total). Items represent three comprehension types: literal recall, inferential reasoning, and evaluative analysis. Comprehension Blueprint Analysis: Comprehension questions align to CCSS Reading Anchor Standards. Typical distribution per passage (approximate):

- Literal (Standards 1–3): 8–10 items
- Inferential (Standards 1, 3–6): 6–8 items
- Evaluative/Integrative (Standards 6–9): 4–6 items

This distribution supports increasing cognitive demand across grades.

Vocabulary tasks measure depth of word knowledge and context-based meaning, supporting RL/RI.4 and Language standards (L.4–6).

Vertical Construct Progression

The blueprint demonstrates a coherent progression:

- Kindergarten–Grade 1: Phonemic Awareness → Phonics → Fluency
- Grades 2–3: Fluency → Vocabulary → Basic Comprehension
- Grades 4–5: Vocabulary depth + Analytical Comprehension
- Grades 6–8: Advanced Comprehension, Text Integration, Evaluation

This progression reflects both NRP evidence and CCSS grade-level expectations. Timed CBM measures provide growth-sensitive screening data, while untimed comprehension measures capture higher-order literacy skills. Together, the tasks represent the full construct of reading. The easyCBM® K–8 Reading Assessment Blueprint demonstrates strong alignment to the National Reading Panel framework and the Common Core State Standards. Foundational skills are measured directly and efficiently in early grades, while later grades emphasize vocabulary and comprehension complexity. The design ensures construct coverage, vertical coherence, and psychometric defensibility for screening and benchmark purposes.

Table 4. Standards Crosswalk with CCSS, NRP, and easyCBM®

CCSS Anchor Standard	NRP Component	easyCBM® Task	Construct Evidence
RF.K–2 Foundational Skills	Phonemic Awareness / Phonics	PS, LS, LN	Timed segmentation and sound production
RF.1–5.4 Fluency	Fluency	WRF, PRF	Words correct per minute
RL/RI.1–8.1–3 Key Ideas	Comprehension	Proficient Reading	Literal & inferential questions
RL/RI.4 Vocabulary	Vocabulary	Vocabulary Task	Context-based meaning
RL/RI.8–9 Integration	Comprehension	Proficient Reading	Evaluative reasoning

The easyCBM® reading assessment system aligns with evidence-based reading instruction frameworks (National Reading Panel, 2000) and the Common Core State Standards (NGA & CCSSO, 2010). It operationalizes phonemic awareness, phonics, fluency, vocabulary, and comprehension across K–8. In the technical documents included in the remainder of this report, scoring rules are presented for all tasks. The early reading skills (LN, LS, PS) and the fluency measures (WRF and PRF) contain explicit administration and scoring rules in the User Manual. These measures all require production responses so standardized scoring is critical. The other measures (Vocabulary, and both Basic and Proficient Reading Comprehension use selected responses so scoring rules are built into the computer-based administration. These latter three measures are also the most important tasks requiring a blueprint related to the CCR expectations. These measures also contain explicit references in the count of items related to various types of comprehension. Processes for ensuring content accuracy and avoiding technical/editorial flaws are not only referenced, but a QA workflow is also provided in the individual technical reports. The expected distribution of content within and across ELA domains and the distribution of score points by domain/subdomain is specified in the technical documents on test development. In the early foundation skills of LN, LS, and PS, the distribution is normal but changes over the year. All other measures present normal distributions that show increases in median and mean performance but no changes in variation.

Item Development

Item development is the foundation of any assessment system that intends to support high-stakes decisions, frequent progress monitoring, or both. In curriculum-based measurement (CBM) systems, item development must accomplish two objectives that can be in tension: (a) create content that is instructionally interpretable and aligned to the skills educators teach, and (b) produce scores that are psychometrically defensible across students, forms, and administrations. The technical reports summarized in this document illustrate how these objectives were addressed across a sequence of

easyCBM® reading-related measures spanning early literacy (Kindergarten to Grade 2), passage reading fluency (Grades 5 through 8), vocabulary (Grades 2 through 8), CCSS-aligned reading comprehension (Grades 3 through 8), and Spanish literacy measures (Kindergarten to Grade 2). Together, they represent an integrated approach to designing brief, repeatable measures that can be used for screening and progress monitoring, while also supporting the creation of multiple alternate forms to minimize practice effects and preserve score meaning over time.

First, across the reports, item and form development followed a common set of design principles. First, developers began with explicit test specifications that defined the construct, targeted grade range, intended use (benchmarking versus progress monitoring), timing and administration conditions, and the kinds of student responses required. For early literacy measures, specifications focused on foundational skills underlying the alphabetic principle: letter identification, letter-sound correspondence, and oral segmentation of phonemes. For fluency measures, specifications addressed passage characteristics such as length, genre, readability constraints, formatting, and scoring rules aligned with correct words per minute conventions. For vocabulary and CCSS reading comprehension, specifications addressed the breadth of grade-level standards, item formats (primarily multiple-choice), and cognitive demands to ensure the measures captured targeted knowledge and processes without introducing avoidable barriers.

Second, content was drafted by trained item writers or educators with relevant classroom expertise and then refined through multi-stage review. Reviews typically included checks for alignment to the construct and grade-level expectations, readability and linguistic accessibility, bias and sensitivity, and adherence to formatting and presentation constraints required for standardized delivery (paper-based for some fluency tasks, computer-based for multiple-choice tasks). Universal Design for Assessment principles were emphasized for item sets delivered online, with attention to simplified language where appropriate, clear stems and options, and avoidance of unnecessary complexity that could introduce construct-irrelevant variance. In Spanish literacy development, the review process also reflected the linguistic structure of Spanish, recognizing that syllable-level regularities and orthographic transparency change both the nature of early reading development and the indicators that best represent it.

Third, item pools and draft forms were evaluated empirically through pilot testing. Pilot designs varied by measure type but consistently aimed to collect sufficient data to evaluate item functioning, identify problematic content, and assemble comparable alternate forms. Many reports used overlapping anchor sets (often around 20% common content) to support calibration and equating across forms when groups were nonequivalent or when forms were administered across different samples. Administration procedures were standardized and used trained assessors for individually administered early literacy and fluency tasks, while computer-based platforms enabled efficient capture of large samples for multiple-choice measures. The goal was not only to evaluate average performance, but to obtain evidence about how individual items behaved across a range of student ability.

Fourth, analyses combined classical test theory (CTT) and item response theory (IRT), selected to match the structure of the measure. For early literacy and vocabulary measures, one-parameter logistic (1PL) Rasch models were frequently used to estimate item difficulty, standard errors, and fit statistics such as outfit mean square. These analyses supported decisions about retaining, revising, or removing items and helped assemble forms with comparable difficulty. For reading comprehension measures built around passages with clustered items, testlet-based IRT models were used to acknowledge local dependence among items tied to the same text. For passage reading fluency, each passage served as the scored unit, so CTT-based comparisons of means and correlations across passages provided the most direct evidence for form equivalence and the need for targeted revisions.

Finally, item development was treated as iterative. Empirical findings informed revision cycles, including the removal of items with inestimable parameters or poor distractor functioning, refinement of passages that were too easy or too difficult, and the selection of anchor forms to stabilize interpretation across

administrations. Across measures, a key outcome of the development work was the creation of multiple equivalent alternate forms (often 10 to 20 forms) that support repeated measurement while controlling for unwanted variability. This emphasis on comparability reflects the central requirement for progress monitoring: change in scores should be interpretable as change in student skill, not change in form difficulty.

The technical reports also demonstrate that "item development" in CBM systems extends beyond writing good items. It includes creating a coherent assessment framework in which forms, scoring rules, administration procedures, and psychometric models work together to support specific uses. In early literacy, the emphasis is on efficiency and sensitivity to emerging skills. In fluency, the emphasis is on passage comparability and stable scoring. In vocabulary and comprehension, the emphasis is on aligning to standards while maintaining accessibility and reliable scaling. In Spanish literacy, the emphasis is on selecting indicators that reflect the linguistic properties of the language of instruction. Taken together, the reports illustrate a development approach designed to yield measures that are practical for educators and technically defensible for reporting and decision-making.

In addition, the development work emphasizes the link between item content and the decisions educators make. CBM measures are intended to be actionable, so item development prioritized representativeness (sampling skills students are expected to demonstrate), transparency (clear scoring rules and limited construct-irrelevant complexity), and efficiency (brief administration times). These considerations appear in the early literacy and fluency work, where timing and standardized prompts are essential, and in computer-based measures, where consistent item presentation supports comparability.

The reports also highlight practical trade-offs. Overly strict fit rules can remove items that help content coverage, while overly permissive rules can introduce noise. Likewise, maximizing form equivalence can reduce content variation across forms, but too little variation can increase practice effects. The development approach reflected here treats these issues as iterative design decisions supported by empirical evidence rather than as one-time judgments.

Highlights of Findings from Technical Reports

Across the technical reports, the most consistent finding is that carefully specified development procedures can yield forms that are sufficiently comparable for repeated use in screening and progress monitoring. While the details vary by measure type, the reports collectively show that a disciplined sequence of specification, expert review, piloting, and psychometric evaluation supports the creation of item banks and forms that function as intended across grades and contexts. At the same time, it is important to interpret these findings as illustrative of those reported in each of the individual technical reports. **Note**: The summaries and tables highlight recurring patterns and representative outcomes, but they do not reproduce all analyses, nuances, or caveats from the full reports.

Early literacy development (Kindergarten through Grade 2) provides an example of item-bank based development with Rasch calibration. Letter Names, Letter Sounds, and Phoneme Segmenting items largely exhibited acceptable fit to a 1PL model, with outfit statistics typically within preferred ranges and difficulty estimates that supported differentiation among students. Items that showed extreme misfit or inestimable parameters were flagged for review and, when appropriate, removed. A practical implication is that even when a subset of items performs less ideally, the broader item pool can still support strong form construction if decisions incorporate both fit and measurement contribution (e.g., standard errors and coverage of the ability continuum). Using calibrated pools and anchor designs, the development work supported the assembly of many equivalent alternate forms, which is essential for frequent monitoring.

For passage reading fluency in Grades 2 to 8, the central psychometric question shifted from "How do items function?" to "How comparable are passages when each passage is the unit score?" The reports show that repeated measures designs and inter-form correlations are effective for identifying passages

that align in difficulty and those requiring revision. In all grades, correlations among passages were generally strong, supporting the conclusion that multiple forms can be used interchangeably for progress monitoring with minimal distortion. Mean differences across passages were used diagnostically: passages that consistently deviated from the target difficulty band were revised or set aside. A notable outcome is the deliberate use of anchor passages or benchmark forms that provide stable reference points within a larger set of progress-monitoring passages.

Reading comprehension development and evaluation, including the earlier comprehension adequacy report and the CCSS-aligned Grade 3 through Grade 8 assessments, illustrates the importance of matching analytic methods to test structure. When comprehension items are nested within passages, local dependence is expected; testlet-based IRT modeling provides a principled way to estimate difficulty and fit while accounting for passage effects. Across grades, the reports indicate that most items and testlets demonstrated acceptable model fit, with difficulty values spread to capture meaningful variation among students. Where items were problematic, the issues often involved limited variability, extreme difficulty, or distractor patterns that reduced discrimination. These findings point to the value of including distractor analyses alongside difficulty and fit indices, because a multiple-choice item can appear to fit an IRT model yet fail to provide useful diagnostic information if distractors don't attract responses by lower-performing students.

The sequence of vocabulary development reports (Grades 2 through 8) provides one of the clearest examples of iterative item refinement under a Rasch framework. Across grades, the general pattern was stable calibration, acceptable fit, and the successful assembly of both benchmark and progress-monitoring forms with closely aligned mean measures. Items with underfit or non-functioning distractors were removed, while overfitting items were sometimes retained when distractor functioning indicated that the item behaved predictably rather than misleadingly. The reports also emphasize that universal design considerations are not merely a fairness add-on; they support technical quality by reducing construct-irrelevant variance that can inflate error and reduce interpretability. The practical result is a set of short vocabulary measures that can be administered repeatedly, scaled consistently, and interpreted within an RTI decision-making framework.

The Spanish literacy development report adds an important linguistic and conceptual finding: the strongest early indicators of Spanish reading fluency were often syllable-level measures rather than phoneme-level measures. This is consistent with Spanish orthographic transparency and highlights a key lesson for assessment development: constructs must be operationalized in ways that reflect how skills develop in the language and instructional context. The technical adequacy evidence for the Spanish measures included strong alternate-form correlations for syllable reading and reading fluency measures, reinforcing the feasibility of building reliable progress monitoring tools when development aligns with language structure.

Taken together, the findings support three high-level conclusions. **First**, systematic development procedures can yield technically adequate item pools and forms across diverse reading constructs and grade levels. **Second**, combining CTT and IRT evidence strengthens decision-making because different measures require different kinds of comparability evidence. **Third**, the reports collectively demonstrate that progress monitoring systems benefit from an explicit commitment to form equivalence, iterative refinement, and alignment to both developmental theory and instructional practice. These conclusions, while summarized here, should be read as representative patterns drawn from the set of technical reports rather than as a substitute for the full technical documentation in each report.

The reports also clarify that "acceptable" psychometric evidence is used as a decision framework. Fit thresholds, difficulty targets, and distractor patterns guide concrete actions: retain, revise, replace, or remove. This decision orientation matters in operational settings because it connects technical evidence to the ongoing maintenance of item pools and forms.

An important caveat for this summary is that the results presented here are illustrative of those reported in each of the technical reports. The narrative and tables emphasize recurring patterns (for example, generally acceptable fit, strong inter-form correlations, and successful assembly of multiple alternate forms), but they do not reproduce every statistic, subgroup analysis, or design nuance from the originals. Readers should consult the full reports for details on sampling, administration conditions, and measure-specific decision rules.

Summary of Technical Report 0039: The Development of Early Literacy Measures for Use in a Progress Monitoring Assessment System: Letter Names, Letter Sounds, and Phoneme Segmenting (Alonzo & Tindal, 2007a).

Methods

This report describes the development of three early literacy progress monitoring measures designed for Kindergarten and first-grade students: Letter Names, Letter Sounds, and Phoneme Segmenting. The goal was to create calibrated item banks that could support the construction of multiple equivalent forms for frequent progress monitoring of foundational reading skills.

Item development began with large pools of candidate items. The Letter Names and Letter Sounds measures included uppercase and lowercase letters, while the Phoneme Segmenting measure used words that students segmented into individual phonemes. Pilot testing was conducted in spring 2006 with Kindergarten and first-grade students in a suburban Pacific Northwest school district. Response data were collected from several hundred to more than two thousand students depending on the item.

A common-item nonequivalent groups design was used so that items across different test forms could be calibrated on a single scale. Item parameters were estimated using a one-parameter Rasch model, which provided estimates of item difficulty, standard errors, and item fit statistics. These analyses were used to determine whether items functioned appropriately and should be retained in the item bank.

Results

Rasch analyses produced difficulty estimates for each item across the three measures. For the Letter Names measure, most letters produced stable estimates, although three letters were removed due to unreliable parameters. The remaining items were grouped into easy, moderate, and difficult categories to guide the construction of alternate forms.

For the Letter Sounds measure, several letter-sound items with unstable estimates were excluded. Remaining items were categorized by difficulty and used to assemble equivalent forms. For the Phoneme Segmenting measure, words with poor model fit were removed, leaving 181 items in the final item bank. These items were organized into fourteen difficulty categories based on Rasch difficulty estimates. Using these calibrated item pools, the researchers constructed twenty alternate forms of each measure for Kindergarten and first grade. Items were distributed across forms according to difficulty levels to ensure comparable measurement for progress monitoring.

Table 5. Example of IRT Analysis for Technical Report 0039

Table 2
Results of IRT Analysis of Letter Names Test, Spring Pilot, 2006

Letter	Upper or Lower Case	Count	Measure	Mean Square Outfit
o	Lower case	306	-2.55	.77
X	Upper case	297	-2.14	1.65
A	Upper case	297	-1.67	.24
s	Lower case	1036	-1.64	2.36
O	Upper case	433	-1.64	.39
B	Upper case	306	-1.42	1.29
E	Upper case	433	-1.18	1.70
a	Lower case	433	-1.18	.46
T	Upper case	433	-1.18	.33
x	Lower case	297	-1.17	.55
e	Lower case	297	-.99	.86
r	Lower case	306	-.96	.73
Z	Upper case	306	-.90	1.40
S	Upper case	306	-.77	.81
L	Upper case	297	-.72	.62
t	Lower case	306	-.71	1.17
R	Upper case	433	-.65	.96
N	Upper case	1036	-.60	.41
p	Lower case	1036	-.60	.65
C	Upper case	297	-.54	1.45
m	Lower case	433	-.52	.31
D	Upper case	306	-.49	.67
P	Upper case	297	-.40	.82
n	Lower case	306	-.39	.40
F	Upper case	433	-.39	.32
f	Lower case	306	-.34	.33
I	Upper case	433	-.27	.85
K	Upper case	306	-.20	.54
k	Lower case	297	.05	.98
M	Upper case	297	.22	1.10
i	Lower case	306	.24	.77
c	Lower case	433	.26	.73
G	Upper case	1036	.31	.73
v	Lower case	297	.51	.96
z	Lower case	433	.65	.60
W	Upper case	306	.69	1.33
U	Upper case	297	.81	.80
h	Lower case	1036	.85	.77
Q	Upper case	297	.86	.98
u	Lower case	306	.88	.69
w	Lower case	433	.92	1.54
y	Lower case	730	.98	.62
l	Lower case	433	1.06	1.80

Table 6. Summary Table for Technical Report 0039

Category	Summary
Purpose	Develop early literacy progress monitoring measures for **Letter Names, Letter Sounds, and Phoneme Segmenting**.
Target Population	Kindergarten and Grade 1 students learning foundational reading skills.
Item Development	Large item pools were created using **uppercase/lowercase letters** and **phoneme segmentation words**.
Pilot Testing	Spring 2006 pilot with **hundreds to over 2,000 student responses per item** in a suburban school district.
Item Analysis	**Rasch modeling** used to estimate item difficulty and evaluate item fit; poorly functioning items were removed.
Final Forms	Remaining items were organized by **difficulty levels** and used to construct **20 equivalent alternate forms** for progress monitoring.

Reference

Alonzo, J., & Tindal, G. (2007). *The development of early literacy measures for use in a progress monitoring assessment system: Letter names, letter sounds, and phoneme segmenting (Technical Report 0039)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0040: The Development of Word and Passage Reading Fluency Measures in a Progress Monitoring Assessment System (Alonzo & Tindal, 2007b).

Methods

This report describes the development and initial validation of curriculum-based measures (CBM) for reading fluency, specifically Word Reading Fluency (WRF) and Passage Reading Fluency (PRF), intended for use in kindergarten through Grade 4 as progress-monitoring tools within Response to Intervention (RTI) systems. The goal was to create sets of alternate forms that could be administered repeatedly while maintaining comparable difficulty.

For Word Reading Fluency, researchers created a large bank of candidate words representing a wide range of reading difficulty. Individual words were analyzed using Rasch item-response modeling to estimate item difficulty and to place words on a common measurement scale. Based on the Rasch difficulty estimates, the words were organized into four difficulty strata, which were then used to construct equivalent alternate forms. Each form included a balanced mix of words from the four strata to ensure consistent difficulty across administrations. A total of 20 alternate WRF forms were constructed.

For Passage Reading Fluency, researchers wrote a large pool of candidate reading passages aligned with grade-level expectations for vocabulary, syntax, and readability. Each passage was reviewed by both the report's lead author and an experienced classroom teacher to ensure instructional appropriateness and comparable complexity. From this pool, 20 passages per grade level were selected to serve as alternate forms. Pilot testing data were collected from elementary students, and analyses focused on mean score comparisons and inter-form correlations to determine whether the passages functioned equivalently.

Results

Results indicated that the development procedures produced strong alternate-form equivalence for both reading fluency measures. Rasch analyses confirmed that the word difficulty estimates used to construct the WRF forms produced balanced distributions of item difficulty across forms. Mean performance levels across WRF forms were similar, indicating that no individual form was substantially easier or harder than others.

For the Passage Reading Fluency measures, analyses of pilot data showed strong correlations among alternate passages, demonstrating that the passages measured the same underlying fluency construct. Comparisons of average student performance across passages showed only minor variation, suggesting that the development procedures successfully controlled passage difficulty.

Together, these findings indicate that the 20 alternate forms developed for both WRF and PRF provide technically adequate measures for repeated progress monitoring in early reading. The combination of Rasch modeling, expert review, and empirical form-equivalence analyses supported the conclusion that the measures can be administered frequently to monitor student growth without introducing substantial measurement error due to form difficulty differences.

Table 7. Typical Results of IRT Analysis from Technical Report 0040

Table 2
Results of IRT Analysis of Word Reading Fluency, Spring Pilot, 2006, Grades K-3

Word	Count	Measure	Mean Square Outfit
the	1960	-6.21	1.10
or	1960	-2.81	2.72
will	2654	-3.78	0.85
number	2198	-1.55	0.66
of	1960	-2.93	9.90
about	2654	-1.54	0.58
remain	2654	2.13	3.14
no	2654	-3.71	4.43
and	2654	-5.20	1.15
little	1960	-3.03	4.45
much	1960	-1.80	1.69

Table 8. Example of Descriptive Statistics from Technical Report 0040

Table 4
Descriptive Statistics for Grade 1 Passage Reading Measures

Passage	Title	*n*	*M*	*SD*
Gr1PR_1_C	The New Puppy	60.94	40.00	207
Gr1PR_2_C	Ice Cream Treat	62.98	37.31	205
Gr1PR_3_C	Patty the Pig	55.79	34.76	204
Gr1PR_4_C	The Dog Show	62.20	37.65	205
Gr1PR_5_C	Trains	57.73	36.96	206
Gr1PR_6_C	The Best Story Ever	63.44	42.30	206
Gr1PR_7_C	Making a Pizza	71.17	38.99	206
Gr1PR_8_C	Ocean Fish	59.60	35.70	206
Gr1PR_9_C	The Puppy	58.99	36.23	201
Gr1PR_10_C	The Airplane Ride	64.14	39.44	201
Gr1PR_11_C	The Story for Mom	60.97	40.87	201
Gr1PR_12_C	Loose Tooth	56.11	38.61	200
Gr1PR_13_C	A Prized Garden	62.40	35.52	203
Gr1PR_14_C	The Lake	55.44	36.52	204
Gr1PR_15_C	The Swing	67.68	40.42	204
Gr1PR_16_C	Play Ball	60.18	37.93	202
Gr1PR_17_C	Sick Day	61.14	39.07	195
Gr1PR_18_C	Hannah the Hamster	63.73	40.13	195
Gr1PR_19_C	Playful Sam	60.70	38.99	195
Gr1PR_20_C	The Fair	64.23	41.37	195

Results of Repeated Measures Analysis Across All 20 Passages

Grand Mean	Standard Error	95% Confidence Interval	
		Lower Bound	Upper Bound
63.43	2.89	57.73	69.12

Table 9. Key Findings Summary from Technical Report 0040

Category	Summary
Grade Levels	Kindergarten through Grade 4
Assessment Focus	Word Reading Fluency and Passage Reading Fluency progress monitoring
Statistical Model	Rasch item-response modeling used to estimate word difficulty
Form Equivalence	Mean score comparisons and correlations across forms
Primary Outcome	Alternate forms demonstrated comparable difficulty suitable for repeated CBM progress monitoring

Reference

Alonzo, J. & Tindal, G. (2007). *The development of word and passage reading fluency measures in a progress monitoring assessment system (Technical Report # 0040*). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0043: The Development of Fifth-grade Passage Reading Fluency Measures for Use in a Progress Monitoring Assessment System (Alonzo & Tindal, 2008a).

Methods

This report describes the development of 20 alternate forms of a fifth-grade Passage Reading Fluency (PRF) measure, extending an existing K-4 progress monitoring assessment system. All 20 passages (~250 words each) were written by graduate students enrolled in College of Education courses in spring and summer 2007, targeting a 5.5 grade reading level.

Each passage underwent a four-stage review: (1) the lead author edited for grammatical correctness and grade-level appropriateness; (2) two graduate students divided passages into three paragraphs of even length and verified Flesch-Kincaid readability fell between 5.4 and 5.6 per paragraph; (3) an experienced fifth-grade teacher reviewed each passage for topic, wording, and style appropriateness; and (4) the lead author conducted a final review against test specifications.

Passage difficulty was analyzed using classical statistics. All 20 forms were administered to the same group of students over one week (5 forms per day), with a consistent single assessor per student, to control for tester variability. Repeated-measures correlations and mean comparisons were used to assess cross-form equivalence. Students read each passage aloud for 60 seconds; words read correctly per minute (WRCM) constituted the score.

Results

Pilot testing was conducted in February 2008 with 68-74 fifth-grade students per passage. Mean WRCM scores ranged from 134.96 (Gr5PRF_1, crossword puzzles) to 170.86 (Gr5PRF_20, career choices), indicating meaningful variation in passage difficulty that required correction. Inter-form correlations were uniformly strong across all 20 passages ($r = .76$-$.96$, all $p < .001$), demonstrating that the forms measured the same underlying skill consistently. Using the nine most similarly-difficulty passages as anchors, the remaining 11 passages were revised. The two easiest passages were simplified (e.g., Gr5PRF_1: 'vertically and horizontally' changed to 'up and down and across the paper'; Gr5PRF_2: 'sluggishly' changed to 'slowly'). Passages scoring higher than the anchor group were made more challenging (e.g., Gr5PRF_20 had sentences lengthened and names replaced with longer alternatives). Minor character name changes were used to adjust difficulty in several additional passages (e.g., Gr5PRF_8, Gr5PRF_13, Gr5PRF_14).

These revisions produced 20 comparable alternate forms. Final student copies used size 14 Verdana font; assessor copies included administration directions and a scored version with cumulative word counts. All forms were loaded to the EasyCBM website for web-based access.

Table 10. Example of Descriptive Statistics from Technical Report 0043

Table 2
Descriptive Statistics for Grade 5 Passage Reading Measures

Passage	Topic	*n*	*M*	*SD*
Gr5PRF_1	Crossword puzzles	74	134.96	32.239
Gr5PRF_2	Playing baseball	74	156.00	38.253
Gr5PRF_3	New baby brother	74	164.15	42.690
Gr5PRF_4	Moving to a new country	74	164.53	36.961
Gr5PRF_5	Zoo animals	74	162.54	39.541
Gr5PRF_6	Seating charts	74	161.82	39.808
Gr5PRF_7	Astronaut	74	165.95	34.937
Gr5PRF_8	Jen's crush	74	154.07	40.710
Gr5PRF_9	School pranks	74	168.99	38.482
Gr5PRF_10	Piano recital	74	165.89	43.019
Gr5PRF_11	Spider snack	70	160.26	41.254
Gr5PRF_12	Puppy of her own	70	165.74	42.009
Gr5PRF_13	Time machine	70	153.63	41.202
Gr5PRF_14	Real life hero	70	149.86	36.328
Gr5PRF_15	Spelling quiz	70	160.31	33.414
Gr5PRF_16	Tooth aches	68	164.57	38.255
Gr5PRF_17	Talent show	69	163.43	37.618
Gr5PRF_18	Book club	69	162.86	38.926
Gr5PRF_19	New bike	68	160.31	37.912
Gr5PRF_20	Career choices	69	170.86	38.611

Table 11. Sample Correlations from Technical Report 0043

Table 3
Correlations between Each of the Fifth-Grade Passage Reading Fluency Forms

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1																			
2	.89**	1																		
3	.89**	.94**	1																	
4	.86**	.92**	93**	1																
5	.88**	.93**	.94**	.93**	1															
6	.92**	.94**	.94**	.93**	.92**	1														
7	.89**	.92**	.92**	.89**	.91**	.93**	1													
8	.90**	.91**	.92**	.91**	.90**	.94**	.91**	1												
9	.89**	.94**	.93**	.92**	.93**	.95**	.92**	.96**	1											
10	.86**	.94**	.92**	.92**	.91**	.94**	.93**	.93**	.94**	1										
11	.87**	.93**	.94**	.93**	.93**	.94**	.92**	.93**	.92**	.94**	1									
12	.84**	.90**	.92**	.89**	.90**	.94**	.92**	.92**	.92**	.93**	.94**	1								
13	.82**	.89**	.91**	.88**	.89**	.93**	.85**	.86**	.86**	.88**	.85**	.87**	1							
14	.80**	.83**	.85**	.81**	.84**	.87**	.84**	.86**	.85**	.87**	.89**	.88**	.81**	1						
15	.80**	.85**	.85**	.83**	.84**	.85**	.83**	.84**	.84**	.86**	.85**	.81**	.76**	.79**	1					
16	.83**	.89**	.87**	.85**	.87**	.85**	.91**	.95**	.95**	.91**	.93**	.91**	.89**	.88**	.84**	1				
17	.87**	.93**	.92**	.91**	.89**	.95**	.87**	.87**	.88**	.92**	.85**	.85**	.81**	.79**	.78**	.88**	1			
18	.83**	.85**	.86**	.83**	.82**	.87**	.87**	.87**	.87**	.88**	.93**	.85**	.85**	.81**	.79**	.78**	.88**	1		
19	.85**	.91**	.92**	.89**	.92**	.92**	.89**	.90**	.93**	.93**	.95**	.94**	.90**	.87**	.87**	.90**	.93**	.83**	1	
20	.80**	.89**	.88**	.86**	.89**	.89**	.87**	.88**	.92**	.90**	.89**	.89**	.87**	.83**	.81**	.81**	.88**	.85**	.90**	1

Table 12. Example of Changes Made: Grade 5 Passage Reading Measures in Technical Report 0043

Table 4

Changes Made to Grade 5 Passage Reading Measures

Passage	Changes made to the passage
Gr5PRF_1	Made less challenging by changing Mr. Jacobson to Mr. Jacobs, 'assigned to groups' to 'put in groups', replacing 'vertically and horizontally' with 'up and down and across the paper', and changing several 'the students' to 'they'.
Gr5PRF_2	Made slightly less challenging by changing 'bannister' to 'railing', 'dreamt' to 'had dreamed', and 'sluggishly' to 'slowly'
Gr5PRF_3	No changes made
Gr5PRF_4	No changes made
Gr5PRF_5	No changes made
Gr5PRF_6	No changes made
Gr5PRF_7	No changes made
Gr5PRF_8	Made slightly easier by changing 'Jasmine' to 'Jen'
Gr5PRF_9	Made slightly more difficult by changing Jose to Mario, lengthening a few sentences.

Table 13. Key Findings Summary from Technical Report 0043

Category	Summary
Grade Level	Grade 5 only
Assessment Focus	Passage Reading Fluency progress monitoring
Measure Type	Passage Reading Fluency (PRF) — oral reading of connected narrative text
Alternate Forms	20 forms, each approximately 250 words, piloted February 2008
Readability Target	Flesch-Kincaid level 5.4–5.6 per paragraph (3 paragraphs per passage)
Statistical Model	Classical statistics: repeated-measures correlations and mean comparisons
Pilot Sample	68–74 students; each completed 5 passages per day over one week
Form Equivalence	Strong inter-form correlations ($r = .76–.96$, all $p < .001$); targeted revisions applied
Grand Mean Range	134.96 (lowest: crossword puzzles) to 170.86 (highest: career choices) WRC/min
Primary Outcome	20 comparable alternate forms for Grade 5 progress monitoring, loaded to EasyCBM®

Reference

Alonzo, J., & Tindal, G. (2008). *The development of fifth-grade passage reading fluency measures for use in a progress monitoring assessment system (Technical Report 0043)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0046: The Development of Middle School Passage Reading Fluency Measures for Use in a Progress Monitoring Assessment System (Alonzo, Park, et al., 2008).

This technical report describes the **development of middle school Passage Reading Fluency** measures for students in grades 6 through 8 as part of an expanded progress monitoring assessment system. The measures were designed to extend oral reading fluency assessment beyond the elementary grades and to provide educators with reliable tools for monitoring reading development during middle school.

Methods

This report describes the development of Passage Reading Fluency (PRF) progress monitoring measures for middle school students in Grades 6 through 8. The goal of the project was to create multiple alternate forms that could be administered repeatedly within a curriculum-based measurement (CBM) system while maintaining comparable difficulty across passages.

Researchers initially developed a large pool of reading passages aligned with middle school reading expectations. Passages were written by graduate students and reviewed by the lead author as well as panels of graduate students and classroom teachers to ensure appropriate readability, vocabulary level, and content quality. From this pool, twenty alternate PRF passages were developed for use in a progress monitoring system.

To evaluate form equivalence, the researchers conducted pilot testing with middle school students and analyzed passage difficulty using classical statistical procedures. The analyses focused on mean score comparisons and correlations among passage scores to determine whether alternate forms measured the same reading fluency construct. Passage difficulty estimates were used to identify a subset of passages with similar difficulty levels that could serve as anchors for the set of forms.

Results

Analyses of the pilot data indicated that the alternate passages showed acceptable levels of consistency for use in progress monitoring. Correlation analyses among passage scores demonstrated that student performance across passages was strongly related, suggesting that the passages measured the same underlying reading fluency construct.

Difficulty analyses identified nine passages that were highly similar in difficulty and could serve as a core reference set for the measure. The remaining passages were revised and adjusted so that their difficulty more closely matched the anchor passages. Revisions included modifications to vocabulary, sentence structure, and overall passage length to improve comparability.

After these revisions, the set of twenty passages demonstrated improved consistency in mean performance levels and correlations across forms. The results indicated that the passages functioned as comparable alternate forms suitable for repeated administration in a middle school reading progress monitoring system. Overall, the development process produced a technically adequate set of PRF measures that can be used to track reading fluency growth for students in Grades 6–8.

Table 14. Illustrative Descriptive Statistics from Technical Report 0046

Table 2
Descriptive Statistics for Grade 6 Passage Reading Measures

Passage	Topic	*n*	*M*	*SD*
Gr6PRF_1	Finding Money	55	185.70	35.18
Gr6PRF_2	Being Bilingual	55	166.20	38.20
Gr6PRF_3	Flying	55	171.96	34.19
Gr6PRF_4	Shopping Mall	55	186.20	39.59
Gr6PRF_5	New School	55	183.43	37.14
Gr6PRF_6	Cousin Envy	55	164.85	31.40
Gr6PRF_7	Pet Rabbit	55	184.26	32.95
Gr6PRF_8	Lost Keys	55	185.09	40.14
Gr6PRF_9	Singing	55	177.91	33.13
Gr6PRF_10	New Video Game	55	195.72	32.60
Gr6PRF_11	Embarrassment	55	186.17	38.02
Gr6PRF_12	Grandma	55	176.28	33.22
Gr6PRF_13	Sign Language	55	189.20	35.97
Gr6PRF_14	Lost Bike	55	199.74	38.60
Gr6PRF_15	Budget Woes	54	193.92	36.09
Gr6PRF_16	Asking for a Date	54	183.75	37.98
Gr6PRF_17	Going to College	54	203.00	44.55
Gr6PRF_18	Group Project	54	177.62	36.50
Gr6PRF_19	Roller Coasters	54	184.00	37.65
Gr6PRF_20	Saturday Morning	53	201.58	45.77

Table 15. Typical Correlation from Technical Report 0046

Table 3
Correlations between Each of the Sixth-Grade Passage Reading Fluency Forms

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1																			
2	.90**	1																		
3	.88**	.90**	1																	
4	.93**	.92**	.92**	1																
5	.93**	.91**	.91**	.94**	1															
6	.77**	.76**	.76**	.82**	.80**	1														
7	.83**	.82**	.78**	.85**	.87**	.79**	1													
8	.90**	.89**	.88**	.89**	.90**	.80**	.85**	1												
9	.88**	.86**	.85**	.86**	.85**	.76**	.81**	.88**	1											
10	.91**	.88**	.88**	.93**	.90**	.80**	.83**	.89**	.86**	1										
11	.90**	.89**	.90**	91**	.90**	.81**	.87**	.90**	.90**	.92**	1									
12	.91**	.88**	.91**	.92**	.90**	.81**	.87**	.89**	.91**	.91**	.93**	1								
13	.88**	.83**	.87**	.87**	.88**	.75**	.85**	.88**	.90**	.86**	.92**	.92**	1							
14	.89**	.84**	.86**	.89**	.89**	.78**	.83**	.90**	.88**	.91**	.89**	.88**	.90**	1						
15	.88**	.87**	.86**	.89**	.88**	.82**	.86**	.89**	.94**	.90**	.93**	.92**	.93**	.93**	1					
16	.87**	.84**	.83**	.88**	.86**	.80**	.84**	.88**	.90**	.90**	.88**	.92**	.89**	.88**	.92**	1				
17	.89**	.86**	.83**	.90**	.89**	.84**	.84**	.88**	.88**	.92**	.89**	.93**	.88**	.89**	.91**	.92**	1			
18	.85**	.80**	.84**	.87**	.86**	.80**	.85**	.86**	.86**	.87**	.89**	.92**	.90**	.87**	.91**	.90**	.89**	1		
19	.89**	.87**	.87**	.89**	.91**	.84**	.87**	.90**	.88**	.89**	.89**	.91**	.89**	.89**	.91**	.92**	.91**	.92**	1	
20	.89**	.84**	.86**	.87**	.88**	.81**	.85**	.89**	.88**	.91**	.90**	.90**	.91**	.93**	.92**	.88**	.92**	.88**	.90**	1

Table 16. Key Findings Summary from Technical Report 0046

Category	Summary
Grade Levels	Grades 6–8
Assessment Focus	Passage Reading Fluency
Statistical Model	Classical Test Theory
Forms Developed	17–20 forms per grade
Primary Outcome	Reliable fluency measures for middle school progress monitoring

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2008). *The development of middle school passage reading fluency measures for use in a progress monitoring assessment system (Technical Report 0046)*. Eugene, ORL University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0041: Examining the Technical Adequacy of Reading Comprehension Measures in a Progress Monitoring Assessment System (Alonzo et al., 2007).

This technical report evaluates the reliability and performance of reading progress monitoring measures, focusing on methodological design and psychometric outcomes. The study aimed to determine whether the measures provide reliable and valid information for educational decision-making.

Methods

The sample consisted of students across multiple grade levels, including both general education students and those receiving targeted instructional support. Assessments were administered in standardized settings by trained examiners. Data collection protocols emphasized consistent administration procedures, scoring accuracy, and repeated measurement to support reliability analyses.

Analytical methods included classical test theory statistics such as internal consistency estimates, as well as item response theory modeling to examine item-level performance. Item difficulty, discrimination, and fit indices were evaluated. Additional analyses examined score distributions and stability to assess suitability for progress monitoring.

Results

Findings indicated that the reading measures demonstrated moderate to strong reliability across grade levels. Item difficulty values were generally centered within acceptable ranges, indicating appropriate calibration for the intended populations. Most items exhibited acceptable fit statistics, with few extreme or poorly functioning items identified. Overall, the results support the use of these measures for monitoring student progress and informing instructional decisions.

Table 17. Example of Item Statistics from Technical Report 0041

Table 4

Item Statistics, Entry Order, Gr3MC1

Item Number	Raw Score	Count	Measure	Model Standard Error	Mean Square Outfit
1	78	88	-1.29	.35	.89
2	54	87	0.48	.24	.96
3	75	88	-0.96	.31	.80
4	45	87	0.96	.23	1.17
5	73	88	-0.77	.30	1.10
6	62	88	0.03	.25	.94
7	67	89	-0.24	.26	.91
8	75	88	-0.95	.31	.79
9	79	89	-1.30	.35	.98
10	27	89	1.98	.25	1.17
11	69	89	-0.39	.27	.71
12	54	89	0.53	.23	1.10
13	69	88	-0.39	.27	.78
14	53	89	0.54	.23	1.22
15	75	88	-0.87	.31	.55
16	59	87	0.22	.24	.86
17	59	89	0.17	.25	1.14
18	33	89	1.64	.23	1.16
19	45	89	1.01	.23	1.05
20	69	89	-0.39	.27	.82

Table 18. Typical Distractor Analysis from Technical Report 0041

Table 5
Distractor Analysis, Gr3MC1, Continued

Entry #	Data Code	Score Value	Count	%	Average Measure	S.E. Mean
11	B	0	5	6	0.08	.28
	C	0	15	17	0.30	.17
	A	1	69	78	1.33	.12
	Missing	**				
12	B	0	4	4	0.57	.48
	C	0	31	35	0.84	.11
	A	1	54	61	1.27	.16
	Missing	**				
13	C	0	10	11	0.29	.27
	A	0	10	11	0.33	.18
	B	1	69	78	1.31	.12
	Missing	**				
14	B	0	15	17	0.61	.17
	C	0	20	23	1.07	.20
	A	1	53	60	1.20	.15
	Missing	**	1	1	2.66	
15	A	0	8	9	-0.20	.20
	B	0	6	7	0.13	.26
	C	1	75	84	1.30	.11
	Missing	**				
16	A	0	25	28	0.44	.16
	C	0	4	5	0.45	.44
	B	1	59	67	1.41	.12
	Missing	**	1	1	0.75	
17	B	0	8	9	0.72	.27
	A	0	20	23	0.85	.15
	C	1	59	68	1.22	.15
	Missing	**	2	2	0.95	.21
18	C	0	22	25	0.80	.16
	A	0	34	38	0.92	.13
	B	1	33	37	1.45	.22
	Missing	**				
19	C	0	18	20	0.69	.16
	B	0	26	29	0.81	.17
	A	1	45	51	1.41	.16
	Missing	**				
20	A	0	13	15	0.30	.20
	B	0	7	8	0.48	.33
	C	1	69	78	1.30	.12
	Missing	**				

Summary of Technical Report 0807: Examining the Technical Adequacy of Fifth Grade Reading Comprehension Measures in a Progress Monitoring Assessment System (Alonzo & Tindal, 2008b).

This technical report presents the development and evaluation of reading progress monitoring measures, with emphasis on methodological design and empirical performance. The study examined whether the measures yield reliable, instructionally meaningful data for diverse student populations.

Methods

Participants included students sampled from 5th grade, representing both general education populations and students receiving additional instructional support. Assessments were administered under standardized conditions by trained examiners. Data collection procedures emphasized consistency in administration, scoring, and timing to support repeated-measures analyses and ensure score comparability. Analyses incorporated both classical test theory and item response theory approaches. Internal consistency estimates were calculated to evaluate reliability. Item-level analyses examined difficulty, discrimination, and fit statistics to assess measurement quality. Additional analyses explored score distributions and stability across administrations to evaluate progress monitoring sensitivity.

Results

Results indicated that the reading measures demonstrated moderate to strong reliability across grade levels. Item difficulty values were generally centered within acceptable ranges, supporting differentiation among students with varying skill levels. Most items exhibited acceptable fit statistics, with few items identified as poorly functioning. Overall, the findings support the technical adequacy of the measures for monitoring student progress and informing instructional decision-making.

Table 19. Example of Item Statistics from Technical Report 0807

Table 3
Item Statistics, Entry Order, Gr5MC1

Item Number	Raw Score	Count	Measure	Model Standard Error	Mean Square Outfit
1	155	176	-2.27	0.26	2.64
2	105	176	0.01	0.19	0.61
3	106	176	-0.03	0.20	0.52
4	103	176	0.09	0.19	0.58
5	145	176	-1.70	0.23	2.96
6	143	176	-1.59	0.22	1.20
7	84	176	0.78	0.19	0.69
8	111	176	-0.22	0.20	0.46
9	129	176	-0.95	0.21	1.09
10	101	175	0.15	0.19	0.71
11	108	175	-0.11	0.20	0.56
12	43	174	2.27	0.20	1.73
13	75	174	1.10	0.19	0.98
14	113	174	-0.32	0.20	0.64
15	65	174	1.46	0.19	1.47
16	99	174	0.22	0.19	0.98
17	108	174	-0.13	0.20	0.50
18	114	174	-0.36	0.20	0.73
19	93	174	0.44	0.19	0.95
20	73	174	1.17	0.19	1.72

Table 20. Illustrative Distractor Analysis from Technical Report 0807

Table 4
Distractor Analysis, Gr5MC1

Entry #	Data Code	Score Value	Count	%	Average Measure	S.E. Mean
1	A	0	12	7	-0.67	0.30
	C	0	9	5	0.75	0.41
	B	1	155	88	0.81	0.15
	Missing	**				
2	A	0	51	29	-1.25	0.13
	C	0	20	11	0.03	0.23
	B	1	105	60	1.79	0.14
	Missing	**				
3	C	0	55	31	-1.12	0.13
	B	0	15	9	-0.35	0.26
	A	1	106	60	1.81	0.14
	Missing	**				
4	A	0	59	34	-1.04	0.14
	C	0	14	8	-0.09	0.27
	B	1	103	59	1.82	0.14
	Missing	**				
5	B	0	4	2	-0.44	0.51
	A	0	27	15	0.50	0.23
	C	1	145	82	0.78	0.16
	Missing	**				
6		0	12	7	-0.54	0.14
		0	21	12	-0.41	0.27
		1	143	81	0.98	0.16
	Missing	**				
7	B	0	60	34	-0.97	0.14
	A	0	32	18	0.37	0.23
	C	1	84	48	2.04	0.15
	Missing	**				
8	C	0	47	27	-1.28	0.11
	B	0	18	10	-0.55	0.23
	A	1	111	63	1.76	0.14
	Missing	**				
9	B	0	22	13	-1.21	0.25
	C	0	25	14	-0.51	0.24
	A	1	129	73	1.27	0.15
	Missing	**				
10	B	0	45	26	-1.11	0.15
	A	0	29	16	-0.17	0.22
	C	1	101	57	1.8	0.15
	Missing	**	1	1	-1.51	0

Table 21. Summary of Key Findings from Technical Report 0807

Category	Summary
Sample	The sample consisted of fifth-grade students drawn from participating schools, with sample sizes for individual analyses generally ranging from approximately 48 to 50 students per form. Students represented typical general education classrooms used to evaluate progress monitoring measures.
Assessment Forms	Multiple fifth grade reading comprehension forms were developed and analyzed. Forms were designed to be equivalent in structure and difficulty to support repeated administration for screening and progress monitoring.
Analysis Method	Statistical analyses included classical test theory methods and Rasch modeling. Analyses focused on form equivalence, within-subjects comparisons across forms, and item-level performance.
Items Analyzed	Reading comprehension items embedded within multiple assessment forms were analyzed. Items were designed to reflect grade-level comprehension skills and varied in difficulty to capture a range of student ability.
Problematic Items	A small number of items exhibited weaker performance, including reduced variability or fit concerns. These items were identified for potential revision or further investigation in subsequent studies.
Item Fit	Rasch and descriptive analyses indicated that most items demonstrated acceptable fit and functioned as intended. Item difficulty was generally centered within recommended ranges, supporting differentiation among students.
Overall Conclusion	Results supported the technical adequacy of the fifth grade reading comprehension measures for use in a progress monitoring assessment system. While reliability was acceptable for screening and instructional use, the report notes that larger samples and additional analyses could further strengthen measurement precision.

Reference

Alonzo, J., & Tindal, G. (2008). *Examining the technical adequacy of fifth grade reading comprehension measures in a progress monitoring assessment system (Technical Report 0807).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0808: Examining the Technical Adequacy of Second Grade Reading Comprehension Measures in a Progress Monitoring Assessment System (Alonzo, Liu, et al., 2008).

Methods

This report describes the development and pilot testing of second grade reading comprehension measures intended for use in a progress monitoring assessment system. The goal was to create multiple equivalent forms that could be administered repeatedly to track student growth.

Reading passages were written by elementary and secondary teachers and graduate students. Comprehension questions were then developed by an assessment design expert. Items were organized according to cognitive attributes (e.g., types of comprehension processes) and estimated difficulty levels to ensure coverage of different comprehension skills.

A total of 20 alternate test forms were created. The forms were piloted using a common person / common item design. Groups of students each completed five forms, with overlapping forms across groups to allow comparisons across administrations. This design allowed the researchers to evaluate item performance and establish relationships among the forms. Content and bias reviews were conducted to ensure that passages and questions were appropriate for second-grade students and free from unnecessary linguistic or cultural bias. Item-level analyses were then conducted using Rasch modeling, which estimated item difficulty based on student performance and allowed the researchers to examine how well individual items functioned within the measurement model.

Results

0808 provided difficulty estimates for each comprehension item, allowing researchers to evaluate how items performed relative to student ability levels. The analysis supported the creation of a calibrated pool of items across the 20 forms.

Overall, the results indicated that the items covered a range of difficulty levels appropriate for second-grade readers. The Rasch modeling also allowed the researchers to examine the comparability of the different test forms. Items that appeared to function less effectively were retained for additional examination in future studies with larger samples. The pilot results demonstrated that the forms could be used to support a progress monitoring system for second grade reading comprehension, with item-level information helping guide future refinement of the measures.

Table 22. Sample Item Statistics from Technical Report 0808

Table 3

Item Statistics, Entry Order, Gr2MC1

Item Number	Raw Score	Count	Measure	Model Standard Error	Mean Square Outfit
1	100	118	-1.49	0.27	0.76
2	78	118	-0.28	0.21	0.75
3	71	118	0.03	0.21	1.09
4	53	118	0.77	0.20	1.16
5	68	118	0.15	0.21	1.00
6	59	118	0.52	0.20	1.19
7	79	118	-0.32	0.21	0.75
8	71	118	0.03	0.21	0.88
9	61	118	0.44	0.20	1.27
10	72	117	-0.04	0.21	1.02
11	88	117	-0.80	0.23	0.75
12	47	117	1.00	0.21	1.10

Table 23. Illustrative Distractor Analysis from Technical Report 0808

Table 4
Distractor Analysis, Gr2MC1

Entry #	Data Code	Score Value	Count	%	Average Measure	S.E. Mean
1	B	0	5	4	-0.60	.52
	C	0	13	11	-0.40	.21
	A	1	100	85	0.87	.13
	Missing	**				
2	B	0	14	12	-0.37	.25
	A	0	26	22	-0.21	.14
	C	1	78	66	1.15	.15
	Missing	**				
3	C	0	20	17	-0.09	.19
	B	0	27	23	0.31	.24
	A	1	71	60	1.02	.16
	Missing	**				
4	A	0	30	25	-0.02	.18
	C	0	35	30	0.56	.14
	B	1	53	45	1.13	.21
	Missing	**				
5	A	0	25	21	-0.25	.13
	B	0	25	21	0.37	.30
	C	1	68	58	1.11	.15
	Missing	**				
6	C	0	18	15	-0.44	.19
	A	0	41	35	0.66	.19
	B	1	59	50	1.01	.17
	Missing	**				
7	C	0	10	8	-0.49	.32
	B	0	29	25	-0.23	.14
	A	1	79	67	1.15	.14
	Missing	**				
8	A	0	21	18	-0.22	.17
	B	0	26	22	0.04	.24
	C	1	71	60	1.16	.15
	Missing	**				
9	A	0	26	22	0.27	.15
	C	0	31	26	0.45	.26
	B	1	61	52	0.95	.18
	Missing	**				
10	B	0	18	15	-0.06	.22
	C	0	27	23	0.18	.16
	A	1	72	62	1.04	.17
	Missing	**	1	1	0.25	

Table 24. Summary of Key Findings from Technical Report 0808

Category	Summary
Purpose	Develop and pilot second-grade reading comprehension progress monitoring measures.
Item Development	Teachers created passages; comprehension questions written and grouped by cognitive skill and difficulty.
Design	20 alternate forms piloted using a common person / common item design.
Analysis Method	Rasch modeling used to estimate item difficulty and examine item functioning.
Key Finding	Items covered appropriate difficulty levels and supported development of comparable progress monitoring forms.

Reference

Alonzo, J., Liu, K., & Tindal, G. (2008). *Examining the technical adequacy of second-grade reading comprehension measures in a progress monitoring assessment system (Technical Report No. 0808)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Table 25. Summary of Key Findings from Technical Report 0041, 0807, and 0808

The methods and results for TR 0041 are virtually identical for all the Categories listed below in two other Technical Reports: Alonzo, Liu, et al. (2008) and Alonzo and Tindal (2008b).

Category	Summary
Sample	The sample included a range of reading ability levels suitable for evaluating item performance and measurement adequacy.
Assessment Forms	Twenty forms were designed to be equivalent in structure and difficulty to support repeated administration within a progress monitoring framework.
Analysis Method	Item Response Theory (IRT) was used to report fit statistics, difficulty parameters, and discrimination indices as indices of technical adequacy.
Items Analyzed	Items spanned a range of difficulty levels intended to capture variation in student reading ability.
Item Fit	Most items exhibited acceptable fit statistics and functioned as intended.
Overall Conclusion	Results supported the technical adequacy of the reading comprehension measures for progress monitoring purposes. Most items functioned well, with only minor revisions needed to improve measurement precision and instructional utility.

Reference

Alonzo, J., Liu, K., & Tindal, G. (2007). *Examining the technical adequacy of reading comprehension measures in a progress monitoring assessment system (Technical Report 0041)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, Liu, et al. (2008). *Examining the technical adequacy of second-grade reading comprehension measures in a progress monitoring assessment system (Technical Report # 0808)*. Eugene, OR, Behavioral Research and Teaching, University of Oregon.

Alonzo and Tindal (2008b). *Examining the technical adequacy of fifth-grade reading comprehension measures in a progress monitoring assessment system (Technical Report # 0807).* Eugene, OR, Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1102: The Development and Technical Adequacy of Seventh-Grade Reading Comprehension Measures in a Progress Monitoring Assessment (Park et al., 2011).

Methods

This report describes the development and pilot testing of seventh grade reading comprehension measures designed for use in the easyCBM® progress monitoring and screening system. The goal was to create a set of technically sound comprehension assessments appropriate for middle school students and suitable for repeated administration.

Each assessment form consisted of an original narrative passage approximately 1,600–1,900 words long followed by 20 multiple-choice comprehension questions. The questions were designed to measure different levels of comprehension, including literal, inferential, and evaluative understanding. Each item included three response options: a correct answer, a near-correct distractor, and a clearly incorrect distractor. This structure was intended to capture gradations in student understanding while maintaining efficient scoring. Passages and items were developed specifically for the assessment system and reviewed to ensure appropriate reading level, clarity, and alignment with comprehension constructs. The measures were piloted with seventh-grade students using the easyCBM® online platform. Data from the pilot administrations were used to evaluate item performance and the comparability of the assessment forms.

To evaluate item functioning and test structure, the researchers applied Rasch measurement modeling. Rasch analysis allowed the authors to estimate item difficulty and examine whether individual items functioned consistently within the measurement model. This approach also allowed the researchers to determine whether items covered an appropriate range of difficulty for the target population and whether the forms could support progress monitoring.

Results

The Rasch analyses produced difficulty estimates for each comprehension item and allowed the researchers to evaluate the functioning of the assessment forms. Overall, the items demonstrated an appropriate range of difficulty for seventh-grade students, suggesting that the measures could distinguish among different levels of reading comprehension ability.

The analyses also indicated that the test forms functioned comparably and that most items fit the Rasch model adequately. These results supported the technical adequacy of the assessment forms for use in a progress monitoring system. The calibrated item information provides a foundation for constructing multiple assessment forms capable of tracking student performance over time. Overall, the findings indicate that the seventh-grade comprehension measures developed for easyCBM provide a structured and psychometrically sound approach to assessing reading comprehension and monitoring student growth in middle school reading skills.

Table 26. Example Item Statistics from Technical Report 1102

Item Statistics, Entry Order, Gr7MC1.

Item Number	Raw Score	Count	Measure	Model Standard Error	Mean Square Outfit
1	61	64	-2.55	0.60	0.32
2	42	64	0.07	0.28	1.02
3	53	64	-1.00	0.35	0.80
4	10	64	2.76	0.36	1.25
5	58	64	-1.76	0.44	0.64
6	41	64	0.15	0.28	0.91
7	16	64	2.10	0.31	1.75
8	54	64	-1.12	0.36	0.78
9	43	64	-0.02	0.29	0.72
10	30	64	0.97	0.27	1.06
11	62	64	-2.99	0.73	0.28
12	51	64	-0.76	0.33	0.96
13	33	64	0.75	0.27	1.18
14	26	64	1.27	0.28	0.88
15	33	64	0.75	0.27	1.17
16	14	64	2.30	0.32	1.66
17	38	64	0.38	0.28	1.19
18	48	64	-0.46	0.31	0.93
19	50	64	-0.66	0.32	0.83
20	45	64	-0.19	0.29	1.03

Table 27. Key Findings Summary from Technical Report 1102

Category	Summary
Purpose	Develop and evaluate seventh grade reading comprehension measures for the easyCBM® progress monitoring system.
Assessment Design	Passages of 1,600–1,900 words with 20 multiple-choice comprehension items assessing literal, inferential, and evaluative understanding.
Data Collection	Pilot testing conducted with seventh-grade students using the easyCBM® online platform.
Analysis Method	Rasch modeling used to estimate item difficulty and evaluate item functioning.
Key Results	Items showed an appropriate range of difficulty and acceptable model fit, supporting use for comprehension assessment and monitoring.

Reference

Park, B. J., Alonzo, J., & Tindal, G. (2011). *The development and technical adequacy of seventh-grade reading comprehension measures in a progress monitoring assessment (Technical Report # 1102)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1209: The Development of CBM Vocabulary Measures: Grade 2 (Alonzo, Anderson, et al., 2012a).

Methods

This technical report describes the development and piloting of curriculum-based vocabulary measures designed for second-grade students within a response-to-intervention (RTI) framework. Approximately 1,200 students from schools across the United States participated, with data collected during spring benchmark administrations via an online assessment platform. Items were presented individually on computer screens with responses recorded automatically.

Item development was guided by principles of universal design for assessment to ensure accessibility for diverse learners, including students with disabilities and English language learners. Items were written to target specific grade-level vocabulary standards and underwent extensive review prior to piloting. Anchor items were embedded across forms to support calibration. Statistical analyses employed a one-parameter logistic (1PL) Rasch model using Winsteps software. Item difficulty estimates, fit statistics, and standard errors were examined for all items. Overfitting items were retained when distractor analyses confirmed appropriate functioning; no underfitting items were identified.

Results

Rasch analysis was used to evaluate the psychometric properties of all items on the Gr2Vocab1 form. The mean item measure for this form was -1.85 logits, indicating that, on average, items were relatively easy for the sampled population of second-grade students. Item difficulty (measure) values ranged from -3.91 (Item 1) to -0.99 (Item 8), reflecting a moderate spread of difficulty across the form. All items were on the easier end of the logit scale, consistent with expectations for a Grade 2 vocabulary assessment administered in the spring. Item response counts ranged from 77 (Item 10) to 95 (Item 11), indicating adequate sample sizes for estimation across all items. Standard errors of measurement were generally small, ranging from 0.39 to 1.84, with the largest error associated with Item 1 (measure = -3.91), which was the most extreme in difficulty and had the lowest response count (n = 78). Items with more central difficulty estimates tended to have smaller standard errors, consistent with Rasch modeling expectations.

Mean Square Outfit (MSQ Outfit) values were examined to evaluate model fit. Acceptable fit is typically defined as values between 0.50 and 1.50. Most items fell within this range. Item 2 demonstrated notable underfit with an MSQ Outfit of 0.07, and Items 9, 10, and 11 also showed low outfit values (0.16, 0.16, and 0.19 respectively), suggesting these items may have been overly predictable for students of varying ability levels. Despite these low outfit values, distractor analyses confirmed that all items were functioning appropriately and no items were flagged for removal. Item 3 had the highest MSQ Outfit value of 1.45, approaching but remaining within the acceptable threshold.

Discrimination (Disc) values ranged from 0.94 to 1.23, with all items exceeding the minimum acceptable threshold, indicating that each item reliably differentiated between students of higher and lower vocabulary ability. Items targeted a range of grade-level standards, including EL.03, EL.04, EL.05, and EL.06 reading standards related to vocabulary, contextual meaning, figurative language, synonyms, antonyms, homographs, and multiple-meaning words. The breadth of standards covered supports the content validity of the form. Overall, results supported the retention of all 12 items on the Gr2Vocab1 form without revision. These item-level findings were consistent across the other piloted forms, ultimately supporting the creation of ten progress monitoring forms and three benchmark forms with comparable difficulty levels and strong technical adequacy for use in instructional decision-making.

Table 28. Example of Summary Statistics from Technical Report 1209

Table 1a
Item Information and Statistics, Gr2Vocab1 (Mean Measure = -1.85)

Item	Item Name	Standard	Measure	Count	Error	MSQ Outfit	Disc
1	Vocab_Item_55410	EL.05.RE.11 Determine meanings of words using contextual and structural clues.	-3.91	78	1.84	1.00	1.00
2	Vocab_Item_65641	EL.06.RE.12 Interpret figurative language, including similes, metaphors, and words with multiple meanings.	-3.00	93	1.04	0.07	1.15
3	Vocab_Item_35734	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-1.39	94	0.46	1.45	1.07
4	Vocab_Item_45641	EL.04.RE.11 Determine meanings of words using contextual and structural clues.	-1.37	89	0.50	0.21	1.23
5	Vocab_Item_45736	EL.04.RE.13 Apply knowledge of synonyms, antonyms, homographs, and idioms to determine the meaning of words and phrases.	-1.23	87	0.49	1.18	0.99
6	Vocab_Item_45644	EL.04.RE.12 Distinguish and interpret words with multiple meanings (i.e., quarter) by using context clues.	-1.19	89	0.42	1.21	1.06
7	Vocab_Item_35274	EL.03.RE.14 Determine the meanings of words using knowledge of antonyms, synonyms, homophones, and homographs.	-1.08	94	0.39	0.91	0.94
8	Vocab_Item_35151	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-0.99	93	0.45	1.05	1.08
9	Vocab_Item_45645	EL.04.RE.12 Distinguish and interpret words with multiple meanings (i.e., quarter) by using context clues.	-2.41	89	0.74	0.16	1.13
10	Vocab_Item_55650	EL.05.RE.13 Determine the meanings of figurative expressions, such as those in similes and metaphors.	-2.22	77	0.73	0.16	1.16
11	Vocab_Item_35589	EL.03.RE.14 Determine the meanings of words using knowledge of antonyms, synonyms, homophones, and homographs.	-1.94	95	0.55	0.19	1.20
12	Vocab_Item_35109	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-1.50	94	0.50	0.26	1.16

Table 29. Key Findings Summary from Technical Report 1209

Category	Summary
Grade Level	Grade 2
Assessment Focus	Vocabulary
Statistical Model	1PL Rasch
Forms Developed	10 progress monitoring and 3 benchmark forms
Primary Outcome	Reliable vocabulary CBMs for RTI use

Reference

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 2 (Technical Report 1209).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1210: The Development of CBM Vocabulary Measures: Grade 3 (Alonzo, Anderson, et al., 2012b).

Methods

Researchers at the University of Oregon developed vocabulary assessments for grades 2–8 as part of the easyCBM® online progress monitoring system. Item development followed Universal Design for Assessment (UDA) principles, prioritizing accessibility for English language learners and students with disabilities. Eight item writers with backgrounds in education and special education created approximately 800 items per grade level, guided by Oregon State Vocabulary Standards (EL.03.RE.14 and EL.03.RE.15). Items were multiple-choice with three answer options, designed to minimize cognitive load and linguistic complexity. After a thorough review by six researchers from June–August 2009, approximately 5,600 items were prepared for piloting. In spring 2011, roughly 1,200 third-grade students from schools across the United States completed pilot assessments delivered online using a common-item design. Each student was administered anchor items plus randomly selected grade-level items. Results were analyzed using a one-parameter logistic (1PL) Rasch model via Winsteps software, with Mean Square Outfit values (acceptable range: 0.50–1.50) and distractor analysis used to evaluate item quality.

Results

Although no items were originally written specifically for grade 3, extremely easy items identified during piloting were repurposed for this grade. Of the items analyzed, 107 were over-fit (Mean Square Outfit: 0.08–0.49); 90 of these were retained based on satisfactory distractor functioning. An additional 26 under-fit items (Mean Square Outfit: 1.52–3.13) were all removed due to poor distractor performance. In total, 43 items were excluded because higher-ability students did not consistently select the correct answer. The retained items were organized into 13 third-grade vocabulary measures: 10 progress monitoring forms and 3 benchmark screener forms, each containing 20 items. All forms were calibrated to comparable difficulty, with mean measures ranging narrowly from −0.08 to −0.03, confirming strong form equivalence suitable for reliable progress monitoring.

Table 30. Example Item Information and Statistics from Technical Report 1210

Table 1a

Item Information and Statistics, Gr3Vocab1 (Mean Measure = -0.07)

Item	Item Name	Standard	Measure	Count	Error	MSQ Outfit	Disc
1	Vocab_Item_35781	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-1.55	94	0.48	0.43	1.11
2	Vocab_Item_35734	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-1.39	94	0.46	1.45	1.07
3	Vocab_Item_35692	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-0.76	95	0.37	0.55	1.17
4	Vocab_Item_35014	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-0.73	94	0.36	0.70	1.12
5	Vocab_Item_35334	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-0.03	95	0.29	0.57	1.33
6	Vocab_Item_35415	EL.03.RE.14 Determine the meanings of words using knowledge of antonyms, synonyms, homophones, and homographs.	0.01	94	0.32	0.90	0.98
7	Vocab_Item_35776	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	0.80	94	0.27	0.91	1.10
8	Vocab_Item_35868	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	0.93	95	0.26	0.74	1.34
9	Vocab_Item_35508	EL.03.RE.15 Use sentence and word context to find the meaning of unknown words.	-1.21	94	0.44	0.25	1.24
10	Vocab_Item_35859	EL.03.RE.14 Determine the meanings of words using knowledge of antonyms, synonyms, homophones, and homographs.	-0.98	94	0.40	0.44	1.18

Table 31. Typical Distractor Analysis Results from Technical Report 1210

Table 2
Distractor Analysis Results

Item	Answer choice	Score	# of students	Average measure	Standard error	Mean square outfit
v35001	A	1	3414	-4.18	0.02	1.97
v35001	B	0	666	-4.60	0.05	1.39
v35001	C	0	147	-5.06	0.12	1.07
v35002	A	1	88	-3.78	0.03	0.89
v35002	B	0	3	-4.06	0.10	0.91
v35002	C	0	3	-4.17	0.14	0.82
v35003	A	1	87	-3.89	0.04	1.02
v35003	B	0	2	-4.64	0.54	0.58
v35003	C	0	5	-4.51	0.23	0.63
v35004	A	1	80	-3.83	0.04	0.96
v35004	B	0	9	-4.31	0.17	0.77
v35004	C	0	4	-4.59	0.22	0.57
v35005	A	1	62	-3.83	0.04	0.95
v35005	B	0	7	-4.40	0.34	0.83
v35005	C	0	24	-4.15	0.09	0.88
v35006	A	1	92	-3.85	0.04	0.99
v35006	B	0	1	-8.69	0.00	0.00
v35006	C	0	1	-4.89	0.00	0.39
v35007	A	1	90	-3.89	0.04	1.02
v35007	B	0	3	-4.73	0.45	0.57
v35007	C	0	2	-4.57	0.61	0.64
v35008	A	1	80	-3.82	0.04	0.95
v35008	B	0	7	-4.87	0.31	0.50
v35008	C	0	6	-4.65	0.15	0.53
v35009	A	1	69	-3.87	0.05	1.04
v35009	B	0	19	-4.33	0.16	0.82
v35009	C	0	6	-4.67	0.17	0.53
v35010	A	1	90	-3.95	0.05	1.17
v35010	B	0	4	-5.37	0.74	0.43
v35010	C	0	1	-5.17	0.00	0.30
v35011	A	1	62	-3.84	0.04	0.96
v35011	B	0	8	-4.82	0.23	0.50
v35011	C	0	24	-4.05	0.10	0.99
v35012	A	1	86	-3.82	0.04	1.02
v35012	B	0	4	-5.93	0.62	0.23
v35012	C	0	4	-4.75	0.43	0.56
v35013	A	1	81	-3.80	0.03	0.89
v35013	B	0	6	-4.20	0.17	0.83
v35013	C	0	7	-4.37	0.19	0.74
v35014	A	1	81	-3.85	0.03	0.95
v35014	B	0	6	-5.89	0.75	0.51
v35014	C	0	7	-4.69	0.32	0.62

Table 32. Key Findings Summary from Technical Report 1210

Category	Summary
Grade Level	Grade 3
Assessment Focus	Vocabulary curriculum-based measurement
Statistical Model	1PL Rasch
Forms Developed	10 progress monitoring and 3 benchmark forms
Primary Outcome	Technically adequate vocabulary measures for RTI use

Reference

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 3 (Technical Report 1210).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1211: The Development of CBM Vocabulary Measures: Grade 4 (Alonzo, Anderson, et al., 2012c).

Methods

This study developed curriculum-based measurement (CBM) vocabulary assessments for students in grades 2–8, delivered via the easyCBM® online platform. Item development was grounded in the Oregon State Standards for Vocabulary, with eight item writers recruited from Oregon and California — most with Special Education backgrounds — producing approximately 800 items per grade level between October 2008 and May 2009. Writers were instructed to reduce cognitive complexity, use simple and direct language, and cover a range of student ability. Items used a multiple-choice format with three answer options, displayed one at a time on screen with the target word bolded in context. A team of six researchers reviewed all items for clarity, alignment, and formatting from June through August 2009, yielding approximately 5,600 items for piloting. The pilot was conducted in spring 2011 with approximately 1,200 students per grade across U.S. schools, using a common-item design with five anchor items to place all items on a shared scale. Data were analyzed using a one-parameter logistic (1PL) Rasch model via Winsteps software, evaluating each item's difficulty (Measure), Standard Error, and Mean Square Outfit statistic.

Results

Although no items were written specifically for Grade 4, extremely easy items from adjacent grades were identified through piloting and organized into a Grade 4 item bank. Of the items analyzed, 122 were over-fit (Mean Square Outfit: 0.11–0.49; average 0.34) but retained because distractor analysis confirmed appropriate functioning: higher-ability students consistently selected the correct answer. Twenty-six under-fit items (Mean Square Outfit: 1.52–8.78; average 2.00) were removed because distractor analysis revealed reversed performance patterns. One additional item was removed because neither distractor was chosen by any student, resulting in 27 total items excluded. The retained items were used to construct 10 progress monitoring forms and 3 benchmark screener forms (13 forms total), each containing 20 items. Mean item difficulty (Measure) ranged narrowly from −0.09 to −0.05 across all forms, confirming comparable difficulty. Test information functions and characteristic curves confirmed that all forms performed consistently across the ability spectrum.

Table 33. Sample Item Information and Statistics from Technical Report 1211

Table 1a
Item Information and Statistics, Gr4Vocab1 (Mean Measure = -0.08)

Item	Item Name	Standard	Measure	Count	Error	MSQ Outfit	Disc
1	Vocab_Item_45181	EL.04.RE.11 Determine meanings of words using contextual and structural clues.	-2.05	89	0.64	0.12	1.21
2	Vocab_Item_45010	EL.04.RE.12 Distinguish and interpret words with multiple meanings (i.e., quarter) by using context clues.	-1.70	89	0.54	0.26	1.16
3	Vocab_Item_45692	EL.04.RE.12 Distinguish and interpret words with multiple meanings (i.e., quarter) by using context clues.	-0.91	89	0.39	0.82	1.07
4	Vocab_Item_55187	EL.05.RE.11 Determine meanings of words using contextual and structural clues.	-0.86	78	0.47	0.52	1.10
5	Vocab_Item_45694	EL.04.RE.12 Distinguish and interpret words with multiple meanings (i.e., quarter) by using context clues.	0.56	89	0.32	0.86	1.10
6	Vocab_Item_45594	EL.04.RE.11 Determine meanings of words using contextual and structural clues.	0.65	89	0.28	0.79	1.11
7	Vocab_Item_45047	EL.04.RE.11 Determine meanings of words using contextual and structural clues.	1.50	89	0.25	1.21	0.53
8	Vocab_Item_45335	EL.04.RE.13 Apply knowledge of synonyms, antonyms, homographs, and idioms to determine the meaning of words and phrases.	1.66	89	0.25	0.96	1.06
9	Vocab_Item_55144	EL.05.RE.12 Understand and explain frequently used synonyms, antonyms, and homographs.	-1.53	78	0.53	0.45	1.07
10	Vocab_Item_45466	EL.04.RE.13 Apply knowledge of synonyms, antonyms, homographs, and idioms to determine the meaning of words and phrases.	-1.36	89	0.49	0.51	1.04

Table 34. Key Findings Summary from Technical Report 1211

Category	Summary
Grade Level	Grade 4
Assessment Focus	Vocabulary curriculum-based measurement
Statistical Model	1PL Rasch
Forms Developed	10 progress monitoring and 3 benchmark forms
Primary Outcome	Technically adequate vocabulary measures for RTI use

Reference

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 4 (Technical Report 1211).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1212: The Development of CBM Vocabulary Measures: Grade 5 (Alonzo, Anderson, et al., 2012d).

Methods

Researchers at the University of Oregon developed vocabulary Curriculum-Based Measures (CBMs) for grades 2–8 to support Response to Intervention (RTI) screening and progress monitoring. Item development was guided by Universal Design for Assessment (UDA) principles, prioritizing accessibility for English language learners and students receiving special education services. Eight item writers from Oregon and California created approximately 800 items per grade level beginning in October 2008, with completion in May 2009. Items were multiple-choice with three answer options, aligned to Oregon State Vocabulary Standards, and designed to minimize cognitive and linguistic complexity. A six-member research team reviewed all items for clarity, alignment, and formatting from June through August 2009, resulting in approximately 5,600 items for piloting. In spring 2011, roughly 1,200 students per grade across the United States completed pilot assessments administered online via the easyCBM platform. A common-item design using anchor items allowed all items to be calibrated to a single scale. Data were analyzed using a one-parameter logistic (1PL) Rasch model with Winsteps software, examining Mean Square Outfit, Standard Error of Measure, item difficulty, and distractor functioning.

Results

Although no items were originally written specifically for Grade 5, a subset of very easy items from adjacent grades was identified through piloting and organized into Grade 5 measures. Of the Grade 5 items analyzed, 135 were over-fit (Mean Square Outfit: 0.07–0.49; average 0.31), with 134 retained after distractor analysis confirmed appropriate functioning. An additional 31 items were under-fit (Mean Square Outfit: 1.51–4.95; average 2.17); all were removed because higher-ability students were not reliably selecting the correct answer. In total, 32 items were excluded from the Grade 5 item bank. The retained items were organized into 13 alternate forms: 10 progress monitoring forms and 3 benchmark screener forms, each containing 20 items. Mean difficulty measures across progress monitoring forms ranged from −0.24 to −0.19 (average −0.22), and benchmark forms ranged from −0.22 to −0.21 (average −0.22), indicating comparable and consistently accessible difficulty levels suitable for both general education students and those receiving intensive intervention.

Table 35. Example of Item Information and Statistics from Technical Report 1212

Table 1a

Item Information and Statistics, Gr5Vocab1 (Mean Measure = -0.23)

Item	Item Name	Standard	Measure	Count	Error	MSQ Outfit	Disc
1	Vocab_Item_55232	EL.05.RE.12 Understand and explain frequently used synonyms, antonyms, and homographs.	-2.59	78	0.75	0.12	1.18
2	Vocab_Item_65229	EL.06.RE.11 Determine the meaning of unknown words or words with unusual meanings in informational and narrative text by using word, sentence, and paragraph clues.	-2.27	93	0.63	0.70	1.01
3	Vocab_Item_55867	EL.05.RE.12 Understand and explain frequently used synonyms, antonyms, and homographs.	-1.26	78	0.44	1.05	0.90
4	Vocab_Item_55017	EL.05.RE.13 Determine the meanings of figurative expressions, such as those in similes and metaphors.	-1.23	79	0.49	1.11	1.00
5	Vocab_Item_55594	EL.05.RE.12 Understand and explain frequently used synonyms, antonyms, and homographs.	0.54	78	0.30	1.07	0.82
6	Vocab_Item_55452	EL.05.RE.11 Determine meanings of words using contextual and structural clues.	0.58	78	0.29	0.85	1.06
7	Vocab_Item_55732	EL.05.RE.12 Understand and explain frequently used synonyms, antonyms, and homographs.	1.49	78	0.27	0.70	1.48
8	Vocab_Item_55694	EL.05.RE.13 Determine the meanings of figurative expressions, such as those in similes and metaphors.	1.66	78	0.26	0.94	1.19
9	Vocab_Item_65106	EL.06.RE.12 Interpret figurative language, including similes, metaphors, and words with multiple meanings.	-1.79	92	0.51	0.53	1.02
10	Vocab_Item_55463	EL.05.RE.12 Understand and explain frequently used synonyms, antonyms, and homographs.	-1.53	78	0.54	0.49	1.08

Table 36. Key Findings Summary from Technical Report 1212

Category	Summary
Grade Level	Grade 5
Assessment Focus	Vocabulary curriculum-based measurement
Statistical Model	1PL Rasch
Forms Developed	10 progress monitoring and 3 benchmark forms
Primary Outcome	Technically adequate vocabulary measures for RTI use

Reference

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 5 (Technical Report 1212).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1213: The Development of CBM Vocabulary Measures: Grade 6 (Alonzo, Anderson, et al., 2012).

Methods

Participants included approximately 1,200 students drawn from schools across the United States. Data collection occurred during spring benchmark administrations conducted via computer-based delivery. Items were presented individually, and student responses were captured electronically, ensuring standardized administration conditions and consistent scoring procedures across sites.

Item development was guided by principles of universal design for assessment to reduce construct-irrelevant barriers related to language, formatting, and working memory demands. Items were aligned to grade-level vocabulary standards and reviewed extensively by trained researchers prior to piloting. Anchor items were embedded in forms to support calibration and comparability across administrations.

Statistical analyses were conducted using a one-parameter logistic Rasch model implemented in Winsteps. Item difficulty estimates, fit statistics, and standard errors were examined to evaluate item functioning. Distractor analyses were used to identify items with inadequate response patterns. Items exhibiting underfit or non-functioning distractors were removed from the item bank, while overfitting items were retained when distractor functioning was acceptable.

Results

Results supported the creation of ten progress monitoring forms and three benchmark forms with comparable difficulty levels. Mean form measures indicated strong alignment across forms, supporting interpretability for instructional decision-making. Overall findings indicate that the Grade 6 vocabulary measures demonstrate strong technical adequacy and are appropriate for use in benchmarking and progress monitoring.

Table 37. Example of Item Information and Statistics from Technical Report 1213

Table 1a
Item Information and Statistics, Gr6Vocab1 (Mean Measure = -0.23)

Item	Item Name	Standard	Measure	Count	Error	MSQ Outfit	Disc
1	Vocab_Item_75188	EL.07.RE.11 Demonstrate understanding of idioms and comparisons, such as analogies, metaphors, and similes, in prose (informational and literary text) and poetry.	-2.49	69	1.02	0.14	1.13
2	Vocab_Item_75870	EL.07.RE.12 Clarify word meanings through the use of definition, inference, example, restatement, or contrast.	-2.21	68	0.76	0.45	1.04
3	Vocab_Item_75871	EL.07.RE.12 Clarify word meanings through the use of definition, inference, example, restatement, or contrast.	-1.05	69	0.47	0.58	1.10
4	Vocab_Item_65334	EL.06.RE.12 Interpret figurative language, including similes, metaphors, and words with multiple meanings.	-1.01	94	0.44	0.23	1.28
5	Vocab_Item_65407	EL.06.RE.11 Determine the meaning of unknown words or words with unusual meanings in informational and narrative text by using word, sentence, and paragraph clues.	0.33	92	0.30	0.72	1.19
6	Vocab_Item_65605	EL.06.RE.12 Interpret figurative language, including similes, metaphors, and words with multiple meanings.	0.38	92	0.34	0.43	1.42
7	Vocab_Item_65147	EL.06.RE.12 Interpret figurative language, including similes, metaphors, and words with multiple meanings.	1.40	92	0.27	0.81	1.27
8	Vocab_Item_65634	EL.06.RE.11 Determine the meaning of unknown words or words with unusual meanings in informational and narrative text by using word, sentence, and paragraph clues.	1.59	95	0.26	0.80	1.31
9	Vocab_Item_65106	EL.06.RE.12 Interpret figurative language, including similes, metaphors, and words with multiple meanings.	-1.79	92	0.51	0.53	1.02
10	Vocab_Item_75680	EL.07.RE.10 Determine meanings of words using contextual and structural clues.	-1.55	68	0.57	0.44	1.03

Table 38. Key Findings Summary from Technical Report 1213

Category	Summary
Grade Level	Grade 6
Assessment Focus	Vocabulary curriculum-based measurement
Statistical Model	1PL Rasch
Forms Developed	10 progress monitoring and 3 benchmark forms
Primary Outcome	Technically adequate vocabulary measures for RTI use

Reference

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 6 (Technical Report 1213).* Eugene OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1214: The Development of CBM Vocabulary Measures: Grade 7 (Alonzo, Anderson, et al., 2012e).

This technical report documents the development and pilot evaluation of curriculum-based measurement vocabulary assessments designed for use with students in Grade 7. The measures were developed as part of the easyCBM® assessment system to support both benchmarking and progress monitoring within a response-to-intervention framework. Emphasis was placed on producing brief, repeatable measures that were sensitive to growth while remaining accessible to students with a wide range of ability levels.

Methods

Participants included approximately 1,200 students drawn from schools across the United States. Data collection occurred during spring benchmark administrations conducted via computer-based delivery. Items were presented individually, and student responses were captured electronically, ensuring standardized administration conditions and consistent scoring procedures across sites.

Item development was guided by principles of universal design for assessment to reduce construct-irrelevant barriers related to language, formatting, and working memory demands. Items were aligned to grade-level vocabulary standards and reviewed extensively by trained researchers prior to piloting. Anchor items were embedded within forms to support calibration and comparability across administrations.

Statistical analyses were conducted using a one-parameter logistic Rasch model implemented in Winsteps. Item difficulty estimates, fit statistics, and standard errors were examined to evaluate item functioning. Distractor analyses were used to identify items with inadequate response patterns. Items exhibiting underfit or non-functioning distractors were removed from the item bank, while overfitting items were retained when distractor functioning was acceptable.

Results

Results supported the creation of ten progress monitoring forms and three benchmark forms with comparable difficulty levels. Mean form measures indicated strong alignment across forms, supporting interpretability for instructional decision-making. Overall findings indicate that the Grade 7 vocabulary measures demonstrate strong technical adequacy and are appropriate for use in benchmarking and progress monitoring.

Table 39. Example of Item Information and Statistics from Technical Report 1214

Table 1a

Item Information and Statistics, Gr7Vocab1 (Mean Measure = -0.16)

Item	Item Name	Standard	Measure	Count	Error	MSQ Outfit	Disc
1	Vocab_Item_75097	EL.07.RE.10 Determine meanings of words using contextual and structural clues.	-2.03	68	0.63	0.24	1.17
2	Vocab_Item_75766	EL.07.RE.10 Determine meanings of words using contextual and structural clues.	-1.89	69	0.75	0.18	1.16
3	Vocab_Item_75557	EL.07.RE.12 Clarify word meanings through the use of definition, inference, example, restatement, or contrast.	-1.02	68	0.50	1.37	0.87
4	Vocab_Item_85139	EL.08.RE.10 Determine meanings of words using contextual and structural clues.	-0.96	58	0.61	0.38	1.13
5	Vocab_Item_75815	EL.07.RE.10 Determine meanings of words using contextual and structural clues.	0.35	69	0.33	0.79	1.09
6	Vocab_Item_75013	EL.07.RE.11 Demonstrate understanding of idioms and comparisons, such as analogies, metaphors, and similes, in prose (informational and literary text) and poetry.	0.49	69	0.34	1.20	0.91
7	Vocab_Item_75510	EL.07.RE.12 Clarify word meanings through the use of definition, inference, example, restatement, or contrast.	1.54	69	0.30	1.27	0.68
8	Vocab_Item_75691	EL.07.RE.12 Clarify word meanings through the use of definition, inference, example, restatement, or contrast.	1.68	68	0.29	1.19	0.39
9	Vocab_Item_75151	EL.07.RE.12 Clarify word meanings through the use of definition, inference, example, restatement, or contrast.	-1.50	69	0.56	0.45	1.10
10	Vocab_Item_85098	EL.08.RE.11 Analyze idioms and comparisons, such as analogies, metaphors, and similes, to infer the literal and figurative meanings of phrases.	-1.39	58	0.76	0.17	1.23

Table 40. Key Findings Summary from Technical Report 1214

Category	Summary
Grade Level	Grade 7
Assessment Focus	Vocabulary curriculum-based measurement
Statistical Model	1PL Rasch
Forms Developed	10 progress monitoring and 3 benchmark forms
Primary Outcome	Technically adequate vocabulary measures for RTI use

Reference

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 7 (Technical Report 1214).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1215: The Development of CBM Vocabulary Measures: Grade 8 (Alonzo, Anderson, et al., 2012f).

This technical report documents the development and pilot evaluation of curriculum-based measurement vocabulary assessments designed for use with students in Grade 8. The measures were developed as part of the easyCBM® assessment system to support both benchmarking and progress monitoring within a response-to-intervention framework. Emphasis was placed on producing brief, repeatable measures that were sensitive to growth while remaining accessible to students with a wide range of ability levels.

Methods

Participants included approximately 1,200 students drawn from schools across the United States. Data collection occurred during spring benchmark administrations conducted via computer-based delivery. Items were presented individually, and student responses were captured electronically, ensuring standardized administration conditions and consistent scoring procedures across sites.

Item development was guided by principles of universal design for assessment to reduce construct-irrelevant barriers related to language, formatting, and working memory demands. Items were aligned to grade-level vocabulary standards and reviewed extensively by trained researchers prior to piloting. Anchor items were embedded within forms to support calibration and comparability across administrations. Statistical analyses were conducted using a one-parameter logistic Rasch model implemented in Winsteps. Item difficulty estimates, fit statistics, and standard errors were examined to evaluate item functioning. Distractor analyses were used to identify items with inadequate response patterns. Items exhibiting underfit or non-functioning distractors were removed from the item bank, while overfitting items were retained when distractor functioning was acceptable.

Results

Results supported the creation of 10 progress monitoring forms and 3 benchmark forms with comparable difficulty levels. Mean form measures indicated strong alignment across forms, supporting interpretability for instructional decision-making. Overall findings indicate that the Grade 8 vocabulary measures demonstrate strong technical adequacy and are appropriate for benchmarking and progress monitoring.

Table 41. Example of Item Information and Statistics from Technical Report 1215

Table 1a

Item Information and Statistics, Gr8Vocab1 (Mean Measure = -0.15)

Item	Item Name	Standard	Measure	Count	Error	MSQ Outfit	Disc
1	Vocab_Item_85372	EL.08.RE.11 Analyze idioms and comparisons, such as analogies, metaphors, and similes, to infer the literal and figurative meanings of phrases.	-2.49	58	1.03	0.16	1.13
2	Vocab_Item_85640	EL.08.RE.11 Analyze idioms and comparisons, such as analogies, metaphors, and similes, to infer the literal and figurative meanings of phrases.	-2.03	58	1.02	0.32	1.06
3	Vocab_Item_85049	EL.08.RE.10 Determine meanings of words using contextual and structural clues.	-0.94	58	0.52	1.30	0.89
4	Vocab_Item_85369	EL.08.RE.11 Analyze idioms and comparisons, such as analogies, metaphors, and similes, to infer the literal and figurative meanings of phrases.	-0.88	58	0.58	0.22	1.30
5	Vocab_Item_85638	EL.08.RE.11 Analyze idioms and comparisons, such as analogies, metaphors, and similes, to infer the literal and figurative meanings of phrases.	0.04	58	0.41	1.43	1.01
6	Vocab_Item_85453	EL.08.RE.10 Determine meanings of words using contextual and structural clues.	0.16	58	0.42	0.67	1.15
7	Vocab_Item_85417	EL.08.RE.11 Analyze idioms and comparisons, such as analogies, metaphors, and similes, to infer the literal and figurative meanings of phrases.	0.59	58	0.43	1.07	0.86
8	Vocab_Item_85238	EL.08.RE.11 Analyze idioms and comparisons, such as analogies, metaphors, and similes, to infer the literal and figurative meanings of phrases.	0.73	58	0.37	1.17	0.84
9	Vocab_Item_85150	EL.08.RE.12 Verify the meaning of a word in its context, even when its meaning is not directly stated, through the use of definition, restatement, example, comparison, or contrast.	-1.61	57	0.76	0.19	1.18
10	Vocab_Item_85871	EL.08.RE.12 Verify the meaning of a word in its context, even when its meaning is not directly stated, through the use of definition, restatement, example, comparison, or contrast.	-1.22	58	0.59	0.23	1.27

Table 42. Key Findings Summary from Technical Report 1215

Category	Summary
Grade Level	Grade 8
Assessment Focus	Vocabulary curriculum-based measurement
Statistical Model	1PL Rasch
Forms Developed	10 progress monitoring and 3 benchmark forms
Primary Outcome	Technically adequate vocabulary measures for RTI use

Reference

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 8 (Technical Report 1215).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1221: The Development of the easyCBM® CCSS Reading Assessments: Grade 3 (Alonzo, Park, et al., 2012a).

This technical report documents the development and piloting of the easyCBM® Common Core State Standards Reading Assessment for Grade 3. The assessment was designed to support screening and progress monitoring for students at risk for reading difficulties and low-performing students.

Methods

Item development followed principles of Universal Design for Assessment and alignment with CCSS reading standards. Educators with experience teaching struggling readers authored literary, informational, and task-based passages, each paired with multiple-choice comprehension questions. All items underwent expert review for bias, accessibility, and construct relevance. Pilot testing occurred online across multiple U.S. states using convenience samples. Item Response Theory testlet models were used to analyze item difficulty, discrimination, and fit.

Results

Results showed that Read to Perform a Task items were generally less difficult, while literary and informational texts exhibited wider variability. Overall findings support the technical adequacy of the measures for progress monitoring, though subtest-level interpretation is not recommended.

Table 43. Typical Results of IRT Analysis from Technical Report 1221

Table 2
Results of IRT Analysis, Mean Measure of Progress Monitoring Measure 3_2 = 0.30

Test Form	Item #	Reading Type	Standard	Item Name	Measure	Count	Error	Mean Square Outfit	Discrimination
2	1	Literary Text: Demonstrate General Understanding	Identify the speaker or narrator in a selection. Distinguish the order of events or a specific event from a sequence of events. Determine significant events from the story.	35756 to 35760	-0.71	61	0.35	1.25	0.95
2	2	Literary Text: Develop an Interpretation	Determine what characters are like by what they say or do and by how the author or illustrator portrays them. Predict probable future outcomes or actions. Determine and discuss the underlying theme or author's message in literary text. Recognize cause-and-effect relationships in literary text.	35851 to 35855	1.68	65	0.35	0.78	0.99
2	3	Informational Text: Develop an Interpretation	Recall major points in the text and make predictions about forthcoming information. Distinguish cause-and-effect and fact and opinion.	35256 to 35260	0.07	65	0.40	0.43	1.50
2	4	Informational Text: Develop an Interpretation	Recall major points in the text and make predictions about forthcoming information. Distinguish cause-and-effect and fact and opinion.	35706 to 35710	1.18	64	0.26	0.71	1.23
2	5	Read to Perform a Task	Follow simple multiple-step written instructions (e.g., how to assemble a product or play a board game).	35516 to 35520	-0.74	66	0.24	1.87	0.10

Table 44. Key Findings Summary from Technical Report 1221

Category	Summary
Sample	Grade 3 students from multiple U.S. states; convenience sample
Assessment Types	Literary Text, Informational Text, Read to Perform a Task
Analysis Method	Item Response Theory (testlet-based modeling)
Key Finding	Read to Perform a Task items were least difficult
Conclusion	Measures demonstrate strong technical adequacy for monitoring growth

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2012). *The development of the easyCBM® CCSS reading assessments: Grade 3 (Technical Report 1221).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report No 1222: The Development of the easyCBM® CCSS Reading Assessments: Grade 4 (Alonzo, Park, et al., 2012b).

This report describes the development and validation of the Grade 4 easyCBM® CCSS Reading Assessment, created to measure reading comprehension for screening and progress monitoring within an RTI framework. The assessment targets students who struggle with reading comprehension.

Methods

Assessment items were developed by trained educators and aligned to CCSS standards, incorporating literary text, informational text, and task-based reading. Universal Design principles guided item construction to ensure accessibility for diverse learners, including students with disabilities. Pilot data were collected online and analyzed using Item Response Theory.

Results

Results indicated substantial variability in difficulty across passages, with no consistent pattern indicating a single most challenging subtest. Findings support the use of the assessment as a unidimensional measure of reading comprehension, suitable for monitoring student growth over time.

Table 45. Typical Results of IRT Analysis for Technical Report 1222

Table 1
Results of IRT Analysis, Mean Measure of Progress Monitoring Measure 4_1 = 0.52

Test Form	Item #	Reading Type	Standard	Item Name	Measure	Count	Error	Mean Square Outfit	Discrimination
1	1	Literary Text: Demonstrate General Understanding	Identify and/or summarize sequence of events, main ideas, and supporting details in literary selections. Identify the main problem or conflict of the plot, and explain how it is resolved.	45716 to 45720	-2.66	58	0.49	9.90	0.71
1	2	Literary Text: Demonstrate General Understanding	Identify and/or summarize sequence of events, main ideas, and supporting details in literary selections. Identify the main problem or conflict of the plot, and explain how it is resolved.	45846 to 45850	1.28	61	0.41	0.41	1.56
1	3	Informational Text: Demonstrate General Understanding Informational Text: Develop an Interpretation	Identify and/or summarize sequence of events, main ideas, facts, supporting details, and opinions in informational and practical selections. Make and confirm predictions about text by using prior knowledge and ideas presented in the text itself, including illustrations, titles, topic sentences, and important words. Draw inferences or conclusions about an author's meaning supported by facts and events from the text.	45071 to 45075	-0.52	60	0.29	0.93	0.58

Table 46. Key Findings Summary from Technical Report 1222

Category	Summary
Sample	Grade 4 students across multiple U.S. states
Assessment Types	Literary Text, Informational Text, Read to Perform a Task
Analysis Method	Item Response Theory with testlet structure
Key Finding	Wide variability in item difficulty across subtests
Conclusion	Assessment supports valid, reliable progress monitoring

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2012). *The development of the easyCBM® CCSS reading assessments: Grade 4 (Technical Report 1222).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1223: The Development of the easyCBM® CCSS Reading Assessments: Grade 5 (Alonzo, Park, et al., 2012c).

This technical report documents the development and pilot evaluation of the easyCBM® Common Core State Standards–aligned reading comprehension assessments for Grade 5. The assessments were designed to support both universal screening and progress monitoring, with particular emphasis on students who demonstrate persistent difficulty in reading. The measures target comprehension across literary text, informational text, and read-to-perform-a-task formats in alignment with grade-level CCSS expectations.

Methods

Participants included students from multiple states who participated through convenience sampling via the online easyCBM® assessment platform. Item piloting occurred during fall administrations, with students completing subsets of items to support large-scale calibration while minimizing individual testing burden. Administration procedures were standardized and computer-based, allowing automated capture of student responses. Item development followed principles of Universal Design for Assessment to reduce construct-irrelevant variance. Educators with experience teaching struggling readers authored passages

and associated multiple-choice items, which were reviewed for bias, accessibility, and alignment prior to piloting. Each passage and its associated items were treated as testlets to account for local item dependence. Data analyses were conducted using Item Response Theory testlet models. Item difficulty, fit statistics, and discrimination indices were examined to evaluate psychometric functioning. Results indicated that most items demonstrated acceptable fit and contributed to well-targeted forms. Ten progress monitoring forms and three benchmark forms were assembled with comparable difficulty levels to support growth.

Results

Overall findings indicate that the Grade 5 easyCBM® CCSS reading assessments demonstrate strong technical adequacy. The measures provide reliable and interpretable information for identifying students at risk and monitoring growth over time within data-based decision-making frameworks.

Table 47. Key Findings Summary from Technical Report 1223

Category	Summary
Grade Level	Grade 5
Assessment Focus	CCSS-aligned reading comprehension
Sample	Multi-state convenience sample
Statistical Model	IRT testlet modeling
Forms Developed	10 progress monitoring, 3 benchmarks
Intended Use	Screening and progress monitoring

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2012). *The development of the easyCBM® CCSS reading assessments: Grade 5 (Technical Report 1223).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1224: The Development of the easyCBM® CCSS Reading Assessments: Grade 6 (Alonzo, Park, et al., 2012d).

Methods

Researchers developed **vocabulary** Curriculum-Based Measures (CBMs) for use with students in grades 2–8 as part of the easyCBM® online assessment system. Item development was guided by Oregon State Vocabulary Standards and principles of Universal Design for Assessment (UDA), with a focus on reducing cognitive and linguistic complexity to ensure accessibility for students with disabilities, English language learners, and those receiving special education services. Eight qualified item writers created approximately 800 items per grade level between October 2008 and May 2009, producing three-choice multiple-choice items contextualized in brief sentences with bolded target words. A team of six researchers reviewed all items for clarity, alignment, and formatting from June through August 2009, yielding approximately 5,600 items for piloting. Items were administered online to approximately 1,200 students per grade from schools across the United States during spring 2011, using a common-item design with anchor items to enable calibration across forms. Data were analyzed using a one-parameter logistic (1PL) Rasch model in Winsteps, examining Mean Square Outfit, Standard Error of Measure, item difficulty, and distractor functioning.

Results

Although no items were originally written for Grade 6, a set of items identified as extremely easy during piloting were organized into sixth-grade vocabulary measures. Of the Grade 6 items analyzed, 146 were over-fit (Mean Square Outfit: 0.07–0.49; average 0.33), of which 145 were retained based on appropriate distractor functioning. A total of 41 items were under-fit (Mean Square Outfit: 1.51–9.64; average 2.32) and removed due to poor distractor performance, along with one additional item removed on content review grounds, yielding 43 total removed items. Retained items were organized into 13 forms: ten progress monitoring forms and three benchmark screener forms, each comprising 20 items. Mean item difficulty measures across progress monitoring forms ranged from −0.23 to −0.19 (average −0.21), and benchmark forms ranged from −0.22 to −0.21 (average −0.21), indicating comparable difficulty across all alternate forms.

Table 48. Key Findings Summary from Technical Report 1224

Category	Summary
Grade Level	Grade 6
Assessment Focus	CCSS-aligned reading comprehension
Sample	Multi-state convenience sample
Statistical Model	IRT testlet modeling
Forms Developed	10 progress monitoring, 3 benchmarks
Intended Use	Screening and progress monitoring

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2012). *The development of the easyCBM® CCSS reading assessments: Grade 6 (Technical Report 1224).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1225: The Development of the easyCBM® CCSS Reading Assessments: Grade 7 (Alonzo, Park, et al., 2012).

Methods

This study developed reading comprehension assessments for 7th-grade students aligned to the Common Core State Standards (CCSS), with particular focus on persistently low-performing students (approximately 2% of the student population) who do not respond to interventions. The assessment team applied Universal Design for Assessment principles, using the NCEO guidelines and the Test Accessibility and Modification Inventory (TAMI) to maximize fairness and accessibility. Item writers — educators experienced with struggling readers — independently created passages and graphics across three sub-test types: Short Literary Text, Informational Text, and Read to Perform a Task. Each passage was paired with five multiple-choice questions targeting specific CCSS reading standards, with one correct answer and two plausible distractors. A research team of 3–5 members reviewed all 900 items per grade for bias, age-appropriateness, and Universal Design compliance. Items were piloted online in November–December 2011 with convenience samples of students from eight states, randomly assigned to sub-test sets of 25 items to control for class effects and ensure adequate scaling data.

Results

Piloted data were analyzed using Item Response Theory (IRT), treating each passage and its five questions as a testlet. Results yielded 10 progress monitoring forms and 3 benchmark forms (Fall, Winter, Spring), each comprising two Literary Text sub-tests, two Informational Text sub-tests, and one Read to Perform a Task sub-test. IRT measures (item difficulty) for the Grade 7 progress monitoring forms ranged widely: Literary Text from -1.63 to 2.95, Informational Text from -1.66 to 5.89, and Read to Perform a Task from -

0.91 to 3.73. Benchmark form means were approximately 1.14 (Fall), 1.20 (Winter), and 1.17 (Spring). No consistent pattern emerged indicating which sub-test type was most difficult. Discrimination values were generally acceptable, though a small number of items showed poor fit and were excluded (Table 14). Importantly, the authors caution that sub-test scores should not be used to draw conclusions about student mastery of specific text types, as construct validity studies confirm all three sub-tests measure a single underlying reading construct.

Table 49. Key Findings Summary from Technical Report 1225

Category	Summary
Grade Level	Grade 7
Assessment Focus	CCSS-aligned reading comprehension
Sample	Multi-state convenience sample
Statistical Model	IRT testlet modeling
Forms Developed	10 progress monitoring, 3 benchmarks
Intended Use	Screening and progress monitoring

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2012). *The development of the easyCBM® CCSS reading assessments: Grade 7 (Technical Report 1225).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1226: The Development of the easyCBM® CCSS Reading Assessments: Grade 8 (Alonzo, Park, et al., 2012e).

Methods

This study developed and piloted Grade 8 reading comprehension assessments aligned to the Common Core State Standards (CCSS), designed primarily for persistently low-performing students — a subgroup representing roughly 2% of the student population who do not respond consistently to interventions. Assessment items were authored by experienced educators trained in Universal Design for Assessment and CCSS content standards. Item writers independently produced literary passages, informational texts, and graphics-based prompts, each accompanied by five multiple-choice questions targeting specific grade-level reading standards. A graphic artist rendered all visual materials for computer-based delivery. All items — totaling 900 across grades 3–8 — were reviewed by 3–5 trained researchers for bias, grade-level appropriateness, and universal design compliance before piloting. Items were piloted online in November–December 2011 using convenience samples from Oregon, Washington, Montana, Florida, Texas, Illinois, California, and Wisconsin. Each student completed 25 items across five randomly assigned sub-tests to reduce class-level bias. Data were analyzed using Item Response Theory (IRT), treating each passage and its five questions as a testlet.

Results

IRT analyses produced 10 progress monitoring forms and 3 benchmark forms (Fall, Winter, Spring), each containing two Literary Text sub-tests, two Informational Text sub-tests, and one Read to Perform a Task sub-test. Item difficulty (IRT Measure) varied widely across and within sub-test types. For progress monitoring forms, Read to Perform a Task difficulty ranged from -0.71 to 2.08; Literary Text ranged from -3.84 to 3.08; and Informational Text ranged from -4.84 to 3.63. Benchmark assessments showed similar variability. No consistent pattern emerged indicating which sub-test type was most challenging — difficulty appeared to depend more on the specific passage than on text category. The authors caution that sub-test scores should not be interpreted as separate indicators of skill by text type, as construct

validity studies confirm all three sub-tests measure a single underlying reading construct. A small number of items were excluded from final forms due to extreme difficulty values or poor psychometric fit.

Table 50. Key Findings Summary from Technical Report 1226

Category	Summary
Grade Level	Grade 8
Assessment Focus	CCSS-aligned reading comprehension
Sample	Multi-state convenience sample
Statistical Model	IRT testlet modeling
Forms Developed	Benchmark and progress monitoring forms
Intended Use	Screening and progress monitoring

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2012). *The development of the easyCBM® CCSS reading assessments: Grade 8 (Technical Report 1226)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1110: Progress Monitoring Instrument Development: Silent Reading Fluency, Vocabulary, and Reading Comprehension (Nese et al., 2011).

This technical report describes the development of computer-based curriculum-based measurement (CBM) instruments designed to assess silent reading fluency, vocabulary, and reading comprehension in elementary students. The study focused on creating measures that could be administered digitally, addressing limitations associated with traditional oral reading fluency assessments that require individual administration.

Methods

A team of researchers and elementary teachers developed a series of experimental reading measures for Grades 3 through 5. Five types of instruments were created: silent reading fluency sentences, silent reading fluency maze passages, vocabulary maze passages, vocabulary sentence items, and reading comprehension passages. Each measure included multiple alternate forms to support progress monitoring. Items were constructed using grade-appropriate vocabulary and readability targets determined by the Flesch-Kincaid formula. The measures were designed for computer delivery, allowing automated capture of student responses and timing data. For fluency tasks, the system recorded reading time and converted it into words-per-minute estimates. Vocabulary and comprehension measures focused on accuracy of responses to multiple-choice items embedded in passages or sentences. All items were reviewed by a research team to ensure clarity, appropriate difficulty, and absence of bias before being placed into the computer interface.

Results

The development process produced multiple parallel forms of computer-administered reading assessments across the three reading domains. The resulting instruments generated automated data on reading speed, accuracy, and item responses, demonstrating the feasibility of collecting progress-monitoring data through computer-based testing. The measures were designed to maintain student engagement while minimizing cognitive load from distracters so that reading fluency and vocabulary knowledge could be captured efficiently. The report concludes that computerized CBM instruments could improve the efficiency of classroom progress monitoring by allowing groups of students to complete assessments simultaneously while performance data are automatically recorded.

Reference

Nese, J. F. T., Anderson, D., Hoelscher, K., Tindal, G., & Alonzo, J. (2011). *Progress monitoring instrument development: Silent reading fluency, vocabulary, and reading comprehension (Technical Report # 1110).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1301: The Development of easyCBM® Spanish Literacy Assessments for Use in Grades K–2 (Alonzo et al., 2013).

Technical Report 1301 documents the development and evaluation of the easyCBM® **Spanish literacy assessments** for use with students in kindergarten through Grade 2. The purpose of the study was to identify early literacy measures appropriate for students receiving instruction in Spanish, recognizing important linguistic differences between Spanish and English that influence literacy development and assessment. The assessments were intended to support progress monitoring and instructional decision making within Spanish literacy programs.

Methods

Two initial item development studies were conducted with distinct student populations. The first included native English speakers enrolled in a Spanish immersion program, while the second involved native Spanish speakers enrolled in a dual-language bilingual program. Across both studies, students received all literacy instruction in Spanish. Trained researchers individually administered multiple early literacy measures, including letter sounds, syllable sounds, phoneme segmenting, syllable segmenting, word reading fluency, and sentence reading fluency, following standardized protocols.
Analyses focused on examining relationships between early literacy measures and reading fluency outcomes. Correlational analyses and stepwise linear regression models were used to identify which measures best predicted word and sentence reading performance. Across studies and grade levels, performance on syllable-level measures, particularly syllable sounds and syllable reading, consistently emerged as the strongest predictors of reading fluency. Phoneme-level measures showed weaker and less consistent relationships with reading outcomes, especially for younger students.

A subsequent reliability study examined alternate form reliability across kindergarten, Grade 1, and Grade 2. Data were collected from native Spanish-speaking students using multiple forms of each measure. Classical test theory analyses demonstrated very strong correlations among alternate forms of syllable reading, word reading fluency, and sentence reading fluency measures. Regression analyses further indicated that word reading fluency accounted for substantial variance in sentence reading fluency, particularly at Grade 2.

Results

Overall findings support the technical adequacy of the easyCBM® Spanish literacy assessments. Results underscore the importance of syllable-level measures in Spanish literacy development and provide empirical support for assessment systems aligned with the linguistic structure of Spanish for early readers.

Table 51. Example of Descriptive Statistics from Report 1301

Table 2

Descriptive Statistics for Study 1 and Study 2, Grade 1

Measure	Study 1			Study 2		
	n	*M*	*SD*	*n*	*M*	*SD*
Letter Sounds	48	14.00	7.31	35	16.06	9.40
Syllable Sounds	48	20.44	7.89	35	24.94	9.63
Phoneme Segmenting	48	22.35	11.17	35	22.29	9.23
Syllable Segmenting	48	17.31	5.93	35	19.34	4.33
Word Reading	48	14.17	5.17	35	22.94	7.97
Sentence Reading	-NA-	-NA-	-NA-	34	24.12	11.67

Table 52. Typical Set of Relations Among Measures from Report 1301

Table 4

Correlations Among the Grade 1 Measures in Study 1 (n = 48) & Study 2, Grade 1(n =35)

	Letter Sounds		Syllable Sounds		Phoneme Segmenting		Syllable Segmenting		Word Reading	
	1	2	1	2	1	2	1	2	1	2
Syllable Sounds	.46**	.28								
Phoneme Segmenting	.42**	.13	.33*	.01						
Syllable Segmenting	-.11	-.07	.05	-.04	.23	-.06				
Word Reading	.31*	.09	.87**	.56**	.21	.02	.19	.22		
Sentence Reading	NA	.06	NA	.45**	NA	-.10	NA	.11	NA	.57**

*$p < .05$
**$p < .01$

Table 53. Example of Sentence Reading Regression Summary from Technical Report 1301

Table 7

Sentence Reading Regression Summary for Grade 2, Study 1

Independent Variables	Unstandardized Coefficients		Standardized Coefficients	t	Partial and Part Correlations	
	B	Std. Error	Beta		Partial	Part
Syllable Sounds	1.34	0.25	0.57	5.29	0.61	0.52
Syllable Segmenting	0.96	0.37	0.28	2.62	0.36	0.26
Constant	-24.15	8.08		-2.99		

Table 54. Key Findings Summary from Technical Report 1301

Category	Summary
Grades	Kindergarten through Grade 2
Assessment Focus	Spanish early literacy
Sample	Spanish immersion and dual-language bilingual students
Statistical Models	Correlations and linear regression
Primary Finding	Syllable-level measures best predict reading fluency
Intended Use	Progress monitoring and instructional decision making

Reference

Alonzo, J., Gonzalez, M., & Tindal, G. (2013). T*he development of easyCBM® Spanish literacy assessments for use in Grades K–2 (Technical Report 1301).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1901: Initial Norms for the easyCBM® Spanish Language Vocabulary Assessments (Grades 2 – 8) (Alonzo, 2019).

Methods

This technical report describes the development and initial norming of the easyCBM® Spanish Language Vocabulary Assessments for Grades 2–8. The assessments were developed in summer and fall 2016 by a team of native Spanish-speaking educators working collaboratively with University of Oregon assessment developers. Item writers included teachers and paraprofessionals with grade-specific experience in Spanish-language instruction. Prior to writing items, the team participated in a structured training workshop focused on bias reduction, avoidance of regional dialect differences, and the importance of writing original Spanish items rather than translating English vocabulary content.

Items were written in small grade-band teams (Grades 2–3, 4–5, 6–8) and overseen by a Lead Item Developer with expertise in Spanish literature and bilingual education. Each vocabulary item consisted of a sentence stem designed to provide contextual meaning, followed by three answer choices: one correct answer, a near distractor, and a far distractor. After internal review, items were piloted online during the 2017 and 2018 school years. Item Response Theory (IRT) analyses were conducted in spring 2018 to calibrate items and construct 13 comparable alternate forms per grade. Three forms were designated for benchmark screening, with ten retained for progress monitoring. Initial norms were generated using data from Fall 2018 and Winter 2018 benchmark administrations. The sample included students from 19 states (see page 6). The report notes that the sample represents a convenience sample and that winter sample sizes were particularly small in several grades. Benchmark assessments were restricted to on-grade administration within district-designated testing windows.

Results

The initial normative data demonstrate expected developmental and seasonal trends in vocabulary performance across Grades 2–8. Across grades, winter percentile ranks generally reflect higher performance at equivalent raw scores compared to fall administrations, suggesting growth over the school year.

In Grades 2 through 5, raw score distributions show steady upward shifts from fall to winter. Students achieving moderate raw scores in fall often corresponded to substantially higher percentile ranks in winter, reflecting improved performance relative to peers. This pattern is consistent with expected midyear vocabulary development and suggests sensitivity to instructional exposure over time.

In Grades 6 through 8, similar seasonal increases are observed, although variability in percentile distributions is somewhat greater. Some grades display clustering of higher percentile ranks at relatively low raw scores, which may reflect smaller winter sample sizes and limited score spread in the initial dataset. In the upper grades, several raw scores correspond to very high percentile rankings, indicating possible ceiling effects or compressed distributions in the preliminary norming sample.

Overall, the data reflect reasonable growth patterns across seasons and grades but should be interpreted cautiously. Because norms are based on a convenience sample and winter participation was uneven across grades, the author emphasizes that these norms are preliminary. Larger and more representative samples are needed before the benchmarks can be considered fully stable for high-stakes decisions. The report recommends using these norms as provisional interpretive guidance rather than definitive cut scores until further data collection strengthens the normative base.

Reference

Alonzo, J. (2019). *Initial norms for the easyCBM® Spanish language vocabulary assessments (Technical Report No. 1901).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Appendix A: Technical Report Table Titles

Table 1. Tasks for Kindergarten Blueprint

Table 2. Tasks for Grade 1 Blueprint

Table 3. Tasks for Grades 2–8 Blueprint

Table 4. Standards Crosswalk with CCSS, NRP, and easyCBM®

Table 5. Example of IRT Analysis for Technical Report 0039

Table 6. Summary Table for Technical Report 0039

Table 7. Typical Results of IRT Analysis from Technical Report 0040

Table 8. Example of Descriptive Statistics from Technical Report 0040

Table 9. Key Findings Summary from Technical Report 0040

Table 10. Example of Descriptive Statistics from Technical Report 0043

Table 11. Sample Correlations from Technical Report 0043

Table 12. Example of Changes Made: Grade 5 Passage Reading Measures in Technical Report 0043

Table 13. Key Findings Summary from Technical Report 0043

Table 14. Illustrative Descriptive Statistics from Technical Report 0046

Table 15. Typical Correlation from Technical Report 0046

Table 16. Key Findings Summary from Technical Report 0046

Table 17. Example of Item Statistics from Technical Report 0041

Table 18. Typical Distractor Analysis from Technical Report 0041

Table 19. Example of Item Statistics from Technical Report 0807

Table 20. Illustrative Distractor Analysis from Technical Report 0807

Table 21. Summary of Key Findings from Technical Report 0807

Table 22. Sample Item Statistics from Technical Report 0808

Table 23. Illustrative Distractor Analysis from Technical Report 0808

Table 24. Summary of Key Findings from Technical Report 0808

Table 25. Summary of Key Findings from Technical Report 0041, 0807, and 0808

Table 26. Example Item Statistics from Technical Report 1102

Table 27. Key Findings Summary from Technical Report 1102

Table 28. Example of Summary Statistics from Technical Report 1209

Table 29. Key Findings Summary from Technical Report 1209

Table 30. Example Item Information and Statistics from Technical Report 1210

Table 31. Typical Distractor Analysis Results from Technical Report 1210

Table 32. Key Findings Summary from Technical Report 1210

Table 33. Sample Item Information and Statistics from Technical Report 1211

Table 34. Key Findings Summary from Technical Report 1211

Table 35. Example of Item Information and Statistics from Technical Report 1212

Table 36. Key Findings Summary from Technical Report 1212

Table 37. Example of Item Information and Statistics from Technical Report 1213

Table 38. Key Findings Summary from Technical Report 1213

Table 39. Example of Item Information and Statistics from Technical Report 1214

Table 40. Key Findings Summary from Technical Report 1214

Table 41. Example of Item Information and Statistics from Technical Report 1215

Table 42. Key Findings Summary from Technical Report 1215

Table 43. Typical Results of IRT Analysis from Technical Report 1221

Table 44. Key Findings Summary from Technical Report 1221

Table 45. Typical Results of IRT Analysis for Technical Report 1222

Table 46. Key Findings Summary from Technical Report 1222

Table 47. Key Findings Summary from Technical Report 1223

Table 48. Key Findings Summary from Technical Report 1224

Table 49. Key Findings Summary from Technical Report 1225

Table 50. Key Findings Summary from Technical Report 1226

Table 51. Example of Descriptive Statistics from Report 1301

Table 52. Typical Set of Relations Among Measures from Report 1301

Table 53. Example of Sentence Reading Regression Summary from Technical Report 1301

Table 54. Key Findings Summary from Technical Report 1301

Technical Report References

Alonzo, J. (2019). *Initial norms for the easyCBM® Spanish language vocabulary assessments (Technical Report # 1901)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012a). *The development of CBM vocabulary measures: Grade 2 (Technical Report # 1209)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012b). *The development of CBM vocabulary measures: Grade 3 (Technical Report # 1210)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012c). *The development of CBM vocabulary measures: Grade 4 (Technical Report # 1211)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012d). *The development of CBM vocabulary measures: Grade 5 (Technical Report # 1212)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). *The development of CBM vocabulary measures: Grade 6 (Technical Report # 1213)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012e). *The development of CBM vocabulary measures: Grade 7 (Technical Report # 1214)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012f). *The development of CBM vocabulary measures: Grade 8 (Technical Report # 1215)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Gonzalez, M., & Tindal, G. (2013). *The development of easyCBM Spanish literacy assessments for use in Grades K-2 (Technical Report # 1301)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Liu, K., & Tindal, G. (2007). *Examining the technical adequacy of reading comprehension measures in a progress monitoring assessment system (Technical Report # 41)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Liu, K., & Tindal, G. (2008). *Examining the technical adequacy of second-grade reading comprehension measures in a progress monitoring assessment system (Technical Report # 0808)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2008). *The development of middle school passage reading fluency measures in a progress monitoring assessment system (Technical Report # 46)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012a). *The development of the easyCBM CCSS reading assessments: Grade 3 (Technical Report # 1221)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012b). *The development of the easyCBM CCSS reading assessments: Grade 4 (Technical Report # 1222)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012c). *The development of the easyCBM CCSS reading assessments: Grade 5 (Technical Report # 1223)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012d). *The development of the easyCBM CCSS reading assessments: Grade 6 (Technical Report # 1224)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012). *The development of the easyCBM CCSS reading assessments: Grade 7 (Technical Report # 1225)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012e). *The development of the easyCBM CCSS reading assessments: Grade 8 (Technical Report # 1226)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2007a). *The development of early literacy measures for use in a progress monitoring assessment system: Letter names, letter sounds and phoneme segmenting (Technical Report # 39)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2007b). *The development of word and passage reading fluency measures in a progress monitoring assessment system (Technical Report # 40)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2008a). *The development of fifth-grade passage reading fluency measures for use in a progress monitoring assessment system (Technical Report # 43)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2008b). *Examining the technical adequacy of fifth-grade reading comprehension measures in a progress monitoring assessment system (Technical Report # 0807)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nese, J. F. T., Anderson, D., Hoelscher, K., Tindal, G., & Alonzo, J. (2011). *Progress mmonitoring instrument development: Silent reading fluency, vocabulary, and reading comprehension (Technical Report # 1110).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Alonzo, J., & Tindal, G. (2011). *The development and technical adequacy of seventh-grade reading comprehension measures in a progress monitoring assessment (Technical Report # 1102)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Conclusions Supporting Claims for Criterion 1.2: Item and Form Alignment

Text passages, assessment items, and resulting test forms align to the expectations of the ELA domains as outlined by college- and career-ready (CCR) standards.

1.2.a The text passages are of high quality and aligned with the expectations of CCR standards and aligned to assessment design specifications.

Texts are of publishable quality; text complexity is documented; texts are placed at grade level; distribution reflects CCR range; texts support inference; texts align to design specifications. TR 2101 characterizes easyCBM® reading measures as covering a range of text types: Proficient Reading uses original narrative fiction passages written at grade-level readability for target age groups and reviewed by classroom teachers prior to piloting; Basic Reading incorporates Informational Text, Read to Perform a Task, and Short Literary Text, reflecting the CCR range of text types across grades. Passage Reading Fluency texts follow consistent story grammar (introduction, rising action, climax, resolution) to minimize cross-form difficulty variation. Student-facing materials are formatted for clarity, with appropriate font sizes, high contrast, and compatibility with print and digital administration. However, TR 2101 does not document formal text complexity analysis using a quantitative instrument (e.g., Lexile), representing a gap relative to CCR text complexity requirements for grade-level placement evidence.

1.2.b Test items are written to elicit evidence of learning relative to one or more CCR standard/s and aligned to assessment design specifications.

Items can be clearly identified as measuring CCR standards; items align to design specifications; items are content-accurate. TR 2101 coded easyCBM® reading items against ELA standards across all 50 states, documenting alignment to Reading Literature standards RL 2.1–8.3 across Grades 2–8 (Table 5 and Table 6). Alignment ratings of Exact Match, Partial Match, and Deviation are reported by state group (CCSS Adopted, Revised, Unique), with majority Yes alignment across all groups. TR 1227 evaluated Phoneme Segmenting items against CCSS RF Standard 2 (Phonological Awareness), finding strong alignment at Grade 1 (RF.1.2D) with approximately 18% of items flagged as problematic at Kindergarten. TR 1303 evaluated Word Reading Fluency items against RF Standard 3 (Phonics and Word Recognition), documenting strong alignment at K–2 and weaker alignment to multisyllabic and morphological decoding at Grade 3. Content accuracy was confirmed through teacher review prior to piloting.

1.2.c The range of item types and cognitive demand among test events is sufficient to strategically assess the depth and complexity of CCR standards.

Item stimuli reach depth and complexity of CCR standards; there is an appropriate range of cognitive demand; item types and demand align to blueprints; writing claims require extended constructed-response items. TR 2101 documents that easyCBM® reading measures sample skills “from automaticity to depth of understanding.” Timed oral production measures (LNF, LSF, PS, WRF, PRF) assess automaticity and decoding fluency at lower cognitive demand. Proficient Reading comprehension items assess literal, inferential, and evaluative comprehension, representing increasing cognitive demand through the grades, with items mapped to Key Ideas & Details and Craft & Structure strands (Tables 3–6). TR 2101 explicitly states that Writing, Speaking & Listening, and most Language standards are not assessed by easyCBM®, reflecting the CBM design scope. Items are timed oral production or non-timed multiple-choice; no extended constructed-response or writing tasks are included. This distribution is appropriate for the CBM screening rationale but represents an inherent constraint on the full range of cognitive demand required by CCR standards.

1.2.d The assessment is aligned to the reading expectations of CCR standards.

Reading items require textual evidence and close reading; vocabulary items reflect CCR requirements including Tier 2 words and context use. Proficient Reading comprehension items assess literal, inferential, and evaluative comprehension across Grades 2–8, with higher grades requiring citation of text evidence aligned to RL 6.1, RL 7.1, and RL 8.1 (Table 5), consistent with the CCR emphasis on close reading. TR 2101’s strand subtotals (Table 4) show Key Ideas & Details as the dominant strand across all grades, with Craft & Structure items appearing beginning at Grade 5. Basic Reading provides five prompts per form across Informational Text, Read to Perform a Task, and Short

Literary Text formats. Vocabulary items assess grade-level words from a wide range of content materials, field-tested and aligned to grade-level standards; Spanish Vocabulary includes fill-in-the-blank sentences measuring word meaning in context. Integration of Knowledge and Ideas standards are underrepresented relative to their CCR emphasis.

1.2.e The assessment is aligned to the writing, research, and language expectations of CCR standards.

Writing prompts represent CCR distribution; writing tasks are text-based; language score points are in written responses; conventions focus on college and career readiness. TR 2101 explicitly states that Writing, Speaking & Listening, and most Language standards are not assessed by easyCBM® reading measures and are therefore excluded from the alignment review. easyCBM® is a CBM screening and progress monitoring system delivered through timed oral production and multiple-choice formats; it does not include writing prompts, research tasks, text-based writing, constructed-response items, editing tasks, or assessments of language conventions. No score points are awarded for written responses. This design is consistent with the CBM rationale, which prioritizes brief, standardized, and efficiently scored measures of core reading competencies. Writing, research, and language standards as described in CCR do not apply to the current easyCBM® reading measure set, and this indicator is not evaluable based on the scope explicitly defined in TR 2101.

1.2.f Assessment is aligned to the speaking and listening expectations of the CCR standards.

Speaking items assess ability to present findings orally; listening comprehension items assess ability to evaluate text in diverse media forms. TR 2101 explicitly states that Speaking & Listening standards are not assessed by easyCBM® reading measures and are excluded from the alignment review. The assessment does not include speaking tasks requiring oral presentation of findings, collaborative discussion tasks, or listening comprehension items assessing evaluation of text in diverse media formats. Fluency measures (PRF, WRF, LNF, LSF, PS) involve oral reading production scored by the assessor, but these measure oral reading accuracy and rate—not the CCR speaking and listening constructs of presenting, discussing, or listening to diverse media. As a CBM system designed for brief, standardized screening and progress monitoring of reading skills, easyCBM® does not assess speaking or listening competencies. This indicator is therefore not applicable to the current measure set and is not evaluable from the document.

1.2.g Assessment is aligned to foundational skills expectations of reading CCR standards—Grades 3–5.

Prompts and tasks represent CCSS foundational skills distribution; the assessment provides a variety of item formats appropriate for measuring CCR foundational skills. TR 2101 identifies foundational skills measures as having the strongest overall alignment to CCSS among all easyCBM® reading measure types. TR 1227 evaluated Phoneme Segmenting (K–1) against RF Standard 2 (Phonological Awareness): alignment was weak at Kindergarten but strong at Grade 1 for RF.1.2D, with approximately 18% of items identified as problematic. TR 1303 evaluated Word Reading Fluency (K–3) against RF Standard 3 (Phonics and Word Recognition): K alignment was strong (~60% vowel patterns, 74% high-frequency words); Grade 1 showed ~60% regularly spelled one-syllable words; Grade 2 showed ~56% two-syllable long-vowel patterns; Grade 3 showed weaker alignment to multisyllabic and morphological decoding. Item formats—timed oral production of letter names, letter sounds, phonemes, and words—are appropriate for measuring foundational skill automaticity. Coverage extends through Grade 3; Grades 4–5 are not addressed.

ALIGNMENT IN READING

This section of the technical report provides information on various easyCBM® reading measures:

- Reading Skills with early reading (phonemes and words)
- Proficient Reading with literal, inferential, and evaluative comprehension.

Note that the Proficient Reading is used in screening students with disabilities and Basic Reading used in progress monitoring. The reason for this distinction is twofold: (a) when CCSS standards were developed, they provided more sensitive progress information on specific types of text, allowing teachers to target instruction; (b) at the same time, the Proficient Measures were more in line with deep reading of extended text (over 1,000 words) and different types of comprehension more in line with the National Panel on Reading. This section presents data from three Technical Reports posted on the web site for Behavioral Research and Teaching (BRT): https://brtprojects.org where they can be retrieved by placing the TR# into the search screen and downloaded. Note that more specific detail can be obtained from these reports as the findings reported in this summary are exemplary.

Technical Report 1227 (2012) – Phoneme Segmenting Alignment with the Common Core Foundational Skills Standard Two: Grades K-1.
Technical Report 1303 (2012) – Alignment with the Common Core State Standards (CCSS): easyCBM® K-3 Word Reading.
Technical Report 2101 (2021) – The Alignment between easyCBM® Mathematics and Literacy Assessments and State and National Standards.

Reading Skills

Summary of Technical Report 1227: Phoneme Segmenting Alignment with the Common Core Foundational Skills Standard Two: Grades K–1 (Sáez et al., 2012).

This study examined the alignment of the easyCBM® Phoneme Segmenting assessment with the Common Core State Standards (CCSS) Reading Foundational Skills Standard 2 for kindergarten and first grade. The purpose was to determine the extent to which existing phoneme segmenting items, developed prior to adoption of the CCSS, reflect grade-level expectations for phonological awareness.

Participants were nine educators, including kindergarten teachers, first-grade teachers, and special education teachers from multiple states, with varied teaching experience. Teachers were trained via webinar and then reviewed test items using an online platform. Data collection involved item-level analyses in which reviewers identified the number of syllables and phonemes in each word and decomposed words into their component sound units. A total of 169 unique kindergarten words and 158 unique first-grade words were analyzed, representing nearly all words used across assessment forms.

Rather than relying on traditional alignment ratings, the study used descriptive and agreement-based analyses to evaluate alignment. Proportions of single-syllable words, phoneme counts, and consonant–vowel–consonant (CVC) structures were calculated and compared to CCSS expectations. Reviewer agreement rates were computed for syllable and phoneme identification, both with and without a low-accuracy reviewer.

Results indicated weak or absent alignment between kindergarten phoneme segmenting items and CCSS Standard 2 expectations, as few items met the three-phoneme CVC requirement. In contrast, strong alignment was found for first grade Standard RF.1.2D, with approximately 80% of words being single-syllable and requiring full phoneme segmentation. Alignment with other first-grade components was weak. Approximately 18% of words were identified as problematic due to excessive difficulty, pronunciation ambiguity, or limited instructional relevance.

Table 1. Alignment and Analysis of Phoneme Segmenting as Foundations Skills

Aspect	Kindergarten	First Grade
Alignment to CCSS Standard 2	Weak or none	Strong for RF.1.2D only
Word Structure	Many exceed 3 phonemes	Mostly single syllable
Reviewer Agreement	Moderate (high after adjustment)	Moderate (high after adjustment)
Problematic Items	Present	Present (≈18% overall)

Reference

Sáez, L., Irvin, P. S., Alonzo, J., & Tindal, G. (2012*). Phoneme Segmenting Alignment with the Common Core Foundational Skills Standard Two: Grades K–1 (Technical Report 1227)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1303: Alignment with the Common Core State Standards: easyCBM® K-3 Word Reading (Sáez et al., 2013).

This study examined the alignment between the easyCBM® kindergarten through grade 3 (K–3) Word Reading assessment and the Common Core State Standards (CCSS) in English Language Arts, specifically Foundational Skills Standard 3 (Phonics and Word Recognition). The purpose was to determine whether the types of words included in the assessment reflected grade-level expectations for word reading skills.

The subjects for the alignment study were 517 unique word items drawn from existing easyCBM® Word Reading forms (57 kindergarten, 117 grade 1, 172, Grade 2, and 171 grade 3 words). Fifteen experienced teachers from multiple U.S. states participated as expert reviewers, representing kindergarten, grades 1–3, and special education backgrounds. Reviewers were trained via webinar and completed online reviews using a secure Distributive Item Review (DIR) system.

Data collection involved having teachers identify specific word characteristics for each item, including syllable structure, vowel patterns, consonant digraphs, inflectional endings, irregular spellings, and word frequency. Rather than using holistic alignment ratings, alignment was quantified by calculating the frequency with which CCSS-aligned subskills appeared across items. Percent correct agreement among reviewers was also computed to evaluate consistency and certainty of judgments.

Results showed strongest alignment at kindergarten, where approximately 60% of items reflected expected vowel patterns and 74% were among the 500 most frequent words in print. Grade 1 items aligned most strongly with decoding regularly spelled one-syllable words (about 60%), while grade 2 items showed the strongest alignment with regularly spelled two-syllable words containing long vowels (about 56%). In contrast, grade 3 items demonstrated relatively weak alignment with CCSS expectations for multisyllabic and morphologically complex word reading. Overall, the findings indicate that easyCBM® Word Reading aligns well with early phonics expectations, with reduced alignment at higher grades.

Table 2. Alignment of easyCBM® Early Reading with Common Core State Standards

Grade	Primary Alignment Focus	Key Findings
Kindergarten	Vowel patterns and high-frequency words	≈60% long/short vowel words; 74% among 500 most frequent words
Grade 1	Regular one-syllable decoding	≈59.8% of items regularly spelled one-syllable words
Grade 2	Regular two-syllable long-vowel decoding	≈55.5% of items aligned with CCSS expectations
Grade 3	Multisyllabic and morphological decoding	Weak alignment with CCSS Standard 3 expectations

Reference

Sáez, L., Irvin, P. S., Alonzo, J., & Tindal, G. (2013). *Alignment of easyCBM® K–3 word reading to the Common Core State Standards (Technical Report 1303).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 2101: The Alignment between easyCBM® Mathematics and **Literacy** Assessments and State and National Standards (Saez et al., 2021).

This section summarizes the alignment coding framework and conclusions for easyCBM® reading measures relative to state English Language Arts standards. easyCBM® measures included Phoneme Segmenting, Letter Names, Letter Sounds, Word and Passage Reading Fluency, Vocabulary, Basic Reading (formerly CCSS Reading), and Proficient Reading (formerly MCRC Reading). CCSS domains not assessed by easyCBM® included Writing, Speaking & Listening, and most Language standards and were excluded from analysis. Alignment evidence is drawn from structured coding in the Reading alignment spreadsheet, organized by CCSS Adopted, CCSS Revised, and State Unique groups. Coding reflects text-based comparisons between state standards and the Common Core State Standards (CCSS), followed by determination of whether those standards are represented within easyCBM® reading benchmarks and progress-monitoring measures.

Each state standard was coded using mutually exclusive CCSS relationship categories: CCSS Exact Match, CCSS Partial Match, CCSS Deviation, and Non-CCSS (ADDITIONAL). In parallel, easyCBM® alignment was coded as Yes or No based on minimum item-coverage criteria. These codes allow interpretation of both the degree of conceptual overlap with CCSS and the extent to which easyCBM® reading measures reflect those standards.

Across CCSS **Adopted** states, reading standards overwhelmingly fell into the CCSS Exact Match category, resulting in a high proportion of Yes codes for easyCBM® alignment. This pattern was most consistent for Foundational Skills (phonological awareness, phonics, word recognition, and fluency), as well as basic reading comprehension expectations.

CCSS **Revised** states exhibited a mix of Exact and Partial Matches, with partial matches typically reflecting differences in wording or instructional framing rather than substantive changes in task demands. easyCBM® alignment remained largely positive in these states, though isolated CCSS Deviation and Non-CCSS outcomes appeared in comprehension and language-related standards.

State **Unique** standards showed the greatest variability. While many standards still mapped as Partial Matches or Deviations, Non-CCSS outcomes were more frequent. In these cases, easyCBM® alignment was more variable, reflecting the intentional focus of easyCBM® on broadly shared reading competencies rather than state-specific extensions or alternative emphases.

Table 3. easyCBM® Reading Alignment Summary by State Group

State Group	easyCBM® Alignment (Yes)	easyCBM® Alignment (No)	CCSS Exact Match	CCSS Partial Match	CCSS Deviation	Non-CCSS / ADDITIONAL
CCSS Adopted	Majority	Few	High	Low–Mod	Low	Minimal
CCSS Revised	Majority	Few	High	Low–Mod	Low	Mod–High
State Unique	Majority	Few	High	Low–Mod	Low	Mod–High

easyCBM® ELA Item Coverage: Strand Subtotals and Domain Totals by Grade

Counts were computed from the grade item columns in ELA_FormsItemsOnly.xlsx. For each CCSS Code row, the number of numeric item labels listed in each grade's Items cell was counted (Grades 2–8). Tables below provide (a) totals by grade, (b) CCSS domain totals by grade, and (c) a single wide table of CCSS Domain and Code rows with items by Grade.

Table 4. Strand-Level Item Subtotals by Grade for Proficient Reading

Strand	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading Literature	28	54	54	54	49	38	55

Table 5. CCSS Domain Item Totals by Grade

CCSS Domain	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6*	Grade 7	Grade 8
Key Ideas & Details	28	54	54	50	49	31	47
Craft & Structure	0	0	0	4	6	7	8
Integration of Knowledge & Ideas	0	0	0	0	0	0	0
Range of Reading & Level of Text Complexity	0	0	0	0	0	0	0

Table 6. CCSS Domain–Code Item Counts by Grade for Basic and Proficient Reading

Strand	CCSS Domain	Code	Grade 2 (28)
Reading Literature	Key Ideas & Details – Five Ws	RL 2.1	16
Reading Literature	Key Ideas & Details – Character Responses	RL 2.3	12
Strand	**CCSS Domain**	**Code**	**Grade 3 (54)**
Reading Literature	Key Ideas & Details – Text Reference	RL 3.1	28
Reading Literature	Key Ideas & Details – Central Message	RL 3.2	9

Reading Literature	Key Ideas & Details – Character Analysis	RL 3.3	17
Strand	**CCSS Domain**	**Code**	**Grade 4 (54)**
Reading Literature	Key Ideas & Details – Explain Text	RL 4.1	25
Reading Literature	Key Ideas & Details – Determine Theme	RL 4.2	12
Reading Literature	Key Ideas & Details – Story Grammar	RL 4.3	17
Strand	**CCSS Domain**	**Code**	**Grade 5 (54)**
Reading Literature	Key Ideas & Details – Text Explanation	RL 5.1	25
Reading Literature	Key Ideas & Details – Determine Theme	RL 5.2	11
Reading Literature	Key Ideas & Details – Compare/Contrast	RL 5.3	14
Reading Literature	Craft & Structure – Word/Phrase Meaning	RL 5.4	4
Strand	**CCSS Domain**	**Code**	**Grade 6 (49)**
Reading Literature	Key Ideas & Details – Cite Text Evidence	RL 6.1	21
Reading Literature	Key Ideas & Details – Determine a Theme	RL 6.2	19
Reading Literature	Key Ideas & Details – Describe a Plot	RL 6.3	9
Reading Literature	Craft & Structure – Perspective	RL 6.6	6
Strand	**CCSS Domain**	**Code**	**Grade 7 (38)**
Reading Literature	Key Ideas & Details – Cite Text Evidence	RL 7.1	6
Reading Literature	Key Ideas & Details – Determine a Theme	RL 7.2	14
Reading Literature	Key Ideas & Details – Story Interactions	RL 7.3	11
Reading Literature	Craft & Structure – Perspective	RL 7.6	7
Strand	**CCSS Domain**	**Code**	**Grade 8 (55)**
Reading Literature	Key Ideas & Details– Cite Text	RL 8.1	24
Reading Literature	Key Ideas & Details – Determine a Theme	RL 8.2	15
Reading Literature	Key Ideas & Details – Story Interactions	RL 8.3	8
Reading Literature	Craft & Structure – Perspective	RL 8.3	8

The easyCBM® literacy assessments demonstrated substantial alignment with state ELA standards, though alignment strength varied by measure, grade level, and strand. Because most easyCBM® ELA measures were developed between 2006 and 2010—prior to the 2009 release of the CCSS—the researchers anticipated and found some areas of weaker alignment, particularly for reading comprehension measures

easyCBM® measures targeting Foundational Skills—including Phoneme Segmenting, Letter Names, Letter Sounds, and Word and Passage Reading Fluency—demonstrated the strongest alignment with state standards. These skills are consistently represented across both CCSS-aligned and state-unique standards, and the easyCBM® items provided broad coverage of the relevant clusters. Alignment ratings for these measures were predominantly Strong across CCSS Adopted states.

Vocabulary and Basic Reading measures, developed in 2013 after the release of the CCSS, showed relatively stronger alignment compared to earlier ELA measures. These measures were explicitly designed to reflect CCSS expectations and demonstrated Moderate to Strong alignment across the 20 CCSS Adopted states and many of the 24 CCSS Revised states.

Alignment for reading comprehension strands (Reading Literature and Reading Informational Texts) was more variable. Because the Proficient Reading and related measures were developed prior to the CCSS, alignment depended heavily on the degree to which item content incidentally reflected CCSS clusters. Some clusters received Moderate or Limited ratings, while others were rated Insufficient due to inadequate item coverage across benchmark forms.

Reference

Sáez, L., Whitney, M., Swanson, D., & Alonzo, J. (2021). *The alignment between easyCBM® mathematics and literacy assessments and state and national standards (Technical Report 2101).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Conclusion

This document summarizes the alignment of easyCBM® reading measures with state and national standards, primarily focusing on the Common Core State Standards (CCSS). The analysis draws from several technical reports examining various reading skills.

Methods: The alignment of easyCBM® reading measures was evaluated using a structured expert review process. Key methods included:

Expert Review: Experienced teachers and specialists reviewed items to determine their alignment with specific CCSS standards and grade-level expectations. This process involved identifying the number of syllables, phoneme counts, and word structures (e.g., CVC).

Coding Framework: A coding framework was used to categorize alignment, including CCSS Exact Match, CCSS Partial Match, CCSS Deviation, and Non-CCSS (Additional). easyCBM® alignment was coded as Yes or No based on minimum item coverage criteria.

Data Analysis: Alignment was quantified by calculating the frequency of CCSS-aligned subskills across items and reviewer agreement rates. This allowed for the assessment of conceptual overlap and the extent to which measures reflect standards.

Focus on Specific Skills: Alignment was examined for foundational skills (phoneme segmenting, letter names, letter sounds, word reading fluency) and for comprehension (vocabulary, basic reading, proficient reading).

Results: The findings indicated generally strong alignment between easyCBM® reading measures and state/national standards, with some variations across skills and grade levels.

Foundational Skills: Measures targeting foundational skills, such as Phoneme Segmenting, Letter Names, Letter Sounds, and Word and Passage Reading Fluency, demonstrated the strongest alignment. For instance, for Phoneme Segmenting, kindergarten items showed weak or absent alignment with CCSS Standard 2, while first-grade items

showed strong alignment with RF.1.2D (approximately 80% of words were single-syllable and required full phoneme segmentation). Word Reading alignment was strongest at kindergarten (around 60% for vowel patterns) and generally good for regularly spelled one-syllable words in Grade 1 (around 59.8%), but showed weaker alignment for multisyllabic and complex word reading in Grade 3.

Comprehension and Vocabulary: Vocabulary and Basic Reading measures, developed after the CCSS release, showed relatively stronger alignment compared to earlier ELA measures. These measures reflected CCSS expectations and demonstrated Moderate to Strong alignment across CCSS Adopted states. However, alignment for reading comprehension strands (Reading Literature and Reading Informational Texts) was more variable. Some clusters received Moderate or Limited ratings due to inadequate item coverage, while others were rated Insufficient.

Cross-Grade Trends: Alignment generally reflected a coherent progression from early number development in K-2 to more complex concepts like fractions, algebra, and geometry in later grades.

Overall, the easyCBM® reading measures demonstrate substantial alignment with state ELA standards, particularly for foundational skills, providing a solid basis for their use in instruction and assessment.

Appendix A: Technical Report Table Titles

Table 1. Alignment and Analysis of Phoneme Segmenting as Foundations Skills

Table 2. Alignment of easyCBM® Early Reading with Common Core State Standards

Table 3. easyCBM® Reading Alignment Summary by State Group

Table 4. Strand-Level Item Subtotals by Grade for Proficient Reading

Table 5. CCSS Domain Item Totals by Grade

Table 6. CCSS Domain–Code Item Counts by Grade for Basic and Proficient Reading

Technical Report References

Sáez, L., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *Phoneme segmenting alignment with the Common Core Foundational Skills Standard Two: Grades K-1 (Technical Report # 1227)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Sáez, L., Irvin, P. S., Alonzo, J., & Tindal, G. (2013). *Alignment with the Common Core State Standards: easyCBM K-3 word reading (Technical Report # 1303)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saez, L., Whitney, M., Swanson, D., & ALonzo, J. (2021). *The alignment between easyCBM® mathematics and literacy assessments and state and national standards (Technical Report # 2101)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Conclusions Supporting Claims for Criterion 1.3: Fairness and Accessibility

The assessment is fair and accessible for all students in the intended test-taking population

This evaluation applies Criterion 1.3 (Indicators 1.3.a–1.3.c) to all sections of the EGCI_1.3 documentation and finds strong, well-substantiated evidence that the assessment system is fair, accessible, and appropriate for its intended test-taking population.

Indicator 1.3.a: Item and test-event development procedures clearly reflect principles of Universal Design for Assessment (UDA). Across all reading measures, layouts minimize construct-irrelevant variance through high-contrast print, uncluttered presentation, consistent typography, and standardized administration. Items undergo systematic content and bias review by experienced educators and researchers, followed by large-scale field testing and psychometric analyses. Documentation explicitly links fairness and accessibility features to the *Standards for Educational and Psychological Testing*, providing a defensible framework for subgroup appropriateness and interpretability.

Indicator 1.3.b: The documentation provides an unusually comprehensive and transparent accommodations framework. The intended test-taking population is clearly defined, and accommodations are carefully distinguished from adaptations and modifications, preserving the validity of score interpretations. Supported accommodations span setting, administration, and response changes and are aligned with national guidance from NCEO, Smarter Balanced, and NAEP. Clear guidance is provided to educators through administration manuals and technical reports, and explicit restrictions are documented for changes that would invalidate construct measurement (e.g., extended time on fluency measures, read-aloud of reading passages). This clarity strongly supports fair and valid use with ALL students.

Indicator 1.3.c: The range of technology-enhanced administration options supports accessibility without compromising validity. Both online and paper-pencil formats are supported, with platform-agnostic guidance and built-in tools such as timers, live scoring, and adjustable presentation features. Digital interfaces are intentionally minimal and non-distracting, while auditory supports are constrained to directions where appropriate, maintaining construct integrity.

Overall, EGCI_1.3 demonstrates a rigorous, standards-aligned approach to fairness and accessibility, with strong procedural, technical, and documentation evidence supporting equitable access and defensible score interpretations across diverse student populations.

1.3a Reading Measurement Presentation Provides Fair Access

This section provides comprehensive, stand-alone technical documentation for easyCBM® reading measures. Each measure description is aligned explicitly to numbered Standards from the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014)[1]. The section supports technical review, procurement evaluation, research transparency, and validity arguments for screening, benchmarking, and progress-monitoring uses within MTSS frameworks. Note: easyCBM® reading assessments include timed, individually administered measures of early literacy (Letter Names, Letter Sounds, Phoneme Segmenting) and reading fluency (Word Reading Fluency and Passage Reading Fluency), as well as untimed, group-administered multiple-choice measures of vocabulary (Vocabulary) and comprehension (Basic Reading and Proficient Reading).

Table 1. Global Standards Alignment Summary

Standards Domain	Relevant Testing Standards	Application Across Measures
Construct Definition & Interpretation	1.1–1.4	Clear construct articulation; explicit limits on score meaning
Test Design & Content	3.1–3.3	Task formats matched to intended constructs
Fairness & Accessibility	3.6, 7.1–7.4	UDL-informed layout; reduced construct-irrelevant variance
Administration & Scoring	4.1–4.13	Standardized procedures; objective scoring rules
Reliability & Precision	2.1–2.3	Designed for repeated measurement and growth modeling
Validity Evidence	1.0, 2.0	Supports content, relations, and consequences evidence

Letter Names (LN)

- Grades: Kindergarten
- Construct definition. Accuracy and fluency in identifying uppercase and lowercase letter names.
- Task and format. Single-page grid of randomly ordered letters in large sans-serif font with generous spacing.
- Administration procedures. Individually administered; one-minute timed oral task.
- Scoring. Total correct letter names per minute.
- Accessibility and fairness. High-contrast print; minimal visual clutter.
- Intended interpretation and use. Early screening and progress monitoring of alphabet knowledge.
- Limitations. Does not assess letter–sound knowledge or decoding.
- Standards alignment. Testing Standards 1.1, 1.3, 3.1, 3.6, 4.1, 5.1, 7.1

Letter Sounds (LS)

- Grades: Kindergarten–Grade 1
- Construct definition. Knowledge of common letter–sound correspondences.
- Task and format. Grid of lowercase letters; oral sound production.
- Administration procedures. One-minute individually administered task.
- Scoring. Correct sounds per minute.
- Accessibility and fairness. Developmentally appropriate font and spacing.
- Intended interpretation and use. Screening and monitoring phonics-related skills.

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

- Limitations. Does not measure application in connected text.
- Standards alignment. Testing Standards 1.1, 3.1, 4.1, 5.1, 7.2

Phoneme Segmentation (PS)

- Grades: Kindergarten–Grade 1
- Construct definition. Ability to segment spoken words into phonemes.
- Task and format. Auditory-only presentation; oral responses.
- Administration procedures. Assessor-delivered stimuli; standardized scoring.
- Scoring. Total phonemes segmented correctly.
- Accessibility and fairness. Eliminates visual barriers.
- Intended interpretation and use. Diagnostic and screening use for phonological awareness.
- Limitations. Not a decoding or print-based measure.
- Standards alignment. Testing Standards 1.1, 3.2, 4.2, 6.1, 7.3

Word Reading Fluency (WRF)

- Grades: Grades 1–3
- Construct definition. Accuracy and rate of isolated word reading.
- Task and format. List of grade-level words in clear vertical format.
- Administration procedures. One-minute timed oral reading.
- Scoring. Words correct per minute.
- Accessibility and fairness. Uncluttered layout reduces extraneous demands.
- Intended interpretation and use. Monitoring decoding development.
- Limitations. No comprehension inference.
- Standards alignment. Testing Standards 1.1, 3.1, 4.10, 5.1

Passage Reading Fluency (PRF)

- Grades: Grades 1–8
- Construct definition. Oral reading accuracy and rate in connected text.
- Task and format. Single-page passage; one-column layout.
- Administration procedures. Individually administered; timed.
- Scoring. Words Correct Per Minute (WCPM).
- Accessibility and fairness. High-contrast print; consistent spacing.
- Intended interpretation and use. Screening, benchmarking, progress monitoring.
- Limitations. Not a direct comprehension measure.
- Standards alignment. Testing Standards 1.1, 3.1, 4.1, 4.13, 5.1

Vocabulary (Online and Paper-Pencil)

- Grades: Grades 3–8
- Construct definition. Receptive vocabulary knowledge.
- Task and format. Multiple-choice items; minimal graphics.
- Administration procedures. Group or individual administration.
- Scoring. Total correct responses.
- Accessibility and fairness. Readable typography; reduced load.
- Intended interpretation and use. Instructional planning and screening.
- Limitations. Depth of knowledge not assessed.
- Standards alignment. Testing Standards 1.1, 3.2, 4.8, 7.2

Basic Reading Comprehension (Online and Paper-Pencil)

- Grades: Grades 3–8
- Construct definition. Literal understanding of text.
- Task and format. Silent reading with MC items.
- Administration procedures. Standardized administration.
- Scoring. Total correct.
- Accessibility and fairness. Layout minimizes navigation demands.
- Intended interpretation and use. Foundational comprehension screening.
- Limitations. Does not assess inference.
- Standards alignment. Testing Standards 1.1, 3.1, 4.1, 7.1

Proficient Reading Comprehension (Online and Paper-Pencil)

- Grades: Grades 3–8
- Construct definition. Inferential and integrative comprehension.
- Task and format. Parallel format to Basic Comprehension.
- Administration procedures. Standardized procedures.
- Scoring. Total correct.
- Accessibility and fairness. Consistent layout supports fairness.
- Intended interpretation and use. Grade-level proficiency decisions.
- Limitations. Interpret with complementary measures.
- Standards alignment. Testing Standards 1.1, 3.2, 4.8, 7.4

Validity Argument Framework (Messick / Kane)

- Content evidence: Tasks align with clearly articulated reading constructs across development.
- Response processes: Simple, transparent response demands reduce construct-irrelevant variance.
- Internal structure: Measures designed for reliable repeated use and growth modeling.
- Relations to other variables: Intended for linkage with external reading outcomes.
- Consequences: Supports early identification and instructional decision-making.

1.3b Reading Test Accessibility and Accommodations in easyCBM® (From Technical Report # 2510)

The *easyCBM®* assessments were designed using Universal Design for Assessment (UDA) principles. According to Rose (2006)[2], three guiding principles of UDL allow different ways for students to succeed that provide multiple means of representation, expression, and engagement. Their purpose is to ensure fairness, accessibility, and validity for all students, including those with disabilities and English Language Learners (ELLs).

Furthermore, *easyCBM®* follows the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Experienced teachers wrote the test items, which University of Oregon researchers reviewed for content and bias. Items were field-tested nationwide, and Item Response Theory (IRT) analyses determined item difficulty, discrimination, and precision. These analyses guided the creation of equivalent alternate forms for reading and math measures, supporting consistent screening and progress monitoring, and reliability and validity studies confirmed the measures' strong psychometric properties. Technical reports are available through Behavioral Research & Teaching (BRT) and ERIC, with further findings summarized in Swanson & Tindal (2024).

[2] Rose, D. H. (2006). *A Practical Reader in Universal Design for Learning*. Harvard Education Press.

Tindal (2025)[3] emphasizes that standardization ensures fairness, but strict uniformity can limit access for students with disabilities or diverse linguistic and cultural backgrounds. Equity requires flexible testing conditions that preserve validity while removing irrelevant barriers. *easyCBM®* incorporates measurement changes across three areas: (1) setting—changes in lighting, seating, or environment [S]; (2) administration—repeating directions or using visual cues [A]; and (3) response—alternative response methods for selection tasks [R]. These changes maintain fairness without altering what is measured. Distinguishing between informal "adaptations" and formal "accommodations" (listed in IEPs), *easyCBM®* achieves a balance between standardization and flexibility—ensuring all students can accurately demonstrate what they know. We have divided all test changes into three types, varying on a continuum of significance/impact: (a) adaptations considered as universally designated supports that are part of a student's instructional practices, (b) accommodations listed in most state policies and often requiring recommendations for use by an IEP team, and (c) modifications that represent changes to the validity of inferences made from the score.

[3] Tindal, G. (2025). *Rethinking "Standardization" for NAEP to Increase Equity and Access*. University of Oregon: Behavioral Research and Teaching. Technical Report 2510.

The following test changes in *easyCBM®* testing can be used without reference to IEP teams but should be considered with due attention to state testing programs, which may or may not allow these practices. They are considered as extensions of UDA and have been curated from National Center on Educational Outcomes (NCEO) with further changes presented in tables from the Smarter Balanced State Consortium and National Assessment of Educational Progress (NAEP), both digital and paper pencil administration. The comprehensive list is referenced in the technical report by Tindal (2025).

Table 2. Test Changes as Adaptations or Extensions of UDA in Setting [S], Administration [A], or Response [R]

Adaptations	Type	Definition and Clarification
Clarify/Simplify/Explain/Repeat Directions/ Cue to stay on task	A	Clarify/Simplify/Explain/Repeat directions ensures that students hear the directions and problems accurately. Cue/prompt responses as needed.
Color Contrast/Overlays/Templates	A	Color Contrast or Overlays/Templates provide students better access to printed text or text on screens and therefore allow them to better understand and interpret the problem or item.
Extended time (if not fluency measures)	SA	Extended time may function like breaks, allowing students to take more time in reading the problems, reviewing the options (for selected response types), or composing responses in production tasks. This change does not include fluency measures.
Familiar proctor/test administrator	A	Familiar proctor/test administrator provides the student a person with whom they have experience (perhaps in the directions being read in a more appropriate manner or in the prompting to move along) to increase access to a wider range of problems.
Highlighting/Masking	A	Highlighting or masking allows students to make critical text stand out more from other text, thereby reducing the need to mentally sort/focus on critical content.
Large print	A	Presentation of assessment forms in print greater than standard 12- or 14-point font makes the materials more visible. Note that 'Zoom' can be used in the digital environment.
Magnification / Zoom	A	Magnification is a simple strategy to ensure that the student can see/read the item.
Manipulatives/Abacus	A	Manipulatives or Abacus may allow students to organize, sort, or count objects to represent the problem concretely (does not include 'numbers' tables).
Mathematics crib sheet	A	Mathematics crib sheets may function as a scaffold to ensure accurate information can be retrieved (e.g., conversion of measures across different metrics).
Multiple days or breaking test sessions	SA	Multiple days or breaking up the length of testing sessions may allow students to avoid fatigue and maintain attention. This change does not apply to fluency measures.
Noise reduction (buffer)/Volume Adjustment/Amplification	A	Noise reduction (buffers) or Volume Adjustment/Amplification can occur in any manner that allows students to maintain attention, by either reducing excess noise, providing white noise, or playing music from the student's playlist.
Paper format/Print on demand	A	Paper format or print on demand may ensure that students can see/read items and problems without having to access content displayed on a computer screen and potentially reduce glare or avoid scrolling.

Preferential seating	S	Preferential seating may reduce student anxiety or ensure that directions are heard (e.g., sitting in the front of the room).
Read aloud by humans or computers	SA	Read aloud provides students access to problems and options that otherwise may be not read or misread. Includes student reading aloud. Note: Only the Directions can be read aloud in Proficient and Basic Reading measures.
Recorded oral responses	R	Students may have their responses recorded for teachers to score later. This adjustment applies only to production responses (e.g., Letter Names or Sounds, Phoneme Segmenting, either Word or Passage Reading Fluency).
Scratchwork paper / Mark up booklets (e.g., eliminating options).	SAR	Students can have scratch paper when taking any of the measures, so they can record notes or test taking strategies.
Scribe	R	Another person records the student's response. Note that this change only applies to production written responses (e.g., mathematics problems).
Small group or individual administration	S	Small-group and individual administration potentially provide students a less distracting environment.
Speech-to-text / closed captioning	A	(Recorded) Directions can be presented via text so the student hears and sees them. This change includes use of closed captions.
Tactile graphics	A	Graphics can be presented in a tactile format to the student (primarily used in mathematics).
Text-to-speech (computer generated voice)	A	Text-to-speech (computer-generated voice) provides students the directions, which might otherwise be misread (or misinterpreted) by the student. Note: This change follows the same restrictions as Read Aloud (directions only in reading comprehension but allowance in mathematics).
Translations or different language versions.	A	Multiple languages (e.g., English or Spanish) may be used in translations of test directions. Note: If the test is a measure of the primary language, then items and problems must be presented in that language.

Table 3. Test Changes Noted in the IEP as Allowed or Required in Setting [S], Administration [A], or Response [R]

Accommodation	Type	Definition and Clarification
Assistive Technology	SAR	Any assistive devices that allow facilitative changes in presentation to or responses from students.
Braille	AR	Universal English Braille (UEB) is the preferred type of Braille, but whatever presentation type is presented, it should be consistent with that used in instruction.
Signed administration	AR	Signed administration (or response) is designed so that students with hearing impairments or who are deaf can participate in taking most *easyCBM®* measures (those that use multiple choice items). However, for fluency measures that require the student to read out loud, the loss or limit of hearing may impede clear enunciation and make this measure difficult to score; therefore, these measures should not be administered.
Medical Supports	SAR	Various medical or prosthetic supports provided during the test administration.

Some test changes are neither adaptations (universally designed adjustments) nor accommodations but are modifications: They invalidate the score and any decisions that can be made from the student's performance. The reason for this invalidation is that the construct being measured is changed, along with the change in setting (S), administration (A), or response (R).

Table 4. Test Changes Not Allowed in Setting [S], Administration [A], or Response [R]

Modification	Type	Definition and Clarification
Extended Time for Fluency Measures*	A	Extending time for fluency measures misrepresents this construct because the measures are timed for 1-minute to obtain a measure of rate.
Read Aloud of Comprehension Measures	A	Read aloud of passages and questions (with options) changes the construct to 'listening comprehension' and misrepresents the construct of 'reading comprehension'. Note that the test directions may be read aloud.
Calculator for Math Measures	A	Allowing students to use a calculator represents a modification of math problem solving when items require explicit operational problems (e.g., 12 ÷ 4 = __). Although some *easyCBM®* math problems are not explicit operations, allowance of calculators becomes problematic in implementation and oversight.

**Fluency measures: Letter Names, Letter Sounds, Phoneme Segmenting, Word Reading Fluency, Passage Reading Fluency, all Spanish measures except for Vocabulary, and Dyslexia Identification measures (Rapid Automatic Naming and Pseudo-Words).*

Conclusion

The recommendations of these test adaptations with easyCBM® are based on a technical report[4] written under a contract with the NAEP Validity Studies Panel and scheduled to be published in October 2024 when the Trump administration canceled the contract through the American Institutes of Research (AIR). The research upon which the report is based represents a nearly exhaustive review of both published refereed publications and technical reports from various professional organizations over the past 40 years. Before the report was completed, the major conclusions and perspectives had been reviewed by the Panel with eventual agreement that NAEP may need to review and possibly revise its policies on accommodations based on the empirical and logical analysis presented in the report that included the following areas addressed:

- Research on Test Accommodations from NAEP as well as from meta-analyses conducted and published for students with disabilities and English language learners.
- Analyses and summaries of (in)consistencies among state practices.
- Impact of test adaptations on students with disabilities.
- Speculations on form versus function within a behavioral perspective.
- See Lite version for free and released test items.

1.3c UDL with Technology Enhanced Administration and Scoring

Assessment Administration User's Manual (Page 18)

With easyCBM®, educators have the option of assessing via computer-based administration or with paper and pencil.

The untimed tests (Vocabulary, Basic Reading, Proficient Reading, Basic Math, Proficient Math) may be given in a group setting and are optimized for student online testing on a computer or tablet. When a test is administered online, you simply need to monitor the testing environment while students are working. The online testing is student-paced, though we recommend establishing maximum administration times for the benchmark assessments (see Benchmark Testing Time Required by Subject table below) for all students except those whose IEPs indicate they should receive extended time as an accommodation on tests. When students log in during a benchmark window, they will automatically have access to the appropriate benchmark test forms. For progress monitoring tests, you will need to assign the forms in advance.

The timed fluency measures (the early literacy and fluency measures for both English and Spanish) must be administered individually (one-on-one) with each student. You can conduct live scoring by entering student responses (item-level data) directly into the easyCBM® system via a computer or tablet while administering the assessment. A built-in timer is provided for ease of use. We recommend a device with touch screen technology for live scoring, so that students are not aware when an error is being marked (audible 'clicks' to mark errors introduce potential measurement error and should be avoided).

If computer-based administration is not an option, all assessments can be administered on paper-and-pencil, and the results can be entered into easyCBM® after the testing session. The student and assessor (teacher) copies of the test forms are available as PDF files for teachers to download and print. The benchmark measures are organized by grade, benchmark assessment period, and assessment content area. Progress monitoring measures are organized by grade and measure type.

---

[4] Tindal, G. (2025). *Rethinking "Standardization" for NAEP to Increase Equity and Access.* University of Oregon: Behavioral Research and Teaching. Technical Report 2510.

Table 5. Administration Methods – Online and Paper-Pencil

easyCBM Test Administration Methods

Administration Method	Description	Measures Available	
Student online testing	Students take these measures online through the easyCBM website.	Basic Reading Proficient Reading Vocabulary	Basic Math Proficient Math Spanish Vocabulary
Live scoring	Enter student responses into easyCBM as you administer the assessment to the student. This option is available for individually administered measures.	Letter Names Letter Sounds Phoneme Segmenting Word Reading Fluency Passage Reading Fluency	Spanish Syllable Sounds Spanish Syllable Segmenting Spanish Word Reading Fluency Spanish Sentence Reading Fluency
Paper-and-pencil with item-level data	Test students on paper-and-pencil, then enter student responses into easyCBM.	All measures	
Paper-and-pencil with total scores only	Test students on paper-and-pencil, then enter total scores into easyCBM.	All measures, but for benchmark testing only	

User's Manual (page 19)

The amount of testing time required varies by grade and measure. See the table below for the typical amount of time needed for benchmark administration. Please note that these times do not include reading directions or transitioning between tests.

Table 6. Testing Times

Benchmark Testing Time Required by Subject

	Grade	Season	Individually Administered Measures (time per student)		Group-Administered Measures (approximate time per group)	
Reading	K	Fall	LN, LS, PS	3 min	-	
		Winter	LS, PS, WRF	3 min	-	
		Spring	LS, PS, WRF	3 min	-	
	1	Fall	LS, PS, WRF	3 min	-	
		Winter	LS, WRF, PRF	3 min	-	
		Spring	LS, WRF, PRF	3 min	-	
	2*	All	PRF	1 min	Vocabulary Proficient Reading	10 min 20 min
	3–8*	All	PRF	1 min	Vocabulary Proficient Reading	15 min 45 min
Math	K–2*	All	-		Proficient Math	20 min
	3–8*	All	-		Proficient Math	30 min
Spanish	K	All	SP_SSEG, SP_SS	2 min	-	
	1	All	SP_SSEG, SP_SS, SP_WR, SP_SR	4 min	-	
	2	All	SP_WR, SP_SR	2 min	Vocabulary	10 min
	3–8	All	-		Vocabulary	10 min

** Note: In Grades 3 through 8, the optional group-administered Basic Reading assessment takes about 30 minutes to complete. The optional group-administered Basic Math assessment, available for Kindergarten through Grade 8, takes approximately 30 minutes to complete.*

Appendix A: Technical Report Table Titles

Table 1. Global Standards Alignment Summary

Table 2. Test Changes as Adaptations or Extensions of UDA in Setting [S], Administration [A], or Response [R]

Table 3. Test Changes Noted in the IEP as Allowed or Required in Setting [S], Administration [A], or Response [R]

Table 4. Test Changes Not Allowed in Setting [S], Administration [A], or Response [R]

Table 5. Administration Methods – Online and Paper-Pencil

Table 6. Testing Times

Technical Report References

easyCBM User's Manual (2025). Eugene, OR: University of Oregon – Behavioral Research and Teaching.

Conclusions Supporting Claims for Criterion 2.1: Overall Achievement

The interim assessment provides for valid inferences about a student's current overall achievement in the target content domain.

2.1.a Item and form development procedures result in high-quality test events.

Item development, review, and piloting procedures and materials are designed to ensure all newly developed items meet technical quality standards. The document describes a multi-stage approach to developing and evaluating easyCBM® reading items and forms. For fluency measures (LN, LS, PS, WRF, PRF), forms were constructed from common blueprints specifying content balance, difficulty range, and timing rules, then administered under standardized conditions. Technical Report 1403 describes 15 trained research assistants who received structured training and periodic refresher sessions to ensure standardized administration and scoring. Items are evaluated through large-scale piloting: TR 1003 used samples of nearly 2,000 Kindergarten students and TR 1004 used 2,100–2,300 Grade 2 students per measure across three benchmark seasons. These designs provide the response data necessary to evaluate item and form functioning before operational deployment, ensuring that all forms produce dependable scores prior to use.

Assessment design specifications and test development and review procedures ensure test events meet content and statistical quality criteria. Statistical quality criteria are applied systematically across the technical reports. For comprehension measures, Cronbach's alpha, split-half methods, and repeated-measures ANOVA evaluate form equivalence and score consistency. Rasch modeling provides fit statistics to identify items with unexpected response patterns. Generalizability theory analyses partition variance across persons, forms, and occasions, confirming that the largest share reflects student differences—a desirable pattern indicating forms measure stable reading skill rather than passage-specific effects. Decision studies specify how many administrations are needed to meet reliability thresholds for relative versus absolute decisions. Content quality is maintained through blueprint-based alternate-form construction that controls content balance, difficulty distribution, and timing, ensuring test events are both content- and statistically sound.

2.1.b Achievement scores are reliable.

Item/test development and review procedures facilitate the reliability of test scores. Item and form development practices are explicitly oriented toward producing reliable scores. Fluency forms are constructed from a common blueprint—content balance, difficulty range, and timing constraints—and administered by trained examiners under standardized conditions, both of which reduce measurement error. The document notes that strong alternate-form consistency for fluency measures reflects these controlled development choices: when form construction is rigorous, observed score differences more credibly reflect true student performance. For comprehension, the document explicitly identifies test length as a reliability constraint and recommends optimizing item targeting and increasing test information as design-level remedies, directly linking development decisions to reliability outcomes. Review procedures across technical reports are therefore reliability-aware by design.

Procedures for calculating and evaluating reliability are well-documented and appropriate. Multiple reliability methods are documented and applied, each matched to measure type and intended use. Alternate-form reliability (Pearson and Spearman rank correlations) is the primary evidence for fluency measures, complemented by test–retest correlations at one-week intervals. Internal consistency (Cronbach's alpha, split-half with Spearman–Brown and Guttman corrections) is reported for comprehension and vocabulary measures. Generalizability and Phi coefficients accompany D-study results for fluency measures. Standard errors of measurement are calculated from median reliability coefficients and average standard deviations. Reliability of growth (slope) is estimated via two-level hierarchical linear growth models across fall–winter–spring, reported by fall-performance quartile. This multi-method approach is appropriate to the range of measures and uses documented throughout the report.

Obtained reliability indices and estimates of precision are at an appropriate level to support the use of results as intended. Obtained reliability estimates are strong for fluency measures and acceptable for other measure types. Alternate-form reliability for Passage Reading Fluency is consistently .87–.97 across Grades 1–8 (median

coefficients: .93–.96). Word Reading Fluency alternate-form correlations range from .84–.97 with generalizability coefficients of .97–.99. Letter Names and Letter Sounds yield median alternate-form coefficients near .77–.85. Phoneme Segmenting is more variable (.35–.89 across studies), and comprehension alternate-form correlations are more modest (.26–.59), with internal consistency (Cronbach's alpha) for vocabulary in the acceptable-to-strong range. Growth reliability for WRF and PRF in Grades 1–2 is generally .70–.93 across quartiles, though 4th-quartile students show lower slope reliability attributed to ceiling effects. Together, these values support fluency measures for progress monitoring and benchmark screening, while comprehension benefits from supplementary administrations or additional items where higher precision is needed.

Reliability

Reliability refers to the degree to which an assessment yields scores that are consistent, precise, and reproducible for a defined purpose, population, and set of testing conditions. In educational measurement, reliability is not a fixed property of a test "in general"; it is an empirical characteristic of scores produced in a particular administration and interpreted for decisions. When reliability is high, observed score differences are more likely to reflect true differences in student performance on the construct being measured rather than random fluctuations in testing conditions, item sampling, or scoring.

Reliability is often described using observed-score theory, where an observed score is viewed as a combination of a true score plus measurement error. Error can arise from many sources: sampling a limited set of items from a broad content domain; day-to-day variability in student attention, fatigue, or motivation; differences in administration conditions (time limits, directions, access to tools, testing environment); and differences in scoring procedures. Because no single study isolates every potential source of error at once, educational assessment programs typically assemble a "reliability argument" by reporting multiple indices that are well matched to how scores will be used (e.g., screening classification, progress monitoring growth, or program evaluation).

Types of Reliability

Alternate-form reliability (parallel-forms reliability) evaluates the consistency of scores across two equivalent forms designed to measure the same content and skills. Alternate forms are constructed from a common blueprint—content balance, difficulty range, item formats, time limits, and scoring rules—and administered under comparable conditions, often close in time. The correlation between Form A and Form B scores provides evidence that students would obtain similar results regardless of which form they received. In progress monitoring, alternate-form evidence is especially important because students are tested repeatedly and score change should reflect learning rather than differences among forms.

Test–retest reliability evaluates score stability over time by administering the same form (or a highly similar form) to the same students on two occasions separated by a specified interval. The resulting correlation estimates the extent to which students maintain their relative standing over that time window. Test–retest evidence is most informative when the construct is expected to be relatively stable across the retest interval and when administration conditions are held constant. If the interval is long enough for meaningful learning to occur, stability can decrease for substantive reasons; therefore, test–retest designs require careful interpretation and an interval aligned to the intended use (e.g., short-term stability vs. sensitivity to growth).

Internal consistency reliability addresses the extent to which items within a single form function together to measure a common underlying construct. For dichotomously scored items common in mathematics screeners, Cronbach's alpha is frequently reported as an index of how well items cohere and how much precision is gained by aggregating across items. Internal consistency tends to increase when a test includes more informative items, when item difficulties are appropriately distributed (not all too easy or too hard), and when items discriminate among students in the target achievement range. Internal consistency is typically strongest for the total score and lower for short subtests, because fewer items provide less opportunity to average out item-sampling error.

Inter-judge (inter-rater) reliability applies when human judgment contributes to scoring—for example, scoring constructed responses, applying rubric criteria, or recording errors during performance tasks. It quantifies the extent to which different scorers assign the same score to the same student work under standardized scoring rules. Depending on the score scale and scoring design, inter-judge reliability may be estimated using percent agreement, Cohen's kappa, intraclass correlation coefficients, or generalizability approaches. Strong inter-judge reliability is essential because inconsistent scoring introduces error that can overwhelm the precision gained from well-designed items.

These types of reliability are complementary rather than competing. Benchmark screening often prioritizes internal consistency to support dependable total scores at a single time point, while progress monitoring prioritizes alternate-form evidence to ensure equivalence across repeated administrations. Test–retest evidence clarifies short-term stability when the same form is reused or when interpreting change over brief windows. When scoring involves judgment, inter-judge evidence becomes essential because it can limit the maximum reliability attainable by any other design.

Why Reliability is Important

Reliable scores are essential because educational decisions require separating true differences in student performance from random error. When reliability is low, screening cut scores and growth targets become unstable, confidence intervals widen, and students may be misclassified—either missing needed support or being placed in interventions unnecessarily. Reliable measurement reduces overreaction to chance score swings, strengthens progress-monitoring decisions, and improves fairness by providing consistent information across classrooms, schools, and testing occasions. It also increases statistical power for evaluating programs and policies and supports more credible validity arguments about what scores mean and how they should be used[1].

This technical report focuses on reliability of outcomes from different easyCBMs® used to document early reading performance: The summaries and findings represent compilations from previously published technical reports so that reviewers can use this document as a 'quick guide' and then access the full report at the website for Behavioral Research and Teaching (BRT: https://brtprojects.org). Several reliability estimates are provided for different easyCBMs® used to document early reading performance: **Note**: All tables and figures in this summary are examples of those presented in full within the individual Technical Reports but are not exhaustive, just illustrative: The summaries and findings represent compilations from previously published technical reports so that reviewers can use this document as a 'quick guide' and then access the full report at the website for Behavioral Research and Teaching (BRT: https://brtprojects.org). The measures included Letter Names (LN), Letter Sounds (LS), Phoneme Segmenting (PS), Word Reading Fluency (WRF), and Passage Reading Fluency (PRF). Reliability analyses included documenting internal consistency, alternate form, test-retest, and generalizability. The results indicate positive results for the various measures when administered at different benchmark seasons.

Preview of Technical Report Findings

The studies summarized in this document provide an integrated picture of score dependability for easyCBM® reading measures across grades and measure types. These findings are illustrative: this report synthesizes results from primary studies conducted by Behavioral Research and Teaching (BRT) researchers, with the full set of technical reports available at https://brtprojects.org. Across the body of work, reliability is consistently strongest for individually administered, fluency-based measures (Word Reading Fluency and Passage Reading Fluency), while measures that rely on fewer items, group administration, or broader constructs (e.g., multiple-choice reading comprehension) show more modest consistency and benefit from additional design and analytic supports.

A central cross-cutting finding is that fluency measures tend to yield highly stable scores across both alternate forms and short retest intervals. Across grades and studies, correlations among parallel passages or word lists are typically high, indicating that forms behave comparably and that observed differences are largely attributable to

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

student performance rather than passage difficulty. Similarly, one-week test-retest estimates are generally strong, supporting the interpretation that scores are sufficiently stable to support progress-monitoring decisions under standard conditions. Where coefficients are lower, results suggest that reliability is sensitive to the form set, sample size, and classroom-level testing context—factors managed through careful form construction and administration guidance.

A second integrated finding is that generalizability theory analyses reinforce and sharpen the classical reliability results. When persons, forms, and occasions are modeled simultaneously, the largest share of variance is typically attributable to differences among students, with relatively small contributions from forms and occasions. This is desirable: it indicates that the assessments are primarily measuring stable differences in reading skill rather than being driven by the specific passage or testing day. Decision studies extend this insight by showing how precision improves with additional observations. For many fluency applications, one form on one occasion can meet common reliability thresholds for relative decisions, but dependability increases meaningfully with multiple forms and/or occasions, particularly when the goal is an absolute decision (e.g., classification against a benchmark) or when local samples are small.

A third theme concerns how construct breadth and administration mode interact with reliability. Fluency measures involve timed performance on constrained tasks (word reading or passage reading rate/accuracy) and are scored with relatively objective rules, which supports consistency. In contrast, multiple-choice reading comprehension (MCRC) assesses a broader construct and is often group administered, introducing additional sources of error (e.g., differences in testing conditions, attention, and proctoring). Early evidence shows lower alternate-form correlations for comprehension relative to fluency, underscoring that reliability depends on the nature of the score and the conditions under which it is collected.

A fourth integrated finding is that item- and form-level analyses help explain and improve comprehension reliability. Across grade-specific MCRC studies, multiple methods are used to evaluate form equivalence and score consistency (e.g., repeated-measures ANOVA for form comparisons, split-half estimates, top–bottom group checks, and Rasch modeling). These analyses commonly identify that overall item functioning is acceptable, with most items showing reasonable fit and distractor behavior, while isolated items exhibit misfit or unexpected discrimination patterns. A recurring constraint is test length: when forms contain relatively few items, person reliability can be limited, especially when score distributions show ceiling effects or restricted variability. These patterns suggest that reliability for comprehension can be strengthened by optimizing item targeting, increasing information where score variability is limited, and maintaining consistent administration practices.

A fifth theme comes from internal consistency evidence for vocabulary. Using very large benchmark datasets across grades and seasons, internal consistency estimates (e.g., Cronbach's alpha) are stable and generally in the acceptable-to-strong range, supporting the use of vocabulary measures for screening and instructional planning. The consistency of estimates across fall and winter administrations suggests that the measures function reliably across typical seasonal contexts. Split-half evidence and top–bottom comparisons further indicate that items discriminate as expected, with higher-performing students selecting correct options at higher rates.

Taken together, these studies support the conclusion that easyCBM® reading measures can provide dependable scores for key assessment uses when aligned to appropriate designs. Fluency measures show strong evidence for alternate-form consistency and short-term stability, making them well suited to frequent progress monitoring. Comprehension and other item-based measures demonstrate acceptable reliability when evaluated with multiple complementary methods, but their precision is more sensitive to test length, score distributions, and administration conditions; accordingly, they benefit from continued refinement and from measurement designs that reduce error (e.g., additional items, multiple forms, or repeated occasions when needed). Overall, the integrated evidence base emphasizes a practical message: reliability is not only a coefficient, but a function of how measures are built and used, and these reports provide concrete guidance for selecting and administering measures to achieve dependable decisions.

Across the fluency studies that incorporate D-studies, a consistent implication is that measurement precision can be strategically increased by adding parallel forms and/or occasions, with diminishing returns after a small number of administrations. For relative (norm-referenced) decisions, many analyses show that one form on one occasion is often adequate for PRF, while WRF may benefit more from multiple forms, particularly in contexts where classroom samples are small or where the student-by-form interaction is nontrivial. For absolute (criterion-referenced) interpretations, Phi coefficients and associated SEMs highlight that additional administrations can meaningfully reduce error around cut scores, which is directly relevant when schools make risk classifications tied to interventions.

Another integrated pattern is that "problematic" results, when they occur, tend to be form-level or item-level rather than system-wide. For example, occasional lower correlations for a particular passage form suggest that a single passage may be unusually easy, hard, or sensitive to administration nuances. Similarly, isolated comprehension items may show misfit or unexpected top–bottom patterns, indicating the need for revision or replacement. The broader reliability picture remains positive because most forms and items behave as intended, and the analytic approaches used across the technical reports are designed to detect these exceptions so that form pools can be improved over time.

Table 1. Grade-Level Benchmark Measures by Assessment Period

Grade Level	Fall Benchmark	Winter Benchmark	Spring Benchmark
K	Letter Names Letter Sounds Phoneme Segmenting	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Phoneme Segmenting Word Reading Fluency
1	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency
2	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading

The measures in **Grades 3 through 8** consist of Vocabulary, Passage Reading Fluency, and two measures of comprehension: Basic and Proficient.

Summary of Technical Report 0906: Alternate Form and Test–Retest Reliability of easyCBM® Reading Measures (Alonzo & Tindal, 2009).

Method

Approximately 50 students per grade participated in the study, all drawn from a mid-sized K–8 school in the Pacific Northwest. The school included diverse demographics and had reading performance levels slightly above district and state averages. All students present on testing days participated in the study. Measures examined varied by grade level. In Grade 1, the study included Phoneme Segmenting, Letter Names, Letter Sounds, Word Reading Fluency, and Passage Reading Fluency. In Grade 3, Word Reading Fluency, Passage Reading Fluency, and Multiple-Choice Reading Comprehension were examined. In Grades 5 and 8, Passage Reading Fluency and Reading Comprehension were included. Across grades, multiple alternate forms of each measure were selected, representing roughly 40% of the forms available in the easyCBM® assessment system.

Data collection occurred during two testing sessions administered one week apart. In the first session, students completed several forms of each measure type. In the second session, students repeated selected forms from the first session and completed additional alternate forms. Most measures were individually administered by trained research assistants, while reading comprehension tests were group administered by classroom teachers. Scores were recorded according to standardized scoring rules specific to each assessment type.

Results

Analyses included descriptive statistics, bivariate correlations to estimate alternate-form reliability, and correlations across testing sessions to estimate test–retest reliability. Standard errors of measurement were also calculated using median reliability coefficients and average standard deviations. Results indicated strong alternate-form reliability across most reading fluency measures. In Grade 1, alternate-form correlations ranged from .76 to .97 across measures, with the highest coefficients observed for Word Reading Fluency and Passage Reading Fluency (.95–.97). In Grade 3, Word Reading Fluency correlations ranged from .87 to .93 and Passage Reading Fluency from .94 to .95. Passage Reading Fluency alternate-form correlations ranged from .87 to .96 in Grade 5 and .87 to .95 in Grade 8. Reading comprehension alternate forms produced lower correlations (.26–.59). Test–retest correlations ranged from .45 to .97 across measures and grades, with the highest values observed for Word Reading Fluency and Passage Reading Fluency. Overall, the findings demonstrate strong reliability for the easyCBM® fluency measures across grade levels.

Table 2. Reliability for Various Measures Across Grades

Grade	Measures Studied	Alternate Form Reliability (r)	Test–Retest Reliability (r)
Grade 1	SEG, LN, LS, WRF, PRF	.76 – .97	.45 – .97
Grade 3	WRF, PRF, MCRC	.87 – .95 (fluency) .26 (MCRC)	.92 – .94
Grade 5	PRF, MCRC	.87 – .96 (PRF) .59 (MCRC)	.92 – .94
Grade 8	PRF, MCRC	.87 – .95 (PRF) .35 (MCRC)	.91 – .91

Reference

Alonzo, J., & Tindal, G. (2009). *Alternate form and test-retest reliability of easyCBM® reading measures (Technical Report No. 0906)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1003: Technical Adequacy of the easyCBM® Primary-Level Reading Measures (Grades K–1), 2009–2010 Version (Lai et al., 2010).

Methods

A set of studies examine technical adequacy for the easyCBM® early-reading progress monitoring measures used in Kindergarten and Grade 1 during the 2009–2010 school year. Data were drawn from a convenience sample of three Pacific Northwest districts (68 schools contributing Kindergarten data; 71 schools contributing Grade 1 data). All students present during testing windows participated. Sample sizes were large for each measure (Kindergarten: n≈1,984–1,987 for Segmenting and Letter Sounds; n≈1,985 for Word Reading; Grade 1: n≈953 for Segmenting; n≈2,179–2,181 for Letter Sounds and Word Reading; n≈2,180 for Passage Reading). Measures were individually administered 60-second fluency probes, scored as correct phonemes, letter sounds, or words per minute, across fall, winter, and spring. Reliability of growth (slope) was evaluated using two-level hierarchical linear growth models (time at level 1, student at level 2). Growth reliability (for intercept and slope) was summarized by grade and by initial status groups, primarily quartiles based on fall performance (with Word Reading in Kindergarten summarized by lower vs upper 50th percentile).

Results

Kindergarten. For the Kindergarten phoneme segmenting measure, the reliability of the growth slope for students in the first quartile was .67, the reliability for students in the second quartile was .66, the reliability for students in the third quartile was .46, and the reliability for students in the fourth quartile was .09. For the Kindergarten LS measure, the reliability of the growth slope for students in the first quartile was .76, the reliability for students in the second quartile was .76, the reliability for students in the third quartile was .68, and the reliability for students in the fourth quartile was .53. For the Kindergarten WRF measure, the reliability of the growth slope for students in the lower 50th percentile was .82, and the reliability for students in the upper 50th percentile was .71.

Grade 1. For the phoneme segmenting measure, the reliability of the growth slope for students in the first quartile was .28. No other reliability estimates were possible for this measure in grade 1. For the grade 1 LS measure, the reliability of the growth slope for students in the first quartile was .57, the reliability for students in the second quartile was .66, the reliability for students in the third quartile was .72, and the reliability for students in the fourth quartile was .42.

Grade 1. For the WRF measure, the reliability of the growth slope for students in the first quartile was .88, the reliability for students in the second quartile was .82, the reliability for students in the third quartile was .76, and the reliability for students in the fourth quartile was .63. For the grade 1 PRF measure, the reliability of the growth slope for students in the first quartile was .93, the reliability for students in the second quartile was.88, the reliability for students in the third quartile was .89, and reliability for students in the 4th quartile was .60.

Reference

Lai, C.-F., Nese, J. F. T., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). *Technical Adequacy of the easyCBM® Primary-Level Reading Measures (Grades K–1), 2009–2010 Version (Technical Report 1003).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1004: Technical Adequacy of the easyCBM® Grade 2 Reading Measures (2009–2010 version) (Jamgochian et al., 2010).

Methods

This report presents reliability and validity evidence for the easyCBM® Grade 2 reading measures from the 2009–2010 school year. Data were collected from 71 schools across three Pacific Northwest districts. All students in attendance during the assessment windows participated. Seasonal samples were large for the easyCBM® measures

(approximately 2,144–2,301 students for multiple-choice reading comprehension; ~2,154–2,207 for word reading fluency; and ~2,205–2,236 for passage reading fluency across fall, winter, and spring). The Grade 2 measures included: (a) Word Reading Fluency (WRF; 60-second timed word list), (b) Passage Reading Fluency (PRF; 60-second oral passage reading with errors marked), and (c) Multiple-Choice Reading Comprehension (MCRC; a computer-delivered narrative passage with 12 multiple-choice items). Alternate-form reliability and content development were addressed in prior technical reports; this report focused on reliability of performance-level scores (evaluated for MCRC using Cronbach's alpha and split-half methods) and reliability of growth (slope) for WRF, PRF, and MCRC using two-level hierarchical linear growth models across fall–winter–spring. Validity evidence included concurrent and predictive relations with SAT-10 (correlations and regression models), construct validity via confirmatory factor analysis (three-factor model: WRF, PRF, MCRC), and predictive validity of growth (correlating growth-model residuals with SAT-10).

Results

Reliability evidence for performance-level scores was reported for MCRC (12 items). Across fall, winter, and spring, Cronbach's alpha ranged from .679 to .748, indicating moderate internal consistency. Split-half analyses (first 6 vs. last 6 items) yielded correlations from .499 to .599, with Spearman–Brown/Guttman coefficients in the mid-.60s to mid-.70s, again suggesting moderate reliability.

Reliability of growth (slope) was generally moderate and varied by initial status (fall quartiles). For WRF, slope reliability declined from .77 (lowest quartile) to .50 (highest quartile). For PRF, slope reliability was .75, .70, and .68 in quartiles 1–3 but dropped to .25 in quartile 4. For MCRC, slope reliability was .48 (quartile 1) and .62 (quartile 2), could not be estimated in quartile 3 due to convergence problems, and was low in quartile 4 (.14). The discussion attributes the weakest slope reliability in the highest quartile to reduced growth/possible ceiling effects.

Note: Technical Report 1005 also reports on reliability for reading measures used in Grades 3-7: internal consistency for MCRC and slope for WRF, PRF, and MCRC. This report is summarized in 2603-VK8R_ValidityReading with corresponding values in TR2603_VK8R_ValidityReadingTRs21-31.xlsx. The reason for this assignment is that the more important values focus on concurrent validity of these measures with the OAKS readding measures.

Reference

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Sáez. L., Anderson, D., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM® grade 2 reading measures (Technical Report No. 1004).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1403: easyCBM® Beginning Reading Measures: Grades K-1 Alternate Form Reliability and Criterion Validity with the SAT-10 (Wray, Lai, et al., 2014).

Technical Report 1403 examined the alternate-form reliability and criterion-related validity of easyCBM® beginning reading measures for kindergarten and grade 1 students. Participants included 222 kindergarten students and 204 first-grade students from six kindergarten and eight first-grade classrooms across three K–5 schools in a semi-rural Pacific Northwest district. All students present during testing participated. Data collection was conducted by 15 trained research assistants who received structured training and periodic refresher sessions to ensure standardized administration and scoring procedures.

Method

The easyCBM® measures examined included Letter Names (LN), Letter Sounds (LS), Phoneme Segmenting (PS), and Word Reading Fluency (WRF). Students were assessed at five time points throughout the school year (November, December, February, April, and June). At each time point students completed two alternate forms of the same measure presented in counterbalanced order to estimate alternate-form reliability. Assessments were individually administered in quiet testing rooms and typically required about 10 minutes for kindergarten and 7 minutes for grade 1 students. Criterion measures were drawn from the Stanford Early School Achievement Test / Stanford

Achievement Test–10 (SESAT/SAT-10). Kindergarten students completed the Sounds and Letters and Word Reading subtests, while grade 1 students completed Word Reading and Word Study Skills subtests in the spring administration.

Data were screened for outliers using boxplots, histograms, scatterplots, and z-score comparisons. Pearson correlations were used to estimate alternate-form reliability when assumptions of normality were met, while Spearman rank correlations were used when distributional assumptions were violated. Criterion validity was evaluated using standard multiple regression models predicting SAT-10 outcomes from the easyCBM® measures.

Results

Moderate to strong alternate-form reliability was found In kindergarten, reliability coefficients ranged from .31 to .94 across measures and time points, with stronger coefficients for Letter Names (.61–.90), Letter Sounds (.54–.92), and Word Reading Fluency (.74–.94). Phoneme Segmenting showed greater variability (.31–.90). In grade 1, alternate-form coefficients ranged from .27 to .96 with generally strong relations for Letter Names, Letter Sounds, and Word Reading Fluency. Regression analyses indicated that the combined easyCBM measures explained 35–58% of the variance in kindergarten SAT-10 outcomes and 14–32% of the variance in grade 1 SAT-10 outcomes. In addition, grade 1 early literacy measures explained 49–56% of the variance in end-of-year easyCBM Word Reading Fluency performance, supporting their usefulness for screening and progress monitoring.

- Kindergarten: LN coefficients ranged from .61–.90
- Kindergarten: LS coefficients ranged from .54–.92
- Kindergarten: PS coefficients ranged from .31–.90
- Kindergarten: WRF coefficients ranged from .74–.94
- Grade 1: These measures had coefficients that ranged from .26 – .96

Reference

Wray, K. A., Lai, C.-F., Sáez, L., Alonzo, J., & Tindal, G. (2013). *easyCBM® beginning reading measures: K-1 alternate form reliability and criterion validity with the SAT-10 (Technical Report No. 1403)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Table 3. Individual Coefficients from Technical Reports

Grade	Measure	Reliability Type	Coefficients	TR#	Page	Median
K	LN	alternate_form	0.87, 0.82, 0.84, 0.76, 0.77, 0.66, 0.68, 0.61, 0.64, 0.84, 0.87, 0.77, 0.77, 0.69, 0.74, 0.62, 0.67, 0.89, 0.83, 0.81, 0.74, 0.74, 0.65, 0.71, 0.83, 0.81, 0.72, 0.7, 0.64, 0.68, 0.88, 0.8, 0.79, 0.73, 0.79, 0.78, 0.74, 0.74, 0.76, 0.9, 0.84, 0.85, 0.83, 0.84, 0.9	1403	17	0.77
K	LS	alternate_form	0.89, 0.81, 0.86, 0.75, 0.74, 0.64, 0.58, 0.6, 0.57, 0.78, 0.83, 0.72, 0.69, 0.61, 0.53, 0.55, 0.54, 0.92, 0.81, 0.82, 0.67, 0.61, 0.63, 0.62, 0.82, 0.82, 0.71, 0.67, 0.69, 0.66, 0.92, 0.79, 0.77, 0.77, 0.75, 0.83, 0.77, 0.79, 0.75, 0.87, 0.85, 0.88, 0.89, 0.87, 0.91	1403	18	0.77
K	PS	alternate_form	0.81, 0.42, 0.51, 0.4, 0.39, 0.42, 0.39, 0.48, 0.44, 0.49, 0.52, 0.37, 0.35, 0.39, 0.39, 0.49, 0.41, 0.9, 0.43, 0.45, 0.45, 0.41, 0.4, 0.31, 0.41, 0.52, 0.53, 0.48, 0.49, 0.43, 0.83, 0.34, 0.42, 0.38, 0.35, 0.46, 0.51, 0.54, 0.48, 0.83, 0.59, 0.51, 0.57, 0.5, 0.86	1403	19	0.45
K	RAN					
K	WRF	alternate_form	0.84, 0.81, 0.83, 0.75, 0.76, 0.77, 0.75, 0.75, 0.74, 0.77, 0.78, 0.7, 0.77, 0.8, 0.78, 0.79, 0.76, 0.91, 0.86, 0.84, 0.85, 0.84, 0.82, 0.81, 0.89, 0.87, 0.88, 0.85, 0.84, 0.82, 0.9, 0.86, 0.84, 0.85, 0.81, 0.86, 0.86, 0.86, 0.84, 0.92, 0.9, 0.89, 0.91, 0.9, 0.94	1403	20	0.84
1	LN	alternate_form	0.88, 0.84, 0.82, 0.75, 0.74, 0.69, 0.68, 0.67, 0.66, 0.85, 0.82, 0.75, 0.75, 0.69, 0.73, 0.7, 0.67, 0.9, 0.78, 0.77, 0.75, 0.78, 0.75, 0.73, 0.77, 0.75, 0.74, 0.76, 0.7, 0.7, 0.85, 0.81, 0.81, 0.81, 0.8, 0.82, 0.81, 0.81, 0.76, 0.89, 0.84, 0.82, 0.86, 0.85, 0.87	1403	21	0.77
1	LN	alternate_form	0.89, 0.82, 0.85	0906	10	0.85
1	LS	alternate_form	0.87, 0.89, 0.83, 0.82, 0.89, 0.85	1216	12	0.86
1	LS	alternate_form	0.76, 0.76, 0.88	0906	10	0.76
1	LS	alternate_form	0.80, 0.64, 0.40, 0.49, 0.71, 0.49, 0.55, 0.71, 0.75, 0.83	1403	22	0.68

1	LS	g_coefficient	0.87, 0.95	1216	12	0.91
1	LS	test_retest	0.77, 0.83, 0.86, 0.87	1216	12	0.85
1	PRF	alternate_form	0.98, 0.98, 0.98, 0.96, 0.95, 0.93	1216	49	0.97
1	PRF	alternate_form	0.95, 0.96, 0.97	0906	11	0.96
1	PRF	alternate_form	0.95, 0.95, 0.95, 0.96, 0.96, 0.96, 0.96, 0.95, 0.95, 0.96, 0.93, 0.93, 0.92, 0.94, 0.94, 0.93, 0.92, 0.91, 0.93, 0.94, 0.91, 0.95, 0.95, 0.95, 0.96, 0.96, 0.93, 0.94, 0.92, 0.92, 0.92, 0.93, 0.93, 0.91, 0.91, 0.93, 0.93, 0.93, 0.93, 0.94, 0.93, 0.91, 0.91, 0.94, 0.92, 0.93, 0.93, 0.93, 0.93, 0.93, 0.92, 0.9, 0.93, 0.91, 0.96, 0.93, 0.94, 0.92, 0.93, 0.92, 0.92, 0.89, 0.93, 0.92, 0.97, 0.96, 0.92, 0.92, 0.92, 0.92, 0.92, 0.91, 0.91, 0.92, 0.91, 0.95, 0.95, 0.96, 0.91, 0.91, 0.91, 0.92, 0.91, 0.91, 0.88, 0.91, 0.9, 0.95, 0.94, 0.96, 0.94, 0.92, 0.93, 0.92, 0.94, 0.93, 0.92, 0.91, 0.93, 0.91, 0.96, 0.95, 0.97, 0.96, 0.96, 0.94, 0.94, 0.93, 0.94, 0.93, 0.92, 0.91, 0.94, 0.92, 0.96, 0.96, 0.96, 0.95, 0.96, 0.97, 0.91, 0.92, 0.93, 0.93, 0.92, 0.89, 0.9, 0.93, 0.91, 0.93, 0.94, 0.95, 0.96, 0.95, 0.95, 0.95, 0.91, 0.9, 0.91, 0.9, 0.91, 0.89, 0.89, 0.91, 0.88, 0.93, 0.93, 0.93, 0.94, 0.93, 0.94, 0.94, 0.92, 0.95, 0.94, 0.94, 0.94, 0.94, 0.92, 0.92, 0.94, 0.92, 0.96, 0.96, 0.96, 0.95, 0.96, 0.97, 0.97, 0.95, 0.94, 0.94, 0.94, 0.94, 0.94, 0.94, 0.93, 0.92, 0.94, 0.92, 0.96, 0.96, 0.96, 0.96, 0.95, 0.96, 0.96, 0.94, 0.94, 0.97	40	39	0.93
1	PRF	g_coefficient	0.99, 0.91	1216	50	0.95
1	PRF	test_retest	0.98, 0.98, 0.83, 0.95	1216	49	0.97
1	PS	alternate_form	0.89, 0.82, 0.78, 0.62, 0.67	1216	24	0.78
1	PS	alternate_form	0.42, 0.41, 0.14. 0.27, 0.38, 0.31, 0.45, 0.33, 0.41, 0.45	1403	22	
1	PS	alternate_form	0.91, 0.86, 0.91	0906	10	0.91
1	PS	g_coefficient	0.83, 0.50, 0.72	1216	24-35	0.72
1	PS	test_retest	0.56, 0.50, 0.81, 0.58, 0.72, 0.32	1216	23	0.57
1	WRF	alternate_form	0.97, 0.91, 0.95, 0.89	1216	38	0.93
1	WRF	alternate_form	0.96, 0.89, 0.79, 0.73, 0.94, 0.85, 0.80, 0.92, 0.87, 0.94	1403	23	0.88
1	WRF	alternate_form	0.96 0.95, 0.95	0906	11	
1	WRF	g_coefficient	0.98, 0.96	1216	38	0.97
1	WRF	test_retest	0.93, 0.95, 0.87, 0.91	1216	38	0.92

2	PRF	alternate_form	0.93, 0.95, 0.92, 0.94, 0.92, 0.91	1217	23	0.925
2	PRF	alternate_form	0.92, 0.93, 0.94, 0.93, 0.94, 0.94, 0.91, 0.93, 0.93, 0.94, 0.92, 0.94, 0.95, 0.95, 0.94, 0.93, 0.92, 0.93, 0.95, 0.92, 0.95, 0.93, 0.93, 0.94, 0.96, 0.93, 0.96, 0.95, 0.92, 0.93, 0.94, 0.95, 0.93, 0.95, 0.94, 0.94, 0.92, 0.92, 0.93, 0.94, 0.92, 0.94, 0.94, 0.94, 0.95, 0.91, 0.91, 0.92, 0.94, 0.92, 0.94, 0.94, 0.94, 0.94, 0.95, 0.92, 0.92, 0.93, 0.95, 0.92, 0.94, 0.94, 0.95, 0.94, 0.96, 0.95, 0.9, 0.9, 0.91, 0.91, 0.9, 0.92, 0.91, 0.92, 0.91, 0.91, 0.91, 0.91, 0.91, 0.93, 0.94, 0.94, 0.91, 0.94, 0.93, 0.94, 0.94, 0.94, 0.93, 0.93, 0.93, 0.93, 0.93, 0.95, 0.95, 0.92, 0.94, 0.94, 0.95, 0.95, 0.94, 0.94, 0.95, 0.93, 0.96, 0.91, 0.92, 0.94, 0.94, 0.92, 0.94, 0.93, 0.94, 0.93, 0.92, 0.93, 0.93, 0.92, 0.95, 0.95, 0.9, 0.91, 0.92, 0.93, 0.91, 0.94, 0.93, 0.95, 0.92, 0.94, 0.94, 0.95, 0.91, 0.93, 0.94, 0.93, 0.91, 0.91, 0.91, 0.93, 0.89, 0.93, 0.92, 0.93, 0.91, 0.92, 0.92, 0.93, 0.89, 0.91, 0.93, 0.91, 0.94, 0.92, 0.93, 0.93, 0.93, 0.92, 0.94, 0.93, 0.94, 0.93, 0.94, 0.93, 0.93, 0.91, 0.94, 0.94, 0.92, 0.94, 0.93, 0.93, 0.92, 0.92, 0.94, 0.92, 0.95, 0.96, 0.95, 0.94, 0.94, 0.95, 0.95, 0.92, 0.93, 0.95, 0.93, 0.95, 0.94, 0.94	40	41	0.93
2	PRF	g_coefficient	0.97, 0.98	1217	24	0.98
2	PRF	test_retest	0.88, 0.96, 0.93, 0.90, 0.94, 0.95	1217	23	0.95
2	WRF	alternate_form	0.95, 0.92, 0.95, 0.94, 0.92, 0.92	1217	12	0.93
2	WRF	g_coefficient	0.99, 0.96	1217	12	0.98
2	WRF	test_retest	0.94, 0.95, 0.93, 0.89, 0.92, 0.87	1217	11	0.93

Table 4. Median Coefficients from Technical Reports

Grade	Measure	Relability_type	Median	Tech Report	Page
K	LN	alternate_form	0.768	1403	17
K	LS	alternate_form	0.746	1403	18
K	PS	alternate_form	0.490	1403	19
K	WRF	alternate_form	0.830	1403	20
1	LN	alternate_form	0.777	1403	21
1	LN	alternate_form	0.853	0906	10
1	LS	alternate_form	0.858	1216	12
1	LS	alternate_form	0.800	0906	10
1	LS	alternate_form	0.637	1403	22
1	LS	g_coefficient		1216	12
1	LS	test_retest	0.833	1216	12
1	PRF	alternate_form	0.963	1216	49
1	PRF	alternate_form	0.960	0906	11
1	PRF	alternate_form	0.933	40	39
1	PRF	g_coefficient		1216	50
1	PRF	test_retest	0.935	1216	49
1	PS	alternate_form	0.756	1216	24
1	PS	alternate_form	0.357	1403	22
1	PS	alternate_form	0.893	0906	10
1	PS	g_coefficient		1216	24-35
1	PS	test_retest	0.582	1216	23
1	WRF	alternate_form	0.930	1216	38
1	WRF	alternate_form	0.869	1403	23
1	WRF	alternate_form	0.953	0906	11
1	WRF	g_coefficient		1216	38
1	WRF	test_retest	0.915	1216	38
2	PRF	alternate_form	0.928	1217	23
2	PRF	alternate_form	0.931	40	
2	PRF	g_coefficient		1217	24
2	PRF	test_retest	0.927	1217	23
2	WRF	alternate_form	0.933	1217	12
2	WRF	g_coefficient		1217	12
2	WRF	test_retest	0.917	1217	11

Table 5. Grade Group Summary of Coefficients from Technical Reports

Grade	Measure	Reliability	Value
K	LN	alternate_form	0.768
K	LS	alternate_form	0.746
K	PS	alternate_form	0.490
K	WRF	alternate_form	0.830
1	LN	alternate_form	0.777
1	LN	alternate_form	0.853
1	LS	alternate_form	0.858
1	LS	alternate_form	0.800
1	LS	alternate_form	0.673
1	LS	test_retest	0.832
1	PRF	alternate_form	0.963
1	PRF	alternate_form	0.960
1	PRF	alternate_form	0.933
1	PRF	test_retest	0.935
1	PS	alternate_form	0.756
1	PS	alternate_form	0.357
1	PS	alternate_form	0.893
1	PS	test_retest	0.582
1	WRF	alternate_form	0.930
1	WRF	alternate_form	0.869
1	WRF	alternate_form	0.953
1	WRF	test_retest	0.915
2	PRF	alternate_form	0.928
2	PRF	alternate_form	0.931
2	PRF	test_retest	0.927
2	WRF	alternate_form	0.933
2	WRF	test_retest	0.917

Summary

Kindergarten has **one** type of reliability above lower bound of median (.70) for Letter Names (LN), Letter Sounds (LS), and Word Reading Fluency (WRF).

Grade 1 has **two** types of reliability above lower bound of median (.70) for Letter Sounds (LS), Passage Reading Fluency (PRF), and Word Reading Fluency (WRF). **One** type of reliability exists above lower bound of median (.70) for Letter Names and Phoneme Segmenting.

Grade 2 has **two** types of reliability above lower bound of median (.70) for both Passage Reading Fluency (PRF) and Word Reading Fluency (WRF).

S**ummary of Technical Report 1201:** Analyzing the Reliability of the easyCBM® Reading Comprehension Measures: Grade 2 (C. F. Lai, P. S. Irvin, et al., 2012).

This report describes the reliability and technical performance of curriculum-based measurement (CBM) assessments used for screening and progress monitoring within a Response to Intervention (RTI) framework. The study emphasizes methodological design, statistical modeling, and empirical findings related to score consistency and measurement precision.

Methods

Participants were drawn from an extant database of students who completed CBM assessments during a single academic year. The sample included students across multiple grade levels and represented a range of instructional needs, including general education and intervention populations. Students were included in analyses only if they had multiple observations on the same measure, ensuring adequate data for reliability estimation. Assessments were administered under standardized conditions using multiple equivalent forms designed to support repeated measurement.

Data preparation involved screening for incomplete records, removing off-grade testing, and addressing anomalous scores. Scores occurring within short time windows were averaged to reduce measurement error, and observations were organized into evenly spaced time segments to support growth-based analyses. Analytical methods combined descriptive statistics with reliability modeling. Internal consistency and stability of scores over time were examined using correlation-based approaches and growth modeling techniques. When longitudinal data were available, individual growth parameters were estimated and compared across parallel forms to evaluate the reliability of observed change. Model estimation relied on maximum likelihood procedures with robust standard errors to ensure accurate parameter estimation.

Results

Results indicated that the CBM measures demonstrated acceptable to strong reliability across grade levels and assessment types. Score distributions showed sufficient variability to differentiate student performance, and reliability coefficients generally met or exceeded recommended thresholds for screening and progress monitoring. Growth estimates were stable, suggesting that observed changes in scores reflected true changes in student performance rather than measurement error. Overall, the findings support the technical adequacy of the measures for use in RTI decision-making and instructional planning.

Table 6. Tests of Within-Subject Effects and Contrasts from Technical Report 1201

Table 2

Tests of Within-Subjects Effects for Grade 2 Measures 8 to 10

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Sphericity Assumed	14.111	2	7.056	4.040	.021
	Greenhouse-Geisser	14.111	1.856	7.602	4.040	.024
	Huynh-Feldt	14.111	1.941	7.272	4.040	.022
	Lower-bound	14.111	1.000	14.111	4.040	.051
Error(forms)	Sphericity Assumed	143.222	82	1.747		
	Greenhouse-Geisser	143.222	76.105	1.882		
	Huynh-Feldt	143.222	79.563	1.800		
	Lower-bound	143.222	41.000	3.493		

Note. Mauchly's Test of Sphericity: The assumption of sphericity was not violated, Mauchly's W was 0.92, χ^2 (2) = 3.23, $p > .05$.

Table 3

Tests of Within-Subjects Contrasts for Grade 2 Measures 8 to 10

Source	forms	Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Level 1 vs. Level 3	14.881	1	14.881	4.916	.032
	Level 2 vs. Level 3	25.929	1	25.929	5.807	.021
Error(forms)	Level 1 vs. Level 3	124.119	41	3.027		
	Level 2 vs. Level 3	183.071	41	4.465		

Table 7. Sample Split-Half Coefficients from Technical Report 1201

Table 3
Grade 2 Split-Half Coefficients for MCRC Form 9 with N = 12 Items

Cronbach's Alpha	Part 1	Value	.667
		N of Items	6[a]
	Part 2	Value	.520
		N of Items	6[b]
	Total N of Items		12
Correlation Between Forms			.775
Spearman-Brown Coefficient	Equal Length		.873
	Unequal Length		.873
Guttman Split-Half Coefficient			.872

Table 8. Distractor Analysis from Technical Report 1201

Table 10
Distractor Analysis, Grade 2, Form 12

Entry #	Data Code	Score Value	Count	%	Average Measure	S.E. Mean
1	A	1	48	94	3.00	0.22
	B	0	3	6	-0.18	0.54
	C	0	0	0	0.00	0.00
	Missing	**				
2	A	0	10	20	0.91	0.47
	B	0	2	4	0.20	0.68
	C	1	39	76	3.43	0.18
	Missing	**				
3	A	1	46	90	3.14	0.20
	B	0	4	8	-0.14	0.65
	C	0	1	2	-0.48	0.00
	Missing	**				
4	A	0	1	2	-1.46	0.00
	B	0	6	12	0.27	0.60
	C	1	44	86	3.26	0.18
	Missing	**				
5	A	1	50	98	2.87	0.23
	B	0	0	0	0	0
	C	0	1	2	-0.04	0
	Missing	**	0	0	0	0
6	A	0	2	4	1.09	2.03
	B	1	47	92	3.02	0.22
	C	0	2	4	-0.26	0.22
	Missing	**				
7	A	1	48	94	3.05	0.20
	B	0	2	4	-0.97	0.49
	C	0	1	2	-0.94	0.00
	Missing	**				
8	A	0	23	45	2.17	0.28
	B	1	27	53	3.48	0.30
	C	0	1	2	-0.48	0.00
	Missing	**				
9	A	0	3	6	0.75	0.42
	B	0	4	8	-0.25	0.60
	C	1	44	86	3.23	0.20
	Missing	**				
10	A	0	1	2	1.42	0.00
	B	1	43	84	3.11	0.23
	C	0	4	8	1.54	0.76
	Missing	**	3	6	0.69	1.09

Table 9. Summary of Key Findings from Technical Report 1201

Category	Summary
Sample	715 students from two Pacific Northwest public school districts participated. Students were primarily from Grade 2, with broader recruitment across Grades 2–8 to support form comparisons. Grade 2 analyses focused on approximately 44–52 students per form.
Assessment Forms	easyCBM® multiple-choice reading comprehension (MCRC) measures, Grade 2. Nine forms were analyzed (Forms 8–16). Each form included 12 items assessing literal and inferential comprehension.
Analysis Method	One-way repeated measures ANOVA (form equivalence), split-half reliability using the Guttman formula, top–bottom reliability analyses (23rd vs. 78th percentile groups), and Rasch modeling (item difficulty, fit statistics, person and item reliability, distractor analysis).
Items Analyzed	12 items per form (11 items for one form due to low variance). Items were designed with a range of difficulties and ordered from easiest to most difficult.
Problematic Items	A small number of items across forms showed misfit (mean square outfit > 1.5 or < 0.5). Some items exhibited unexpected performance in top–bottom analyses, including one item where lower-performing students outperformed higher-performing students. Most issues were isolated and identified via Rasch and distractor analyses.
Item Fit	Overall item fit was acceptable. Most items fell within the recommended Rasch outfit range (0.5–1.5). Distractor analyses generally showed appropriate functioning, with correct answers selected by higher-ability students.
Overall Conclusion	The Grade 2 easyCBM® MCRC measures demonstrated adequate reliability for screening and progress monitoring. While person reliability was sometimes limited due to short test length and ceiling effects, results support the technical adequacy of the measures, with recommendations for continued refinement and future research using larger samples.

Reference

Lai, C.-F., Irvin, P. S., Alonzo, J., Park, B. J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM® reading comprehension measures: Grade 2 (Technical Report 1201)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1202: Analyzing the Reliability of the easyCBM® Reading Comprehension Measures: Grade 3 (C. F. Lai et al., 2012).

Technical Report 1202 presents an evaluation of curriculum-based measurement (CBM) tools designed for use in screening and progress monitoring within a Response to Intervention (RTI) framework. The report focuses on the methodological procedures used to collect longitudinal student performance data and the statistical analyses conducted to determine the reliability and usefulness of the resulting scores.

Methods

Participants were drawn from an extant dataset consisting of students across multiple grade levels who were administered CBM assessments during a single academic year. The sample included both general education students and students receiving supplemental or intensive instructional support. Inclusion criteria required students to have multiple observations on the same measure to allow for reliable estimation of score stability and growth. Assessments were administered under standardized conditions using alternate forms intended to be of

equivalent difficulty. These forms allowed for repeated measurement without substantial practice effects. Data preparation procedures included removal of incomplete records, exclusion of off-grade administrations, and screening for anomalous scores. When multiple scores occurred within a short time window, values were averaged to reduce random measurement error. Observations were organized into consistent time intervals to support longitudinal analyses. Analyses combined descriptive statistics with reliability-focused modeling techniques. Internal consistency and score stability were evaluated using correlation-based approaches. Where longitudinal data permitted, growth parameters were estimated to examine the reliability of change over time. Statistical models were estimated using maximum likelihood methods with robust standard errors to ensure stable and unbiased parameter estimates.

Results

Results indicated that the CBM measures demonstrated acceptable to strong reliability across grades and assessment types. Score distributions showed sufficient variability to distinguish student performance levels, and reliability coefficients generally met recommended thresholds for screening and progress monitoring. Growth estimates were stable across repeated administrations, suggesting that observed score changes primarily reflected true changes in student performance rather than measurement error. Collectively, the findings support the technical adequacy of the measures for instructional decision-making within RTI systems.

Table 10. Example Tests of Within-Subject Effect from Technical Report 1202

Table 2

Tests of Within-Subjects Effects for Grade 3 Measures 8 to 10

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Sphericity Assumed	13.282	2	6.641	1.249	.293
	Greenhouse-Geisser	13.282	1.983	6.698	1.249	.292
	Huynh-Feldt	13.282	2.000	6.641	1.249	.293
	Lower-bound	13.282	1.000	13.282	1.249	.271
Error(forms)	Sphericity Assumed	404.051	76	5.316		
	Greenhouse-Geisser	404.051	75.357	5.362		
	Huynh-Feldt	404.051	76.000	5.316		
	Lower-bound	404.051	38.000	10.633		

Note. Mauchly's Test of Sphericity: The assumption of sphericity was not violated, Mauchly's W was 0.99, χ^2 (2) = 0.32, $p > .05$.

Table 11. Summary of Outcomes from Technical Report 1202

Category	Summary
Sample	Approximately 700 students drawn from two public school districts in the Pacific Northwest. Grade 3 analyses focused on samples of roughly 44–52 students per form, with additional students included for form comparisons.
Assessment Forms	Grade 3 easyCBM® multiple-choice reading comprehension (MCRC) assessments. Nine forms were analyzed, each designed to be parallel in structure and difficulty for repeated administration.
Analysis Method	One-way repeated measures ANOVA to evaluate form equivalence; split-half reliability using the Guttman formula; top–bottom group comparisons (upper and lower performance percentiles); and Rasch modeling to examine item difficulty, fit statistics, person reliability, and distractor functioning.
Items Analyzed	12 multiple-choice reading comprehension items per form, ordered from easiest to most difficult and designed to assess literal and inferential comprehension skills.
Problematic Items	A small number of items showed misfit (outfit mean square values outside the recommended 0.5–1.5 range) or unexpected performance in top–bottom analyses. These items were flagged for potential revision.

Category	Summary
Item Fit	Overall item fit was acceptable. Most items demonstrated appropriate Rasch fit and effective distractor functioning, with higher-performing students consistently selecting correct responses.
Overall Conclusion	Results supported the technical adequacy of the Grade 3 easyCBM® reading comprehension measures for screening and progress monitoring. Although person reliability was somewhat constrained by short test length, the measures demonstrated acceptable reliability and item functioning, with recommendations for continued refinement and larger-sample validation.

Reference

Lai, C.-F., Irvin, P. S., Alonzo, J., Park, B. J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM® reading comprehension measures: Grade 3 (Technical Report 1202)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1203: Analyzing the Reliability of the easyCBM® Reading Comprehension Measures: Grade 4 (Park, Irvin, Alonzo, et al., 2012).

Technical Report 1203 investigates the technical adequacy of curriculum-based measurement (CBM) assessments used for screening and progress monitoring within a Response to Intervention (RTI) framework. The report focuses on the methodological procedures used to collect longitudinal student performance data and the statistical analyses conducted to evaluate score reliability and interpretability.

Methods

Participants were drawn from a large, extant dataset of students who completed CBM assessments during a single academic year. The sample included students across multiple grade levels and represented both general education populations and students receiving supplemental or intensive instructional services. To ensure sufficient data for reliability and growth analyses, students were included only if they had multiple administrations of the same CBM measure.

Assessments were administered under standardized conditions using multiple alternate forms designed to be of comparable difficulty. These alternate forms allowed for repeated measurement while minimizing practice effects. Data preparation procedures included screening for incomplete records, excluding off-grade administrations, and removing anomalous scores. When multiple scores occurred within short time windows, values were averaged to reduce random measurement error. Observations were then organized into consistent time intervals to support longitudinal analyses. Analytical methods included descriptive statistics and reliability-focused modeling approaches. Internal consistency and score stability were evaluated using correlation-based methods. Where longitudinal data were available, growth parameters were estimated to examine the reliability of change over time. Statistical models were estimated using maximum likelihood procedures with robust standard errors to ensure stable parameter estimation.

Results

Results indicated that the CBM measures demonstrated acceptable to strong reliability across grades and assessment types. Score distributions showed adequate variability to differentiate student performance levels, and reliability coefficients generally met or exceeded recommended thresholds for screening and progress monitoring. Growth estimates were stable across repeated administrations, suggesting that observed changes in scores largely reflected true changes in student performance rather than measurement error. Overall, the findings support the use of these CBM measures for instructional decision-making and RTI implementation.

Table 12. Example Tests of Within-Subjects Effects from Technical Report 1203

Table 2

Tests of Within-Subjects Effects for Grade 4 Measures 8 to 10

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Sphericity Assumed	10.364	2	5.182	1.436	.243
	Greenhouse-Geisser	10.364	1.924	5.385	1.436	.244
	Huynh-Feldt	10.364	2.000	5.182	1.436	.243
	Lower-bound	10.364	1.000	10.364	1.436	.237
Error(forms)	Sphericity Assumed	310.303	86	3.608		
	Greenhouse-Geisser	310.303	82.749	3.750		
	Huynh-Feldt	310.303	86.000	3.608		
	Lower-bound	310.303	43.000	7.216		

Note. Mauchly's Test of Sphericity: The assumption of sphericity was not violated, Mauchly's W was 0.96, χ^2 (2) = 1.68, $p > .05$

Table 13. Example Split Half Coefficients from Technical Report 1203

Table 5

Grade 4 Split-Half Coefficients for MCRC Form 10 with N = 20 Items

Cronbach's Alpha	Part 1	Value	.486
		N of Items	10[a]
	Part 2	Value	.423
		N of Items	10[b]
	Total N of Items		20
Correlation Between Forms			.506
Spearman-Brown Coefficient	Equal Length		.672
	Unequal Length		.672
Guttman Split-Half Coefficient			.672

Table 14. Summary of Key Findings from Technical Report 1203

Category	Summary
Sample	Approximately 700 students drawn from two public school districts in the Pacific Northwest. Grade 4 analyses focused on samples of roughly 44–52 students per form, consistent with other early easyCBM® reliability studies.
Assessment Forms	Grade 4 easyCBM® multiple-choice reading comprehension (MCRC) assessments. Nine parallel forms were analyzed, each designed to support repeated administration for screening and progress monitoring.
Analysis Method	One-way repeated measures ANOVA to evaluate form equivalence; split-half reliability using the Guttman formula; top–bottom group comparisons (upper and lower performance percentiles); and Rasch modeling to examine item difficulty, fit statistics, person reliability, and distractor functioning.
Items Analyzed	12 multiple-choice reading comprehension items per form, ordered from easiest to most difficult and designed to assess both literal and inferential comprehension skills.
Problematic Items	A small number of items showed misfit (outfit mean square values outside the recommended 0.5–1.5 range) or unexpected performance in top–bottom analyses. These items were identified as providing less measurement information and flagged for potential revision.
Item Fit	Most items demonstrated acceptable Rasch fit and appropriate distractor functioning. Higher-performing students consistently selected correct responses at higher rates than lower-performing students.

Category	Summary
Overall Conclusion	Findings support the technical adequacy of the Grade 4 easyCBM® reading comprehension measures for screening and progress monitoring. Although person reliability was constrained by short test length, item functioning and form equivalence were generally acceptable, with recommendations for continued refinement and validation using larger samples.

Reference

Park, B. J., Irvin, P. S., Alonzo, J., Lai, C. F., & Tindal, G. (2012). Analyzing the reliability of the easyCBM® reading comprehension measures: Grade 4 (Technical Report # 1203). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1204: Analyzing the Reliability of the easyCBM® Reading Comprehension Measures: Grade 5 (Park, Irvin, Lai, et al., 2012).

Technical Report 1204 evaluates the technical adequacy of curriculum-based measurement (CBM) assessments used for screening and progress monitoring within a Response to Intervention (RTI) framework. The study emphasizes methodological procedures for large-scale data collection and statistical analyses designed to assess score reliability and interpretability over time.

Methods

Participants were drawn from a large extant dataset of students who completed CBM assessments during a single academic year. The sample included students across multiple grade levels and represented both general education populations and students receiving supplemental or intensive instructional services. Students were included in the analyses only if they had multiple administrations of the same CBM measure, ensuring sufficient data for evaluating score stability and growth.

Assessments were administered under standardized conditions using alternate forms designed to be of comparable difficulty. These forms allowed for repeated measurement while minimizing practice effects. Data preparation procedures included screening for incomplete records, excluding off-grade administrations, and removing anomalous or implausible scores. When multiple scores occurred within short time windows, values were averaged to reduce random measurement error. Observations were then organized into consistent time intervals to support longitudinal analyses. Analyses included descriptive statistics and reliability-focused modeling techniques. Internal consistency and score stability were evaluated using correlation-based methods. Where longitudinal data were available, growth parameters were estimated to assess the reliability of change over time. Statistical models were estimated using maximum likelihood procedures with robust standard errors to ensure accurate and stable parameter estimation.

Results

Results indicated that the CBM measures demonstrated acceptable to strong reliability across grades and assessment types. Score distributions showed sufficient variability to differentiate student performance levels, and reliability coefficients generally met or exceeded recommended thresholds for screening and progress monitoring. Growth estimates were stable across repeated administrations, suggesting that observed changes in scores largely reflected true changes in student performance rather than measurement error. Overall, the findings support the use of these CBM measures for instructional decision-making and RTI implementation.

Table 15. Example of Tests within Subjects from Technical Report 1204

Table 2
Tests of Within-Subjects Effects for Grade 5 Measures 8 to 10

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Sphericity Assumed	16.056	2	8.028	1.932	.151
	Greenhouse-Geisser	16.056	1.964	8.174	1.932	.151
	Huynh-Feldt	16.056	2.000	8.028	1.932	.151
	Lower-bound	16.056	1.000	16.056	1.932	.171
Error(forms)	Sphericity Assumed	390.611	94	4.155		
	Greenhouse-Geisser	390.611	92.313	4.231		
	Huynh-Feldt	390.611	94.000	4.155		
	Lower-bound	390.611	47.000	8.311		

Note. Mauchly's Test of Sphericity: The assumption of sphericity was not violated, Mauchly's W was 0.98, χ^2 (2) = 0.85, $p > .05$.

Table 16. Summary of Key Findings from Technical Report 1204

Category	Summary
Sample	Approximately 700 students drawn from two public school districts in the Pacific Northwest. Grade 5 analyses focused on samples of roughly 44–52 students per form, consistent with earlier grade-level studies in the series.
Assessment Forms	Grade 5 easyCBM® multiple-choice reading comprehension (MCRC) assessments. Nine parallel forms were examined, each designed to support repeated administration for screening and progress monitoring.
Analysis Method	One-way repeated measures ANOVA to assess form equivalence; split-half reliability using the Guttman formula; top–bottom group comparisons (upper and lower performance percentiles); and Rasch modeling to evaluate item difficulty, fit statistics, person reliability, and distractor functioning.
Items Analyzed	20 multiple-choice reading comprehension items per form, ordered from easiest to most difficult and designed to assess literal, inferential, and applied comprehension skills.
Problematic Items	A small number of items demonstrated misfit (outfit mean square values outside the recommended 0.5–1.5 range) or weaker discrimination in top–bottom analyses. These items were identified as providing reduced measurement information and were flagged for potential revision.
Item Fit	Most items demonstrated acceptable Rasch fit and appropriate distractor functioning. Higher-performing students selected correct responses at substantially higher rates than lower-performing students, supporting item validity.
Overall Conclusion	Results supported the technical adequacy of the Grade 5 easyCBM® reading comprehension measures for screening and progress monitoring. Although person reliability was constrained by test length, overall item functioning and form equivalence were acceptable, with recommendations for continued refinement and validation using larger samples.

Reference

Park, B. J., Irvin, P. S., Lai, C. F., Alonzo, J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 5 (Technical Report # 1204)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1205: Analyzing the Reliability of the easyCBM® Reading Comprehension measures: Grade 6 (Irvin, Alonzo, Park, et al., 2012).

This study documents technical adequacy of curriculum-based measurement (CBM) used to screen and progress monitor, emphasizing methodological rigor in evaluating reliability, score stability, and interpretability over time.

Methods

Participants were drawn from a large extant dataset of students who completed CBM assessments during a single academic year. The sample included students across multiple grade levels and represented general education populations as well as students receiving supplemental or intensive instructional support. Inclusion criteria required students to have multiple administrations of the same CBM measure, ensuring sufficient data to support reliability and growth analyses. Assessments were administered under standardized conditions using alternate forms designed to be of comparable difficulty. The use of alternate forms allowed for frequent measurement while minimizing practice effects. Data preparation procedures included screening for incomplete records, excluding off-grade administrations, and removing anomalous or implausible scores. When multiple scores occurred within short time windows, values were averaged to reduce random measurement error. Observations were organized into consistent time intervals to support longitudinal analyses.

Analytical methods combined descriptive statistics with reliability-focused modeling approaches. Internal consistency and score stability were evaluated using correlation-based methods. Where longitudinal data were available, growth parameters were estimated to assess the reliability of change over time. Statistical models were estimated using maximum likelihood procedures with robust standard errors to ensure accurate parameter estimation.

Results

Results indicated that the CBM measures demonstrated acceptable to strong reliability across grade levels and assessment types. Score distributions showed sufficient variability to differentiate student performance, and reliability coefficients generally met or exceeded recommended thresholds for screening and progress monitoring. Growth estimates were stable across repeated administrations, suggesting that observed changes in scores primarily reflected true changes in student performance rather than measurement error. Overall, the findings support the use of these CBM measures for instructional decision-making within RTI systems.

Table 17. Example of Within-Subject Effects from Technical Report 1205

Table 2

Tests of Within-Subjects Effects for Grade 6 Measures 1 to 3

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Sphericity Assumed	312.973	2	156.487	28.916	.000
	Greenhouse-Geisser	312.973	1.931	162.086	28.916	.000
	Huynh-Feldt	312.973	2.000	156.487	28.916	.000
	Lower-bound	312.973	1.000	312.973	28.916	.000
Error(forms)	Sphericity Assumed	530.360	98	5.412		
	Greenhouse-Geisser	530.360	94.615	5.605		
	Huynh-Feldt	530.360	98.000	5.412		
	Lower-bound	530.360	49.000	10.824		

Note. Mauchly's Test of Sphericity: The assumption of sphericity was not violated, Mauchly's W was 0.96, χ^2 (2) = 1.75, $p > .05$.

Table 18. Example of Within-Subject Contrasts from Technical Report 1205

Table 3
Tests of Within-Subjects Contrasts for Grade 6 Measures 1 to 3

Source	forms	Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Level 1 vs. Level 3	5.780	1	5.780	.496	.485
	Level 2 vs. Level 3	414.720	1	414.720	34.485	.000
Error(forms)	Level 1 vs. Level 3	571.220	49	11.658		
	Level 2 vs. Level 3	589.280	49	12.026		

Table 19. Summary of Key Findings from Technical Report 1205

Category	Summary
Sample	Approximately 700 students drawn from two public school districts in the Pacific Northwest. Grade 6 analyses focused on samples of roughly 44–52 students per form, consistent with earlier reports in the series.
Assessment Forms	Grade 6 easyCBM® multiple-choice reading comprehension (MCRC) assessments. Nine parallel forms were examined, each designed to be equivalent in structure and difficulty for repeated administration.
Analysis Method	One-way repeated measures ANOVA to evaluate form equivalence; split-half reliability using the Guttman formula; top–bottom group comparisons (upper and lower performance percentiles); and Rasch modeling to examine item difficulty, fit statistics, person reliability, and distractor functioning.
Items Analyzed	20 multiple-choice reading comprehension items per form, ordered from easiest to most difficult and targeting literal, inferential, and applied comprehension skills appropriate for Grade 6.
Problematic Items	A small number of items demonstrated misfit (outfit mean square values outside the recommended 0.5–1.5 range) or weaker discrimination in top–bottom analyses. These items were flagged for potential revision but did not substantially affect overall form performance.
Item Fit	Most items demonstrated acceptable Rasch fit and appropriate distractor functioning. Higher-performing students selected correct responses at higher rates than lower-performing students, indicating sound item functioning.
Overall Conclusion	Results supported the technical adequacy of the Grade 6 easyCBM® reading comprehension measures for screening and progress monitoring. While person reliability was constrained by test length, item functioning and form equivalence were generally strong, with recommendations for continued refinement and validation using larger samples.

Reference

Irvin, P. S., Alonzo, J., Park, B. J., Lai, C. F., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 6 (Technical Report # 1205)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1206: Reliability of the easyCBM® Reading Comprehension: Grade 7 (Irvin, Alonzo, Lai, et al., 2012).

This report describes the technical adequacy of curriculum-based measurement (CBM) assessments used for screening and progress monitoring within a Response to Intervention (RTI) framework. The report emphasizes methodological procedures for large-scale data collection and statistical analyses designed to evaluate score reliability, stability, and interpretability over time.

Methods

Participants were drawn from a large extant dataset of students who completed CBM assessments during a single academic year. The sample included students across multiple grade levels and represented both general education populations and students receiving supplemental or intensive instructional services. Inclusion criteria required students to have multiple administrations of the same CBM measure to support reliable estimation of score stability and growth. Assessments were administered under standardized conditions using alternate forms designed to be of comparable difficulty. This design allowed for repeated measurement while minimizing practice effects. Data preparation procedures included screening for incomplete records, excluding off-grade administrations, and removing anomalous or implausible scores. When multiple scores occurred within short time windows, values were averaged to reduce random measurement error. Observations were then organized into consistent time intervals to support longitudinal analyses.
Analytical methods combined descriptive statistics with reliability-focused modeling approaches. Internal consistency and score stability were evaluated using correlation-based methods. Where longitudinal data were available, growth parameters were estimated to assess the reliability of change over time. Statistical models were estimated using maximum likelihood procedures with robust standard errors to ensure accurate and stable parameter estimation.

Results

Results indicated that the CBM measures demonstrated acceptable to strong reliability across grade levels and assessment types. Score distributions showed sufficient variability to differentiate student performance levels, and reliability coefficients generally met or exceeded recommended thresholds for screening and progress monitoring. Growth estimates were stable across repeated administrations, suggesting that observed changes in scores largely reflected true changes in student performance rather than measurement error. Overall, the findings support the use of these CBM measures for instructional decision-making within RTI systems.

Table 20. Example Split-Half Coefficients from Report 1206

Table 5

Grade 7 Split-Half Coefficients for MCRC Form 11 with N = 20 Items

Cronbach's Alpha	Part 1	Value	.215
		N of Items	10[a]
	Part 2	Value	.271
		N of Items	10[b]
	Total N of Items		20
Correlation Between Forms			.176
Spearman-Brown Coefficient	Equal Length		.299
	Unequal Length		.299
Guttman Split-Half Coefficient			.296

Table 21. Example Within-Subject Effects from Report 1206

Table 2
Tests of Within-Subjects Effects for Grade 7 Measures 6, 10, and 17

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
forms	Sphericity Assumed	20.222	2	10.111	2.743	.067
	Greenhouse-Geisser	20.222	1.965	10.290	2.743	.068
	Huynh-Feldt	20.222	2.000	10.111	2.743	.067
	Lower-bound	20.222	1.000	20.222	2.743	.102
Error(forms)	Sphericity Assumed	589.778	160	3.686		
	Greenhouse-Geisser	589.778	157.225	3.751		
	Huynh-Feldt	589.778	160.000	3.686		
	Lower-bound	589.778	80.000	7.372		

Note. Mauchly's Test of Sphericity: The assumption of sphericity was not violated, Mauchly's W was 0.98, χ^2 (2) = 1.41, $p > .05$.

Table 22. Summary of Key Findings from Technical Report 1206

Category	Summary
Sample	Approximately 700 students drawn from two public school districts in the Pacific Northwest. Grade 7 analyses focused on samples of roughly 44–52 students per form, consistent with the earlier grade-level studies in the series.
Assessment Forms	Grade 7 easyCBM® multiple-choice reading comprehension (MCRC) assessments. Nine parallel forms were examined, each designed to be equivalent in structure and difficulty for repeated administration.
Analysis Method	One-way repeated measures ANOVA to evaluate form equivalence; split-half reliability using the Guttman formula; top–bottom group comparisons (upper and lower performance percentiles); and Rasch modeling to examine item difficulty, fit statistics, person reliability, and distractor functioning.
Items Analyzed	20 multiple-choice reading comprehension items per form, ordered from easiest to most difficult and targeting literal, inferential, and applied comprehension skills appropriate for Grade 7.
Problematic Items	A small number of items demonstrated misfit (outfit mean square values outside the recommended 0.5–1.5 range) or weaker discrimination in top–bottom analyses. These items were identified as contributing less measurement information and were flagged for potential revision.
Item Fit	Most items demonstrated acceptable Rasch fit and appropriate distractor functioning. Higher-performing students consistently selected correct responses at higher rates than lower-performing students.
Overall Conclusion	Findings supported the technical adequacy of the Grade 7 easyCBM® reading comprehension measures for screening and progress monitoring. Although person reliability was constrained by test length, item functioning and form equivalence were generally strong, with recommendations for continued refinement and validation using larger samples.

Reference

Irvin, P. S., Alonzo, J., Lai, C. F., Park, B. J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 7 (Technical Report # 1206)*. Eugegne, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1216: An Examination of Test-Retest, Alternate Form Reliability, and Generalizability Theory Study of the easyCBM® Reading Assessments: Grade 1 (Anderson, Park, et al., 2012).

This study examined the reliability and measurement dependability of Grade 1 easyCBM® reading progress-monitoring assessments. The analyses focused on four measures of early literacy: Letter Sounds (LS), Phoneme Segmenting (PS), Word Reading Fluency (WRF), and Passage Reading Fluency (PRF). Data were collected in spring 2011 from Grade 1 students in a mid-sized Pacific Northwest school district. Students completed individually administered easyCBM® assessments on two occasions separated by one week. Alternate test forms were used and counterbalanced across administrations to reduce order effects.

Methods

Test-retest reliability was estimated by correlating scores from the same forms administered across the two sessions. Alternate-form reliability was evaluated using correlations among different test forms. After establishing form comparability, generalizability theory (G-theory) analyses were conducted using a fully crossed design that included persons, forms, and occasions as sources of variance. G-studies estimated variance components associated with each facet and their interactions. Decision studies (D-studies) then evaluated how reliability changed under different testing conditions, such as increasing the number of forms or testing occasions. These analyses produced estimates of error variance, standard errors of measurement, and dependability coefficients for each measure.

Results

Results indicated generally strong reliability for the Grade 1 easyCBM® reading measures. Letter Sounds showed moderately strong test-retest reliability, with correlations ranging from .77 to .87, and alternate-form correlations from .82 to .89. Phoneme Segmenting showed more variability, with test-retest correlations ranging from .50 to .81 and alternate-form correlations from .62 to .89.

The strongest reliability estimates were found for the fluency measures. Word Reading Fluency showed test-retest reliability between .87 and .95 and alternate-form correlations between .89 and .97. Passage Reading Fluency also demonstrated strong reliability, with test-retest correlations from .83 to .98 and alternate-form correlations from .93 to .98. Generalizability analyses showed that most variance in the fluency measures was attributable to differences between students, indicating strong measurement precision. Word Reading Fluency showed particularly strong results, with approximately 85–94% of variance attributed to persons. Passage Reading Fluency showed similarly strong results in one analysis. In contrast, Phoneme Segmenting produced weaker results, with larger person-by-occasion interactions suggesting greater sensitivity to differences in administration conditions.

Decision-study analyses indicated that reliability generally increased when additional testing occasions or forms were used, although many fluency measures demonstrated adequate reliability with a single form administered on one occasion. Overall, the findings provide evidence supporting the reliability and technical adequacy of the Grade 1 easyCBM reading measures for progress monitoring.

Figure 1. Example G-Coefficients from Technical Report 1216

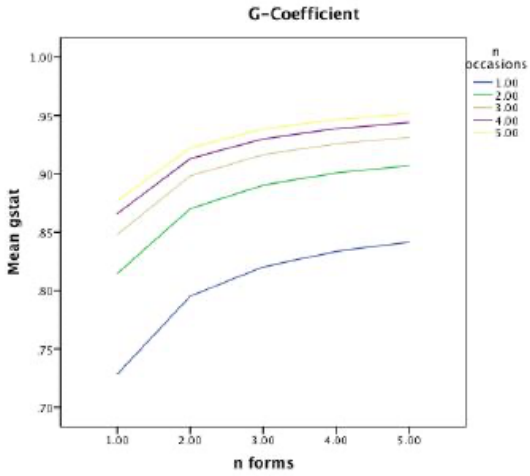


Table 23. Example Test-Retest Reliability from Technical Report 1216

Table 3
Test-retest Reliability of Grade 1 Letter Sound Measures

Test Form	LS1.11.2	LS1.13.2	LS1.14.2	LS1.16.2
LS1.11.1	0.83			
LS1.13.1		0.86		
LS1.14.1			0.87	
LS1.16.1				0.77

Table 24. Example D-Study G and ϕ Coefficients from Technical Report 1216

Grade 1 LS: Forms 11 & 13
D-Study G Coefficients, Ep^2

n forms	*n* occasions				
	1	2	3	4	5
1	0.728	0.815	0.848	0.866	0.877
2	0.795	**0.870**	0.898	0.913	0.922
3	0.820	0.890	0.916	0.93	0.938
4	0.834	0.901	0.926	0.939	0.947
5	0.842	0.907	0.931	0.944	0.952

Grade 1 LS: Forms 11 & 13
D-Study Phi Coefficients, Φ

n forms	*n* occasions				
	1	2	3	4	5
1	0.603	0.730	0.785	0.815	0.835
2	0.652	**0.777**	0.830	0.859	0.877
3	0.670	0.794	0.846	0.874	0.893
4	0.680	0.803	0.854	0.882	0.900
5	0.685	0.808	0.859	0.887	0.905

Reference

Anderson, D., Park, B. J., Lai, C.-F., Alonzo, J., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM® reading assessments: Grade 1 (Technical Report No. 1216)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1217: An Examination of Test–retest, Alternate Form Reliability, and Generalizability Theory Study of the easyCBM® Reading Assessments: Grade 2 (Anderson, Lai, et al., 2012).

Technical Report 1217 presents an examination of test–retest reliability, alternate-form reliability, and generalizability theory analyses for the easyCBM® reading assessments administered at Grade 2. The study forms part of a larger series evaluating the technical adequacy of easyCBM® reading measures for progress monitoring within Response to Intervention (RTI) frameworks.

Methods

Data were collected in spring 2011 from a convenience sample of Grade 2 students nested within classrooms in a mid-sized school district in the Pacific Northwest. Teams of trained University of Oregon research assistants administered the assessments under standardized conditions. Students were assessed during two testing sessions spaced one week apart. Each student completed multiple alternate forms of grade-appropriate Word Reading Fluency (WRF) and Passage Reading Fluency (PRF) measures in one-on-one settings. Administration order was counterbalanced across teachers and sessions to control for order effects while allowing evaluation of both test–retest and alternate-form reliability.

Test–retest and alternate-form reliability were examined using bivariate correlations among forms administered across sessions. Following confirmation of acceptable form equivalence, generalizability theory (G-theory) analyses were conducted using a fully crossed, two-facet design with persons, forms, and occasions as sources of variance. G-studies partitioned variance components attributable to students, forms, occasions, and their interactions. Decision studies (D-studies) were then used to estimate absolute and relative error variances, standard errors of measurement, and dependability coefficients (G-coefficients and Phi coefficients) under varying numbers of forms and testing occasions.

Results

Results indicated strong reliability across both Grade 2 reading measures. Test–retest reliability for Word Reading Fluency ranged from .87 to .95, while Passage Reading Fluency ranged from .88 to .96. Alternate-form reliability was similarly high, with correlations typically between .91 and .95. G-theory analyses showed that the majority of variance was attributable to differences among students, with minimal variance due to forms or occasions. G-coefficients and Phi coefficients generally exceeded .95 for both WRF and PRF measures, indicating highly dependable scores. D-study results demonstrated that even a single form administered on a single occasion yielded reliability exceeding commonly accepted thresholds, with additional forms or occasions further reducing error. Overall, the findings support the technical adequacy of Grade 2 easyCBM® reading measures for progress monitoring and instructional decision-making.

Figure 2. Example G-Coefficients from Technical Report 1217

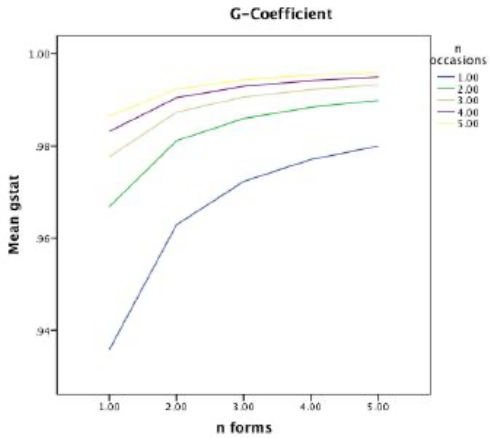


Table 25. Example Test-Retest Reliability Word Reading Fluency Measures Technical Report 1217

Table 3
Test-retest Reliability of Grade 2 Word Reading Fluency Measures

Test Form	WRF2.11.2	WRF2.12.2	WRF2.13.2	WRF2.14.2	WRF2.15.2	WRF2.16.2
WRF2.11.1	0.94					
WRF2.12.1		0.95				
WRF2.13.1			0.93			
WRF2.14.1				0.89		
WRF2.15.1					0.92	
WRF2.16.1						0.87

Table 26. Summary of Key Findings from Technical Report 1217

Category	Summary
Sample	Convenience sample of Grade 2 students drawn from a mid-sized school district in the Pacific Northwest. Students were nested within classrooms and teachers; analytic samples ranged from approximately 20–34 students per G-theory analysis, depending on form set and teacher grouping.
Assessment Forms	easyCBM® Grade 2 reading progress monitoring measures, including **Word Reading Fluency (WRF)** and **Passage Reading Fluency (PRF)**. Two sets of three parallel forms were analyzed (Forms 11–13 and Forms 14–16), administered across two testing occasions.
Analysis Method	Classical test theory analyses (bivariate correlations) for test–retest and alternate-form reliability, followed by **generalizability theory** using a fully crossed two-facet design (persons × forms × occasions). G-studies estimated variance components; D-studies estimated absolute and relative error, standard errors of measurement, and G- and Phi-coefficients under varying numbers of forms and occasions.
Items Analyzed	Fluency scores rather than discrete items. WRF scores reflected correct words read; PRF scores reflected passage reading rate. Multiple forms per measure were treated as random samples from a universe of possible forms.
Problematic Items	No specific forms or items were identified as problematic. Form equivalence was supported, with negligible variance attributable to forms or occasions in G-theory analyses.
Item Fit	Both WRF and PRF measures demonstrated strong test–retest reliability (WRF ≈ .87–.95; PRF ≈ .88–.96) and strong alternate-form correlations (generally ≈ .91–.95). G-theory results showed the majority of variance attributable to persons (≈ 87%–92%), with low error variance and high dependability coefficients.
Overall Conclusion	Results provide strong evidence for the reliability and dependability of Grade 2 easyCBM® reading fluency measures. Test–retest, alternate-form, and G-theory analyses support the use of WRF and PRF for progress monitoring, with reliable measurement achievable even with a single form administered on a single occasion.

Reference

Anderson, D., Lai, C.-F., Park, B. J., Alonzo, J., & Tindal, G. (2012). *An examination of test–retest, alternate form reliability, and generalizability theory study of the easyCBM® reading assessments: Grade 2 (Technical Report 1217)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1218: An Examination of Test–retest, Alternate Form Reliability, and Generalizability Theory Study of the easyCBM® Word and Passage Reading Fluency Assessments: Grade 3 (Park, Anderson, et al., 2012).

Technical Report 1218 examines the reliability and technical adequacy of the easyCBM® Word Reading Fluency (WRF) and Passage Reading Fluency (PRF) assessments for Grade 3 students. The study is part of a broader series evaluating easyCBM® reading measures using traditional reliability methods and generalizability theory to determine the dependability of scores for progress monitoring within Response to Intervention (RTI) frameworks.

Methods

Data were collected in the spring of 2011 from a convenience sample of Grade 3 students nested within classrooms in a mid-sized school district in the Pacific Northwest. Teams of trained University of Oregon research assistants administered the assessments under standardized conditions. Students participated in two assessment sessions spaced one week apart. During each session, students completed multiple alternate forms of grade-appropriate WRF and PRF assessments in one-on-one testing environments. Administration order was counterbalanced across sessions and teachers to control for order effects and support test–retest and alternate-form reliability analyses.

Initial analyses examined descriptive statistics for each form and session. Test–retest reliability was evaluated using bivariate correlations between identical forms administered across sessions, while alternate-form reliability was assessed through correlations among different forms administered within and across sessions. Following confirmation of acceptable form equivalence, generalizability theory (G-theory) analyses were conducted using a fully crossed, two-facet design with persons, forms, and occasions as sources of variance. Variance components were estimated through G-studies, and decision studies (D-studies) were used to estimate absolute and relative error variances, standard errors of measurement, and dependability coefficients (G-coefficients and Phi coefficients) under varying numbers of forms and testing occasions.

Results

Results for Word Reading Fluency indicated moderately strong test–retest reliability, with coefficients ranging from approximately .67 to .92, and moderate to high alternate-form correlations ranging from about .72 to .92. Passage Reading Fluency demonstrated stronger reliability overall, with test–retest correlations between .84 and .94 and alternate-form correlations consistently above .90. G-theory analyses showed that a substantial proportion of score variance was attributable to differences among students, although results varied by form set and teacher grouping. Dependability coefficients for some WRF analyses fell below commonly accepted thresholds when only one form and one occasion were used, indicating the need for multiple forms and/or occasions to achieve adequate reliability. In contrast, PRF measures generally achieved high G- and Phi-coefficients, often exceeding .90 with limited testing occasions. Overall, the findings support the use of Grade 3 easyCBM® PRF measures for progress monitoring, while suggesting more cautious and structured administration for WRF measures.

Figure 3. Example G-Coefficients from Technical Report 1218

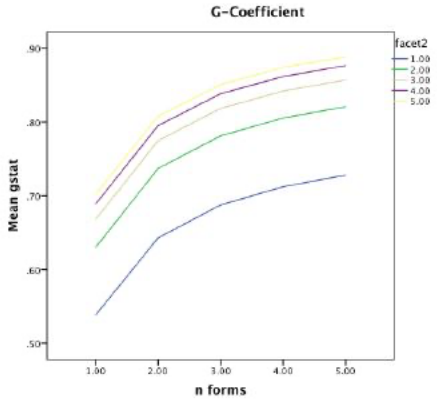


Table 27. Example Test-Retest Reliability Word Reading Fluency Measures Technical Report 1218

Table 3
Test-retest Reliability Results: Word Reading Fluency

Test Form	WRF3.11.2	WRF3.12.2	WRF3.13.2	WRF3.14.2	WRF3.15.2	WRF3.16.2
WRF3.11.1	0.92					
WRF3.12.1		0.90				
WRF3.13.1			0.83			
WRF3.14.1				0.74		
WRF3.15.1					0.67	
WRF3.16.1						0.69

Table 28. Summary of Key Findings from Technical Report 1218

Category	Summary
Sample	Convenience sample of Grade 3 students from a mid-sized school district in the Pacific Northwest. Students were nested within classrooms; analytic samples for G-theory ranged from approximately 9 to 31 students depending on teacher and form set.
Assessment Forms	easyCBM® Grade 3 **Word Reading Fluency (WRF)** and **Passage Reading Fluency (PRF)** measures. Multiple parallel forms were administered across two occasions, including Forms 11–13 and Forms 14–16, with counterbalanced administration.
Analysis Method	Classical test theory analyses (bivariate correlations) for test–retest and alternate-form reliability, followed by **generalizability theory** using a fully crossed persons × forms × occasions design. G-studies estimated variance components; D-studies estimated error variances, standard errors of measurement, and G- and Phi-coefficients under varying numbers of forms and occasions.
Items Analyzed	Fluency scores rather than discrete items. WRF scores reflected correct words read; PRF scores reflected passage reading rate. Forms were treated as random samples from a universe of possible forms.
Problematic Items	No individual items were identified as problematic. Only Form 14 appeared in analyses with weaker variance structures, suggesting potential form-level instability warranting further investigation.
Item Fit	Test–retest reliability was moderate to strong (WRF ≈ .67–.92; PRF ≈ .84–.94). Alternate-form reliability was strong (WRF ≈ .72–.92; PRF ≈ .92–.96). G-theory analyses showed 48%–81% of variance attributable to persons, with minimal variance due to forms or occasions.
Overall Conclusion	Results provide moderate to strong evidence for the reliability of Grade 3 easyCBM® WRF and PRF measures. While some G-theory analyses indicated that a single form on a single occasion may be insufficient for dependable relative decisions, administering multiple forms across occasions substantially improved reliability. Overall findings support use for progress monitoring, with guidance on optimal administration conditions.

Reference

Park, B. J., Anderson, D., Alonzo, J., Lai, C.-F., & Tindal, G. (2012). *An examination of test–retest, alternate form reliability, and generalizability theory study of the easyCBM® word and passage reading fluency assessments: Grade 3 (Technical Report 1218).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1219: An Examination of Test–retest, Alternate Form Reliability, and Generalizability Theory Study of the easyCBM® Word and Passage Reading Fluency Assessments: Grade 4 (Alonzo et al., 2012).

Methods

Data were collected in spring 2011 from a convenience sample of Grade 4 students nested within classrooms in a medium-sized Pacific Northwest school district. University of Oregon research assistants individually administered nine alternate forms of the easyCBM® Passage Reading Fluency (PRF) measure to students in one-on-one sessions across two testing occasions held one week apart. Administration followed standardized protocols, and forms were counterbalanced across sessions to control for order effects. Select forms were repeated across sessions to permit test-retest analyses. Sample sizes varied by form, ranging from 20 to 74 students per session.
Bivariate correlations were used to calculate test-retest and alternate form reliability. Generalizability (G) and Decision (D) studies employed a two-facet fully crossed design (persons x forms x occasions), analyzing variance components attributable to persons, forms, occasions, and their interactions.

Results

Test-retest reliability was examined by correlating scores from the same PRF form administered one week apart. Results indicated moderately strong reliability across all forms, with correlations ranging from .86 to .96. The lowest coefficient was observed for Form 16 (r = .86), while Forms 9, 12, and others yielded coefficients at or near .96. These findings indicate that the Grade 4 PRF measures produce consistent scores over time, supporting their use for tracking student progress across successive testing occasions.

Alternate form reliability was assessed through bivariate correlations between different PRF forms administered within the same testing session. Correlations ranged from .83 to .98, reflecting moderate to strong relationships across forms. The highest coefficients (e.g., .97-.98) were observed among forms such as PRF4.9.2 and its alternates, while slightly lower correlations (e.g., .70-.74) appeared in select form pairings. Overall, the pattern of results supports adequate form equivalence, meaning that scores from different forms can be meaningfully compared when monitoring student progress over time.

Two G-study analyses were conducted: one examining Forms 11, 12, and 13 (n = 39 students) and one examining Forms 14 and 16 (n = 48 students). In both analyses, the large majority of score variance was attributable to individual differences among persons: 88% for the first analysis and 80% for the second. Variance associated with forms was 0% in both analyses, confirming strong form equivalence. Variance attributable to occasions was small but non-trivial (approximately 4-5%), suggesting modest differences in performance across the two testing sessions.
G-coefficients for relative decisions (appropriate for norm-referenced interpretations such as benchmarking) were .98 and .94 for the two analyses, respectively. Phi coefficients for absolute decisions (appropriate for criterion-referenced or progress-monitoring decisions) were .96 and .91. These values indicate a high degree of measurement dependability under the conditions of the study.

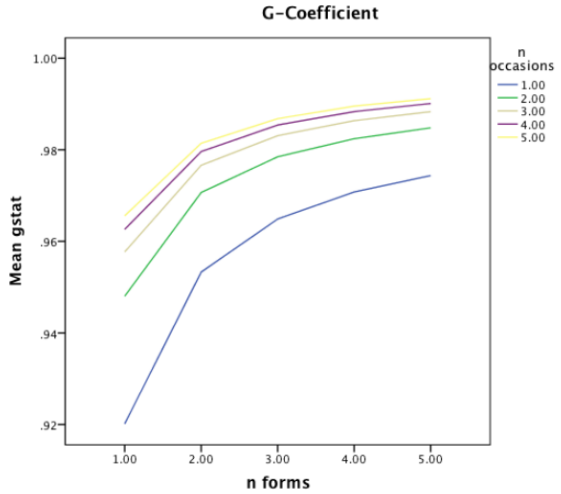
D-study prophecy analyses examined how varying the number of forms and testing occasions would affect reliability. Results consistently showed that administering a single form on a single occasion yielded phi coefficients above .80, the conventional threshold for acceptable dependability. Increasing the number of forms or occasions produced incremental gains in reliability, but the improvements were modest relative to the baseline single-form, single-occasion design. This finding suggests that the easyCBM® Grade 4 PRF measures can support dependable measurement decisions even in typical classroom contexts where administration of multiple forms per occasion may be impractical. Absolute standard errors under the observed design conditions were 8.004 (Forms 11-13) and 9.917 (Forms 14 & 16), reflecting relatively low measurement error given that mean scores ranged from approximately 120 to 169 words correct per minute.

Table 29. Example Correlations for Passage Reading Fluency Measures in Technical Report 1219

Table 4
Correlation between Alternate Forms of Grade 4 Passage Reading Fluency Measure

Test Form	PRF4.11.2	PRF4.12.2	PRF4.13.2	PRF4.14.2	PRF4.15.2	PRF4.16.2
PRF4.8.2	0.95	0.97	0.93			
PRF4.9.2	0.98	0.97	0.97			
PRF4.10.2				0.95	0.92	0.94
PRF4.11.2		0.96	0.94	0.95		0.74
PRF4.12.2			0.95			
PRF4.13.2				0.92		0.70
PRF4.14.2					0.89	0.83
PRF4.15.2						0.88

Figure 4. Example G-Coefficients from Technical Report 1219



Alonzo, J., Lai, C.-F., Anderson, D., Park, B. J., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM® passage reading fluency assessments: Grade 4* (Technical Report No. 1219). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1220: Test–retest, Alternate Form Reliability: Grade 5 (C. F. Lai, B. J. Park, et al., 2012).

Technical Report 1220 examines the test–retest reliability, alternate-form reliability, and generalizability theory (G-theory) results for the easyCBM® Passage Reading Fluency (PRF) assessments administered at Grade 5. This report is part of a larger series evaluating the technical adequacy of easyCBM® reading measures for use in progress monitoring within Response to Intervention (RTI) frameworks.

Methods

Data were collected in the spring of 2011 from a convenience sample of Grade 5 students nested within classrooms in a mid-sized school district in the Pacific Northwest. Teams of trained University of Oregon research assistants administered the assessments under standardized conditions. Students participated in two one-on-one testing sessions spaced one week apart. Across sessions, students completed multiple alternate forms of grade-appropriate PRF assessments. Administration order was counterbalanced across teachers and sessions to control for order effects while supporting both test–retest and alternate-form reliability analyses.

Test–retest reliability was evaluated using bivariate correlations between identical PRF forms administered across sessions. Alternate-form reliability was examined through correlations among different PRF forms administered within and across sessions. Following confirmation of form equivalence, generalizability theory analyses were conducted using a fully crossed two-facet design with persons, forms, and occasions as sources of variance. G-studies partitioned variance attributable to students, forms, occasions, and their interactions. Decision studies (D-studies) were then used to estimate absolute and relative error variances, standard errors of measurement, and dependability coefficients (G-coefficients and Phi coefficients) under varying numbers of forms and testing occasions.

Results

Results indicated moderate to strong test–retest reliability for most Grade 5 PRF forms, with correlations generally ranging from .88 to .94, although one form demonstrated lower stability. Alternate-form correlations were typically strong, ranging from approximately .85 to .98, indicating high comparability across forms. G-theory analyses showed that the majority of score variance (approximately 57%–89%) was attributable to differences among students, with minimal variance due to forms or occasions. Dependability coefficients for relative decisions ranged from approximately .67 to .91 when one form was administered on one occasion, with reliability increasing substantially as additional forms or occasions were included. Overall, the findings support the technical adequacy of the Grade 5 easyCBM® PRF measures for progress monitoring, while highlighting conditions under which measurement precision is optimized.

Figure 5. Example G-Coefficients from Technical Report 1220

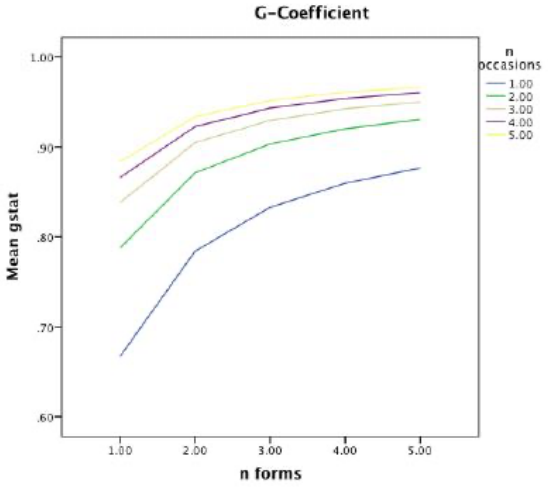


Table 30. Example Test-Retest Reliability Results Technical Report 1220

Table 3
Test-retest Reliability Results

Test Form	PRF5.8.2	PRF5.9.2	PRF5.10.2	PRF5.11.2	PRF5.12.2	PRF5.13.2	PRF5.14.2	PRF5.15.2
PRF5.8.1	0.94							
PRF5.9.1		0.91						
PRF5.10.1			0.88					
PRF5.11.1				0.93				
PRF5.12.1					0.90			
PRF5.13.1						0.90		
PRF5.14.1							0.54	
PRF5.15.1								0.91

Table 31. Example Correlation Results Technical Report 1220

Table 4
Correlation between Alternate Forms of Grade 5 Passage Reading Fluency Measure

Test Form	PRF5.9.2	PRF5.10.2	PRF5.11.2	PRF5.12.2	PRF5.13.2	PRF5.14.2	PRF5.15.2
PRF5.8.2	0.94	0.95	0.98	0.94	0.97		
PRF5.9.2		0.92		0.93	0.86	0.94	
PRF5.10.2				0.97	0.91	0.85	0.73
PRF5.11.2				0.95	0.91	0.88	0.92
PRF5.12.2					0.97		
PRF5.13.2						0.89	0.92
PRF5.14.2							0.92

Table 32. Summary of Key Findings from Technical Report 1220

Category	Summary
Sample	Convenience sample of Grade 5 students from a mid-sized school district in the Pacific Northwest. Students were nested within classrooms; G-theory analyses included between 11 and 18 students per teacher-level analysis, depending on form set.
Assessment Forms	easyCBM® **Passage Reading Fluency (PRF)** measures for Grade 5. Multiple parallel forms were examined across four teacher-specific analyses, including Forms 8–15, administered across two occasions using counterbalanced designs.
Analysis Method	Classical test theory analyses (bivariate correlations) for test–retest and alternate-form reliability, followed by **generalizability theory** using a fully crossed persons × forms × occasions design. G-studies estimated variance components; D-studies estimated absolute and relative error variances, standard errors of measurement, and G- and Phi-coefficients under varying numbers of forms and occasions.
Items Analyzed	Fluency scores rather than discrete items. PRF scores reflected passage reading rate, with forms treated as random samples from a universe of possible passages.
Problematic Items	No individual items were flagged as problematic. One form (PRF5.14) demonstrated weaker test–retest reliability (r ≈ .54), suggesting potential form-level instability rather than systematic measurement error.
Item Fit	Test–retest reliability was generally strong (r ≈ .88–.94 for most forms). Alternate-form correlations were moderate to strong (≈ .85–.98). G-theory analyses showed 57%–89% of variance attributable to persons, with minimal variance due to forms or occasions.
Overall Conclusion	Results provide moderate to strong evidence for the reliability and dependability of Grade 5 easyCBM® Passage Reading Fluency measures. While reliability varied across teacher-level analyses, findings support use for progress monitoring, with reliability substantially improving when multiple forms or occasions are used. In most cases, one form on one occasion yielded acceptable reliability for relative decisions.

Reference

Lai, C.-F., Park, B. J., Anderson, D., Alonzo, J., & Tindal, G. (2012). *An examination of test–retest, alternate form reliability, and generalizability theory study of the easyCBM® reading assessments: Grade* 5 (*Technical Report 1220).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1406: Internal Consistency of easyCBM® Vocabulary Measures: Grades 2–8 (Wray, Alonzo, et al., 2014).

Technical Report 1406 examines the internal consistency and split-half reliability of the easyCBM® Vocabulary measures for Grades 2 through 8. Using a very large extant dataset, the study evaluates whether the vocabulary measures produce stable and internally consistent scores suitable for benchmark screening and progress monitoring within a Response to Intervention (RTI) framework.

Methods

Data were drawn from the easyCBM® database and consisted of fall and winter benchmark vocabulary assessments administered during the 2013–2014 school year. The fall benchmark included 171,180 students across Grades 2–8, while the winter benchmark included 173,043 students across the same grades. Samples at each grade level ranged from approximately 17,000 to over 30,000 students. The dataset reflected substantial demographic diversity with respect to race, ethnicity, English learner status, disability status, and gender. Assessments were administered in group settings using either computer-based or paper-and-pencil formats.

Each easyCBM® Vocabulary assessment consisted of multiple-choice items in which students selected the correct meaning of a target word embedded in a sentence. Grade 2 measures contained 12 items, while Grades 3–8 contained 20 items. Three alternate forms per grade were used for benchmark screening, and all forms were designed to be of equivalent difficulty. Prior to analysis, records for students who did not respond to any items were removed, and datasets were screened for out-of-range values. Internal consistency was evaluated using Cronbach's alpha for each grade and season. Split-half reliability was examined by dividing each measure into first and second halves and computing reliability coefficients for each half, as well as correlations between halves. Additional item-level analyses compared performance of students in the top and bottom 27th percentiles to evaluate item discrimination.

Results

Reliability estimates were highly consistent across fall and winter administrations within each grade. Split-half reliability coefficients ranged from .61 to .75 for individual halves, with median values of .66 for first halves and .69 for second halves. Correlations between halves ranged from .58 to .72, indicating moderate to strong internal consistency. Item-level analyses showed that higher-performing students consistently answered items correctly at higher rates than lower-performing students, supporting appropriate item functioning. Overall, results provide strong evidence that easyCBM® Vocabulary measures demonstrate adequate internal consistency and reliability for use in educational screening and progress monitoring.

Table 33. Cronbach's Alpha and Split-Half Reliability from Technical Report 1406

Table 5

Internal Reliability: easyCBM© Vocabulary

Grade/Time	Cronbach's Alpha	Split-half Reliability		
		1st Half	2nd Half	Correlation
2/F	.83	.69	.75	.70
2/W	.84	.73	.73	.72
3/F	.84	.72	.75	.69
3/W	.82	.70	.72	.67
4/F	.82	.71	.72	.65
4/W	.81	.65	.71	.64
5/F	.79	.62	.68	.63
5/W	.78	.61	.68	.59
6/F	.80	.67	.69	.62
6/W	.81	.65	.64	.65
7/F	.79	.66	.67	.64
7/W	.78	.64	.66	.61
8/F	.79	.66	.69	.61
8/W	.76	.65	.64	.58

Table 34. Summary of Key Findings from Technical Report 1406

Category	Summary
Sample	Extant benchmark data from the easyCBM® database collected during the 2013–2014 school year. The sample included over **170,000 students** across Grades 2–8 for fall and winter administrations, representing a diverse population in terms of ethnicity, ELL status, disability status, and gender.
Assessment Forms	easyCBM© **Vocabulary** benchmark measures for Grades 2–8. Each grade included **three benchmark forms** and **ten progress-monitoring forms**, all designed to be of equivalent difficulty and administered via computer or paper-and-pencil formats.
Analysis Method	Internal consistency was evaluated using **Cronbach's alpha**. **Split-half reliability** was calculated by comparing first-half and second-half item sets (divided at the median), along with correlations between halves. Item-level analyses examined performance differences between the top and bottom 27th percentiles.
Items Analyzed	Vocabulary items consisting of a sentence with a bolded target word and three multiple-choice options. Grade 2 measures included **12 items**; Grades 3–8 included **20 items** per form.
Problematic Items	No items were identified as problematic. All items demonstrated expected discrimination patterns, with higher-performing students answering items correctly at higher rates than lower-performing students. Zero-variance items appeared only where all top-performing students answered correctly, consistent with expectations.
Item Fit	Internal consistency was strong across grades, with **Cronbach's alpha ranging from .76 to .84** and a median of **.81**. Split-half reliability ranged from **.61 to .75**, with median correlations between halves of **.64**, indicating acceptable internal structure across forms and grades.
Overall Conclusion	Results provide strong evidence for the internal consistency and reliability of the easyCBM® Vocabulary measures across Grades 2–8. Findings support their use for benchmark screening and progress monitoring, with consistent performance across grades, seasons, and forms indicating sound measurement quality and form equivalence.

Reference

Wray, K. A., Alonzo, J., & Tindal, G. (2013). *Internal consistency of the easyCBM® vocabulary measures: Grades 2–8 (Technical Report 1406).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1407: Internal Consistency of the easyCBM® CCSS Reading Measures Grades K-8 (Guerreiro et al., 2014).

Technical Report 1407 examines the internal consistency and split-half reliability of the easyCBM® CCSS Reading benchmark assessments for Grades 3 through 8. The purpose of the study was to evaluate the reliability of the reading measures used for benchmarking and progress monitoring, particularly given their role in high-frequency instructional decision making. The CCSS Reading measures were designed to assess comprehension across literary text, informational text, and read-to-perform-a-task formats aligned with Common Core State Standards.

Methods

Data for the study were drawn from an extant dataset collected during the 2013–2014 school year and included fall and winter benchmark scores from more than 150,000 students across grades 3 through 8. Sample sizes varied by grade and benchmark period, with fall administrations including approximately 75,900 students and winter administrations including approximately 81,200 students. The dataset reflected substantial demographic diversity, including variation in race, ethnicity, English learner status, disability status, and gender, although some demographic variables were missing for grades and testing windows.

The easyCBM® CCSS Reading measures consist of five passage-based testlets per form, each containing five multiple-choice items, for a total of 25 items per assessment. Students receive one point for each correct response, yielding a maximum possible score of 25. Measures are administered online in group settings, with automated scoring and data capture. All benchmark and progress monitoring forms within each grade were developed concurrently and designed to be of equivalent difficulty. Prior to analysis, records with no item responses and out-of-range values were removed. Internal consistency was evaluated using Cronbach's alpha, while split-half reliability was examined by comparing first-half and second-half item sets and computing correlations between halves. Item-level analyses comparing performance of students in the upper and lower 27th percentiles were also conducted to examine discrimination patterns.

Results

Results indicated strong internal consistency across grades and benchmark periods. Cronbach's alpha values ranged from .83 to .90, with a median of .87 across all measures. Split-half reliability coefficients ranged from .73 to .86, with median values of .76 for first halves and .82 for second halves. Item-level analyses demonstrated consistent discrimination, with higher-performing students answering items correctly more frequently than lower-performing students. Overall findings provide strong evidence supporting the reliability and technical adequacy of the easyCBM CCSS Reading measures for use in benchmarking and progress monitoring.

Table 35. Summary of Results for Technical Report 1407

Category	Summary
Grades	Grades 3–8
Assessment Focus	CCSS-aligned reading comprehension
Sample Size	Over 150,000 students across fall and winter benchmarks
Reliability Evidence	Median Cronbach's alpha = .87; strong split-half reliability
Statistical Methods	Cronbach's alpha and split-half reliability
Item Performance	Consistent discrimination across upper and lower ability groups
Intended Use	Benchmarking and progress monitoring

Summary of Technical Report 1807: Supplementary report on easyCBM MCRC Measures: A Follow-up to previous technical report (Alonzo & Anderson, 2018).

A large extant data set was used. Reliability results demonstrated adequate internal consistency across grades. Cronbach's alpha values ranged from approximately .66 (Grade 8) to .79 (Grade 4), with most grades clustering between .71 and .79. Split-half reliability coefficients showed similar patterns, ranging from approximately .60 to .80. These values indicate acceptable reliability for screening and benchmarking purposes, particularly given the time-limited nature of the assessments. Subgroup analyses further supported reliability, with higher internal consistency observed in general education and special education subgroups, often exceeding .80. Gender and ethnicity subgroup analyses demonstrated consistent reliability estimates across groups, supporting fairness and generalizability of the measures

References

Guerreiro, M., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM® CCSS reading measures, Grades 3–8 (Technical Report No. 1407).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Anderson, D. (2018). *Supplementary report on easyCBM MCRC Measures: A Follow-up to previous technical report (Technical Report # 1807).* Eugene OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1803: Reliability of the Slope of the easyCBM® Reading Measures (Nese et al., 2018).

Methods

The analytic sample comprised students in Grades K–8 who were identified by their school districts as needing intensive intervention and who had completed easyCBM® reading progress monitoring assessments during the 2014–2015, 2015–2016, and 2016–2017 school years. Inclusion required a minimum of 10 assessment scores per measure with at least 20 weeks between the first and last administration, resulting in sample sizes that varied by grade and measure. Six reading measures were examined: Letter Names (LN), Phoneme Segmenting (SEG), Letter Sounds (LS), Word Reading Fluency (WRF), Passage Reading Fluency (PRF), Vocabulary (VOC), Multiple Choice Reading Comprehension (MCRC), and CCSS Reading (CCSS). Slope reliability was evaluated using two approaches: (1) Pearson split-test correlation analysis, in which each student's assessments were split into odd- and even-numbered subsets and OLS growth slopes were correlated, and (2) a reliability-of-slope approach, defined as the ratio of true score variance (random slope variance from a mixed-effects growth model) to total variance, computed in R.

Results

Results varied systematically by measure, grade level, and analytic approach, with the most consistent and favorable reliability estimates observed for early literacy measures in the lower grades. For the Pearson split-test correlation approach, the early/emergent literacy measures—Letter Names, Phoneme Segmenting, and Letter Sounds—demonstrated reasonable reliability ($r \geq .60$, with lower confidence interval bounds meeting or exceeding .40) in Kindergarten and Grade 1. Similarly strong results were obtained for Word Reading Fluency in Grade 1 ($r = .88$) and Passage Reading Fluency in Grade 1 ($r = .85$) and ranged from .25 to .42 in Grades 2–8. Vocabulary and MCRC measures showed largely weak and inconsistent correlations across grades.

The reliability-of-slope approach yielded somewhat more favorable results overall, particularly for the fluency-based measures. Word Reading Fluency demonstrated strong reliability through Grade 2 ($r = .75$–1.00), and Passage Reading Fluency showed acceptable reliability through Grade 3 ($r = .56$–$.88$). The early literacy measures again performed well in Grades K–1, with coefficients ranging from .72 to .87. In contrast, Vocabulary, MCRC, and CCSS Reading showed weak reliability across grades under this approach as well, with many coefficients below .40 and wide confidence intervals, reflecting high variability and small sample sizes for several grade levels.Across both analytic approaches, a clear developmental pattern emerged: slope reliability was strongest in the earliest grades and for measures targeting foundational reading skills, and weakest in the later grades (Grade 4 and above) and for higher-order comprehension and vocabulary measures.

Table 36. Illustrative Results from Technical Report 1803

Table 2
Reliability of Slope Analyses Results

Measure	Grade	*n*	Correlation coefficient (*r*)	Confidence Interval	
				Lower	Upper
Letter Names (LN)	K	253	.87**	.74	1.00
	1	44	.79**	.50	1.00
Phoneme Segmenting (SEG)	K	42	.83**	.53	1.00
	1	91	.72**	.51	.98
Letter Sounds (LS)	K	237	.86**	.72	1.00
	1	199	.72**	.57	.89
Word Reading Fluency (WRF)	K	10	1.00**	.45	1.00
	1	348	.92**	.81	1.00
	2	155	.75**	.58	.94
	3	60	.54	.29	.87
Passage Reading Fluency (PRF)	1	75	.88**	.65	1.00
	2	412	.62**	.52	.74
	3	387	.56**	.45	.68
	4	163	.53	.37	.72
	5	263	.42	.29	.56
	6	64	.48	.23	.81
	7	11	.03	NA	NA
	8	137	.40	.22	.62
Vocabulary (VOC)	2	41	.07	.00	.42
	3	7	.47	.00	1.00
	4	22	.48	.08	1.00
	5	17	.77	.33	1.00
	6	7	.44	.00	1.00
	7	1	NA	NA	NA
	8	32	.08	.00	.55

Note. NA = Not analyzed. **Lower bound of the confidence interval around the median correlation estimate falls below 0.50 but meets or exceeds 0.40.

References

Nese, J. F. T., Anderson, D., Irvin, P. S., & Alonzo, J. (2018). *In-brief: Reliability of the slope of the easyCBM® reading measures (Technical Report No. 1803).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1111: easyCBM® Slope Reliability: Letter Names, Word Reading Fluency, and Passage Reading Fluency (Patarapichayatham et al., 2011).

Methods

Participants were students in Kindergarten through Grade 8 drawn from an extant database of approximately 170,000 students per grade who used the easyCBM® system during the 2010–2011 school year. After data cleaning—excluding students with fewer than three scores, off-grade-level testing, scores at or above the 50th percentile on the first progress monitoring measure, and zero scores—final sample sizes ranged from 122 to 1,146. Three curriculum-based measures (CBMs) were examined: Letter Names (K–1), Word Reading Fluency (K–3), and Passage Reading Fluency (grades 1–8), each administered in 60-second timed trials across up to 20 alternate forms. Raw scores were averaged within weeks and then collapsed into bi-weekly segments. Using structural equation modeling (SEM) in Mplus, a parallel-process linear growth model was applied by alternating bi-weekly segments into two parallel series (Process A and Process B). Individual growth slopes were estimated as empirical Bayes factor scores within each process, and the Spearman-Brown formula was applied to the inter-process slope correlation to yield a reliability coefficient for the full measurement period.

Results

The results of this study provided strong evidence for the reliability of growth slope estimates across all three easyCBM® reading measures and the grades in which they were administered. Across all analyses, Spearman-Brown corrected slope reliability coefficients were universally above .80, with the majority exceeding .85, indicating that the rate of change captured by these measures is highly stable and replicable across equivalent time samples within a school year.

For Letter Names, reliability estimates were .859 for kindergarten and .753 for grade 1, indicating good-to-strong slope stability in the early literacy domain. Although the grade 1 coefficient was the lowest among Letter Names estimates, it still exceeded the conventional threshold of .70 generally considered acceptable for educational measurement purposes.

Word Reading Fluency yielded slope reliabilities of .812 (grade 1), .881 (grade 2), and .853 (grade 3), reflecting strong consistency across the three grades examined. These findings suggest that this measure effectively captures stable growth trajectories in students' word-level reading skills over the course of a school year.
Passage Reading Fluency demonstrated robust slope reliability across grades 1 through 7, with coefficients ranging from .711 to .869. Notably, grades 2 through 6 all produced estimates above .77, with several grades yielding reliabilities exceeding .82. Grade 1 produced the lowest estimate in this measure (.711), likely reflecting greater variability in early oral reading fluency growth trajectories. The one clear exception to the otherwise high reliability pattern was grade 8, which yielded a substantially lower coefficient of .495. This grade 8 estimate falls below acceptable thresholds and suggests that the Passage Reading Fluency measure may not reliably track individual growth slopes among older, more advanced readers who are closer to the ceiling of the measurement scale.
Taken together, these findings support the use of easyCBM® measures as reliable tools for monitoring student learning trajectories within a response to intervention (RTI) framework, particularly for students in the earlier grades or those performing below grade-level expectations. The sole notable concern is the grade 8 Passage Reading Fluency reliability, which warrants caution when interpreting slope-based decisions for students at that level. The overall pattern reinforces the value of these measures for identifying whether individual students are responding to instructional interventions across the academic year.

Table 37. Illustrative Results from Technical Report 1111

Table 14
Passage Reading Fluency Descriptive Statistics: Grade 7

Time Point	N	Minimum	Maximum	Mean	Std. Deviation
1A	51	44	150	112.82	24.775
1B	43	85	175	132.47	20.536
2A	109	36	170	112.83	26.431
2B	175	30	175	116.14	27.849
3A	118	23	187	120.32	26.762
3B	81	28	203	121.96	27.377
4A	**192**	**4**	**185**	**108.18**	**34.597**
4B	**73**	**40**	**178**	**107.96**	**30.366**
5A	**86**	**44**	**177**	**108.37**	**29.759**
5B	106	16	170	115.05	31.199
6A	141	22	192	128.26	30.802
6B	140	19	183	114.52	28.089
7A	128	45	175	119.52	27.167
7B	91	34	189	116.29	31.463
8A	90	40	186	128.24	26.770
8B	107	63	185	125.76	27.152
9A	149	23	196	123.50	33.089
9B	67	24	180	110.44	31.384
10A	**14**	**60**	**154**	**110.29**	**30.603**
10B	**0**				
Valid N (listwise)	0				

Reference

Patarapichayatham, C., Anderson, D., Irvin, P. S., Kamata, A., Alonzo, J., & Tindal, G. (2011). *easyCBM® slope reliability: Letter names, word reading fluency, and passage reading fluency (Technical Report No. 1111)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 2602: Reliability Analyses for easyCBM® Measures in Grades 3-8: Total Scores for Vocabulary, Proficient Reading, and Proficient Math (Tindal & Nese, 2026).

Methods

TR2602 reports internal-consistency reliability for three easyCBM® benchmark measures in Grades 3–8: Vocabulary, Proficient Reading, and Proficient Math. Student records were drawn from districts that permitted research use of easyCBM® server data. Each record included grade and available demographics plus item-level 0/1 responses (0 = incorrect, 1 = correct).

Methods

Analyses were limited to Benchmark administrations and split into Fall and Winter benchmark files. Results were reported by grade (3–8) and season, alongside demographics and descriptive statistics (n, mean, SD, range, median, skewness, kurtosis) to describe score distributions.

Reliability was estimated using two complementary approaches. First, Rasch marginal reliability (IRT) was computed from expected a posteriori (EAP) theta estimates and their standard errors. Rasch models were fit in R with the TAM package using marginal maximum likelihood; marginal reliability summarizes the proportion of variance in estimated ability attributable to true differences rather than expected error. Median coefficients and 95% confidence intervals were obtained using 500 bootstrap replications for each grade-by-season cell. Second, Cronbach's alpha (KR-20 for dichotomous items) was estimated in R with the psych package and similarly summarized with bootstrapped medians and 95% intervals. Findings are presented separately for each measure. Both indices were interpreted as evidence of score precision for screening uses.

Results: Vocabulary

Vocabulary (20 items) showed strong reliability across grades and both benchmark seasons. Fall means rose from Grade 3 (M = 14.30, SD = 4.49) to Grades 6–8 (M ≈ 16.40–16.51), with medians of 16–18 and negative skewness. Winter means were slightly higher for most grades (e.g., Grade 3 M = 15.16; Grade 6 M = 16.46; medians 17–18).

Rasch marginal reliability medians ranged from 0.74 to 0.81. The highest estimate was Grade 3 Fall (median = 0.81; 95% CI 0.81–0.82). Most grade-by-season cells clustered between 0.74 and 0.76 (e.g., Grade 5 Winter median = 0.74; CI 0.73–0.75; Grade 8 Fall median = 0.75; CI 0.74–0.75).

Cronbach's alpha converged with the IRT results. Median alphas ranged from 0.80 to 0.86, highest in Grade 3 (Fall 0.85; Winter 0.86) and lowest in Grade 5 Winter (median = 0.80; lower bound = 0.79). All other alpha lower bounds were at least 0.80. The report concludes Vocabulary total scores provide dependable internal consistency for benchmark screening in Grades 3–8 in Fall and Winter. Confidence intervals were narrow because of large n. Both indices indicate consistent total-score measurement. across grades and benchmark seasons overall.

Results: Proficient Reading

Proficient Reading (20 items) showed adequate-to-strong reliability, but more variability than Vocabulary. Fall means ranged from Grade 3 (M = 10.76, SD = 3.97) to Grade 6 (M = 13.70, SD = 3.53), with medians 11–15. Winter means ranged from Grade 3 (M = 9.87, SD = 3.21) to Grade 5 (M = 14.71, SD = 3.98), with medians 10–16.

Rasch marginal reliability medians ranged from 0.60 to 0.78. Grade 3 Winter was lowest (median = 0.60; 95% CI 0.59–0.61). Most other cells fell in the low-to-mid 0.70s (e.g., Grade 4 Fall 0.78; Grade 5 Winter 0.77; Grade 7 Fall 0.75), with lower values in Grade 6 (Fall 0.68; Winter 0.69) and Grade 8 Winter (0.69).

Cronbach's alpha mirrored this pattern (medians 0.60–0.81). Grade 3 Winter was again lowest (median = 0.60; lower bound = 0.59) and Grade 5 Winter highest (median = 0.81; 95% CI 0.80–0.81). Grade 8 Winter was the only grade/season with an alpha lower bound below 0.70 (0.69). The report concludes Proficient Reading generally supports benchmark screening, while highlighting Grade 3 Winter (and modestly lower Grade 6/8 winter bounds) for monitoring. Confidence intervals were similarly narrow for most grades.

Tindal, G., & Nese, J. F. T. (2026). *Reliability Analyses for easyCBM® Measures in Grades 3-8: Total Scores for Vocabulary, Proficient Reading, and Proficient Math (Technical Report 2602)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1701: Oral reading fluency: Outcomes from 30 years of research (Tindal, 2017).

Methods

Technical Report 1701 synthesizes approximately three decades of empirical and methodological research on Oral Reading Fluency (ORF) as a curriculum-based measurement approach used for screening, benchmarking, and progress monitoring in reading. The report's purpose is to document how ORF has been operationalized across studies, summarize evidence regarding growth, reliability, and validity, and describe measurement design features that influence score interpretation for instructional decision making.

ORF is described as a standardized, brief passage-reading task in which students read aloud for a fixed duration, most commonly 60 seconds, and performance is scored as words correct per minute (WCPM). Scoring conventions emphasize accurate word identification, with errors recorded for mispronunciations, substitutions, omissions, and hesitations; self-corrections are typically not counted as errors. Passage development and selection are treated as central to score comparability. The report reviews approaches used to construct passage sets that approximate within-grade difficulty targets, including the use of readability indicators, empirical calibration, and procedures intended to manage passage-to-passage variation that can obscure true change.

Evidence summarized in the report draws from multiple datasets and published studies spanning elementary through secondary grades. The reviewed work includes small investigations of special populations as well as large, regionally or nationally stratified samples used to generate norms and seasonal growth expectations, commonly covering Grades 1 through 8 and extending to Grades 9 through 12 in later work. Data collection practices described across studies typically involved universal screening during fall, winter, and spring benchmark windows, with additional administrations at shorter intervals when ORF was used for progress monitoring. Administration procedures were generally standardized, including directions, timing rules, and scoring guidelines designed to support comparability across examiners and occasions.

Analytic methods highlighted in the synthesis include classical test theory (CTT) indices of reliability and stability, generalizability analyses used to partition error attributable to passages and occasions, and item response theory (IRT) models used in selected research to evaluate passage difficulty and growth. The report cites prior work using a Rasch model to analyze ORF growth and to support development of norms across grades. Validity evidence is summarized through correlational and predictive analyses linking ORF to broader reading outcomes, including associations with standardized reading tests and large-scale assessments, as well as studies examining sensitivity to instructional change.

Results

Across the body of evidence, ORF scores exhibit systematic growth patterns, with growth rates varying by grade level and season. Reliability is generally sufficient for screening and progress monitoring when multiple passages or repeated administrations are used, whereas single-passage estimates are more sensitive to passage-to-passage variability. Overall, the report concludes that ORF provides an efficient indicator of reading proficiency and growth when passage sets, administration procedures, and interpretive benchmarks are aligned with intended uses in data-based decision-making frameworks.

Table 38. Summary of Results from Technical Report 1701

Area	Oral Reading Fluency (ORF) is positioned as a curriculum-based indicator for screening, benchmarking, and progress monitoring of reading proficiency.
Item Functioning	Passage selection and calibration are identified as primary determinants of score comparability; passage-to-passage variability can meaningfully affect WCPM unless managed through design and administration controls.
Reliability	Across the research base, ORF demonstrates adequate reliability for intended uses when multiple passages or repeated administrations are used; single-passage estimates are less stable.
Validity Evidence	ORF scores show moderate to strong relationships with broader reading outcomes and standardized measures, supporting convergent and predictive validity for decision making in tiered support systems.
Population Coverage	Evidence spans elementary through secondary grades and includes both general and special populations, with interpretive guidance emphasizing alignment between norms, grade level, and administration season.

Reference

Tindal, G. (2017). *Oral reading fluency: Outcomes from 30 years of research (Technical Report No. 1701).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Appendix A: Technical Report Table and Figure Titles

Table 1. Grade-Level Benchmark Measures by Assessment Period
Table 2. Reliability for Various Measures Across Grades
Table 3. Individual Coefficients from Technical Reports
Table 4. Median Coefficients from Technical Reports
Table 5. Grade Group Summary of Coefficients from Technical Reports
Table 6. Tests of Within-Subject Effects and Contrasts from Technical Report 1201
Table 7. Sample Split-Half Coefficients from Technical Report 1201
Table 8. Distractor Analysis from Technical Report 1201
Table 9. Summary of Key Findings from Technical Report 1201
Table 10. Example Tests of Within-Subject Effect from Technical Report 1202
Table 11. Summary of Outcomes from Technical Report 1202
Table 12. Example Tests of Within-Subjects Effects from Technical Report 1203
Table 13. Example Split Half Coefficients from Technical Report 1203
Table 14. Summary of Key Findings from Technical Report 1203
Table 15. Example of Tests within Subjects from Technical Report 1204
Table 16. Summary of Key Findings from Technical Report 1204
Table 17. Example of Within-Subject Effects from Technical Report 1205
Table 18. Example of Within-Subject Contrasts from Technical Report 1205
Table 19. Summary of Key Findings from Technical Report 1205
Table 20. Example Split-Half Coefficients from Report 1206
Table 21. Example Within-Subject Effects from Report 1206
Table 22. Summary of Key Findings from Technical Report 1206
Table 23. Example Test-Retest Reliability from Technical Report 1216
Table 24. Example D-Study G and ϕ Coefficients from Technical Report 1216
Table 25. Example Test-Retest Reliability Word Reading Fluency Measures Technical Report 1217
Table 26. Summary of Key Findings from Technical Report 1217
Table 27. Example Test-Retest Reliability Word Reading Fluency Measures Technical Report 1218
Table 28. Summary of Key Findings from Technical Report 1218
Table 29. Example Correlations for Passage Reading Fluency Measures in Technical Report 1219
Table 30. Example Test-Retest Reliability Results Technical Report 1220
Table 31. Example Correlation Results Technical Report 1220
Table 32. Summary of Key Findings from Technical Report 1220
Table 33. Cronbach's Alpha and Split-Half Reliability from Technical Report 1406
Table 34. Summary of Key Findings from Technical Report 1406
Table 35. Summary of Results for Technical Report 1407
Table 36. Illustrative Results from Technical Report 1803
Table 37. Illustrative Results from Technical Report 1111
Table 38. Summary of Results from Technical Report 1701

Figure 1. Example G-Coefficients from Technical Report 1216
Figure 2. Example G-Coefficients from Technical Report 1217
Figure 3. Example G-Coefficients from Technical Report 1218
Figure 4. Example G-Coefficients from Technical Report 1219
Figure 5. Example G-Coefficients from Technical Report 1220

Appendix B: Guide to Spreadsheet Technical Report Value Displays

See Riverside Insights or BRT to access exact values for TR Summaries
2603-RK8R_ReliabilityReadingTables.xlsx

- TR0906
- TR1003
- TR1004
- TR1403
- TR1201
- TR1202
- TR1203
- TR1204
- TR1205
- TR1206
- TR1216
- TR1217
- TR1218
- TR1219
- TR1220
- TR1406
- TR1407
- TR1807
- TR1803
- TR1111
- TR2602
- TR1701…See Tech Report 2603-RK8R-Reliability Reading - Page 44

Technical Report References

Alonzo, J., & Anderson, D. (2018). *Supplementary report on easyCBM MCRC Measures: A Follow-up to previous technical report (Technical Report # 1807)*. Eugene OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Lai, C. F., Anderson, D., Park, B. J., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 4 (Technical Report # 1219)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). *Alternate form and test-retest reliability of easyCBM reading measures (Technical Report # 0906)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Lai, C. F., Park, B. J., Alonzo, J., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 2 (Technical Report # 1217)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Park, B. J., Lai, C., F., Alonzo, J., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 1 (Technical Report # 1216)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Guerreiro, M., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM CCSS reading measures grades K-8 (Technical Report # 1407)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Alonzo, J., Lai, C. F., Park, B. J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 7 (Technical Report # 1206)*. Eugegne, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Alonzo, J., Park, B. J., Lai, C. F., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 6 (Technical Report # 1205)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Sáez, L., Anderson, D., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM Grade 2 reading measures (Technical Report # 1004)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Irvin, P. S., Alonzo, J., Park, B. J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 2 (Technical Report # 1201)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 3 (Technical Report # 1202)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Nese, J. F. T., Jamgochian, E. M., Kamata, A., Anderson, D., Park, B. J., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM primary-level reading measures (Grades K-1), 2009-2010 Version. (Technical Report # 1003)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon..

Lai, C. F., Park, B. J., Anderson, D., Alonzo, J., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 5 (Technical Report # 1220)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nese, J. F. T., Anderson, D., Irvin, P. S., & Alonzo, J. (2018). *In-Brief: Reliability of the slope of the easyCBM® reading measures (Technical Report # 1803)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Anderson, D., Alonzo, J., Lai, C. F., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 3 (Technical Report # 1218)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Alonzo, J., Lai, C. F., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 4 (Technical Report # 1203)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Lai, C. F., Alonzo, J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 5 (Technical Report # 1204)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Patarapichayatham, C., Anderson, D., Irvin, P. S., Kamata, A., Alonzo, J., & Tindal, G. (2011). *easyCBM slope reliability: Letter names, word reading fluency, and passage reading fluency (Technical Report # 1111)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G. (2017). *Oral reading fluency: Outcomes from 30 years of research (Technical Report # 1701)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., & Nese, J. F. T. (2026). *Reliability Analyses for easyCBM® Measures in Grades 3-8: Total Scores for Vocabulary, Proficient Reading, and Proficient Math (Technical Report # 2602)*. Eugene, OR.: Behavioral Research and Teaching, University of Oregon.

Wray, K., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM vocabulary measures Grades 2-8 (Technical Report # 1406)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Wray, K., Lai, C. F., Sáez. L., Alonzo, J., & Tindal, G. (2014). *easyCBM beginning reading measures: Grades K-1 alternate form reliability and criterion validity with the SAT-10 (Technical Report # 1403)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Conclusions Supporting Claims for Criterion 2.1: Overall Achievement

The interim assessment provides for valid inferences about a student's current overall achievement in the target content domain.

2.1.c Achievement scores support intended interpretations of student performance.

Evidence is provided; equating/linking procedures are described; item specifications ensure consistency; there is empirical evidence and an active research agenda. Evidence supporting intended interpretations is provided across multiple validity frameworks. Criterion-related evidence is primary: early literacy measures (Letter Names, Letter Sounds, Phoneme Segmenting, Word Reading Fluency) demonstrate strong correlations with DIBELS and TOWRE (rs = .75–.95, TR 1309), and regression models predicting SAT-10 Word Reading explain 48–77% of variance in Grades K–1 (TR 1003, TR 1403). Internal structure evidence from TR 1302 supports a theoretically coherent three-factor model (Early Literacy, Fluency, Comprehension) in Grades K–2 via CFA, with strong factor loadings for fluency and foundational literacy indicators. Rasch calibration in TR 0030 and TR 0025 places item difficulty estimates on a common logit scale, supporting form comparability across administrations. Item development specifications in TR 0030 include eight explicit design criteria, standardized timed fluency formats, and post-pilot IRT review, ensuring consistency in item presentation and format across test events. The document reflects an active research agenda spanning 2004–2018 across Grades K–8, integrating criterion validity, classification accuracy, internal structure, and growth modeling.

2.1.d Achievement scores are appropriate for supporting their intended uses.

Intended uses are clearly articulated; sufficient theoretical and empirical evidence supports the intended uses. The intended uses of easyCBM® reading achievement scores are consistently and clearly articulated throughout the document. The introduction and all TR summaries frame the measures as tools for universal screening, benchmark monitoring, and instructional decision-making within Multi-Tiered Systems of Support (MTSS) and Response to Intervention (RTI) frameworks, with the explicit goal of identifying students at risk of reading failure and supporting tier placement decisions. Sufficient theoretical and empirical evidence supports these uses: criterion-related validity from regression studies (TR 0910, TR 1003, TR 1004, TR 1403, TR 1806) documents meaningful relationships between easyCBM® scores and external reading measures across Grades K–8. Classification accuracy evidence (TR 1106, TR 1107, TR 1801, TR 1805) documents AUC values of .75–.95 for most K–2 measures and acceptable discrimination for Grades 3–8. The document's conclusion states that cumulative evidence supports the interpretation of easyCBM® reading scores for screening, progress monitoring, and instructional decision-making within MTSS systems.

Conclusions Supporting Claims for Criterion 2.2: Predicted Student Performance

The interim assessment provides valid information regarding predicted student performance on a state's summative assessment or other intended criterion measure(s).

2.2.a The design of the interim assessment supports its use in predicting performance on one or more external measures.

Sufficient information evaluates construct similarity; the intended use does not invalidate prediction; evidence for specific assessments is provided. The design of easyCBM® reading measures supports their use in predicting performance on multiple external criterion measures. The introduction documents that easyCBM® reading targets early literacy, oral reading fluency, vocabulary, and comprehension—constructs closely aligned with those assessed by the SAT-10, OAKS, MSP, and SBAS ELA criterion measures. Measures vary by grade to maintain construct alignment: foundational literacy tasks (K–2) predict SAT-10 reading subtests, while fluency and comprehension measures (Grades 3–8) predict state proficiency assessments. The intended use of easyCBM® as a CBM screening tool is consistent with—not contradictory to—its use for predicting state assessment performance; CBM screening tools are designed to correlate with criterion outcomes. Although not aligned to ACT or SAT, the measures are explicitly validated for Oregon (OAKS), Washington (MSP), and national accountability (SBAS ELA) criterion

measures, with regression and ROC-based prediction evidence across these assessments documented in TR 0910, TR 1106, TR 1107, TR 1806, and the cross-validation reports.

2.2.b Predicted results are reliable.

Procedures for calculating reliability are documented; reliability is consistent with intended inferences; predictions demonstrate sufficient reliability. Procedures for evaluating the reliability of predicted results are well-documented across technical reports. For classification-based predictions, ROC analyses in TR 1106, TR 1107, TR 1801, and TR 1805 report AUC statistics with 95% confidence intervals, sensitivity, specificity, and overall classification rates. For regression-based predictions, TR 0910, TR 1003, TR 1004, and TR 1403 report R, R^2, adjusted R^2, standard error of estimate, and model F statistics. These procedures are appropriate to their respective prediction types and consistent with the inferences they support: ROC-based indices are appropriate for binary RTI tier-placement decisions; R^2 is appropriate for continuous predictive relationships. Cross-validation studies in TR 1108 (Oregon) and TR 1109 (Washington) directly evaluate cut score stability across independent split samples, with overlapping AUC confidence intervals confirming that classification predictions are replicable. PRF cut score differences averaged 1–2 CWPM across Oregon split groups and approximately 7 CWPM in Washington, indicating acceptable stability for screening purposes.

2.2.c Predicted results reflect a student's likely performance on the state summative assessment or other intended criterion measure(s).

Data and procedures are documented; procedures supporting intended interpretations are clearly articulated; studies support the appropriateness of predicted results. Data and procedures used to establish predictive relationships are documented and reasonable. Sample characteristics are described in each TR: TR 0910 used Grades 3–8 students from two Oregon districts (hundreds to over 1,000 per grade); TR 1806 used Grades 3–8 students from Pacific Northwest districts during 2014–2015; TR 1107 used approximately 900 students per grade from two Washington districts. Prediction procedures are clearly articulated throughout: regression models predicting OAKS from easyCBM® measures (TR 0910) explain R^2 of .50–.65 across grades; PRF correlates .57–.68 with SBAS ELA in Grades 3–8 (TR 1806); concurrent PRF-DIBELS ORF correlations are very strong (.88–.95). ROC-based classification decision rules are explicitly documented in TR 1106 and TR 1107, with sensitivity-prioritized thresholds consistent with RTI false-negative minimization goals. Cross-validation of cut scores in Oregon and Washington (TR 1108, TR 1109) demonstrates stable AUC values and minimal cut score variation across independent samples.

2.2.d Predicted results are appropriate for supporting their intended uses.

Intended uses for predicted results are clearly articulated; sufficient theoretical and empirical evidence supports the appropriateness of intended uses. The intended uses for predicted results are clearly and consistently articulated throughout the document. All TR summaries describe predictions as supporting RTI/MTSS screening decisions: identifying students at risk of failing state assessments, placing students in tiered intervention services, and informing benchmark monitoring. Cut scores are framed as tools for risk identification with sensitivity prioritized to minimize false negatives in high-stakes RTI contexts. The document appropriately notes that cut scores are criterion-dependent and should not be generalized across states without validation, reflecting awareness of use boundaries. Sufficient theoretical and empirical evidence supports these uses: oral reading fluency as a predictor of general reading competence is grounded in long-established CBM theory, supported by updated ORF norms spanning 1992–2017 (TR 1702). Empirical evidence across Grades K–8 documents consistent predictive relationships between easyCBM® measures and SAT-10, OAKS, MSP, and SBAS ELA. The document's conclusion integrates these findings to support defensible, MTSS-aligned predictions across Grades K–8.

Introduction

In the *Standards for Educational and Psychological Testing* (2014)[1], "validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of the test. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests…It is the interpretations of the test for proposed uses that are evaluated, not the test scores themselves (p. 11)" (American Educational Research Association et al., 2014). The Standards emphasize a unified view of validity: diverse evidence is integrated into a coherent argument that links the construct to the score meaning, the uses, and the decisions made. The *Standards* describe several major sources of validity evidence. Evidence based on test content evaluates how well the tasks, items, and scoring align with the intended construct and domain. Evidence based on response processes examines the cognitive, behavioral, or rater processes engaged by examinees and scorers and whether they match the construct definition. Evidence based on internal structure evaluates the relationships among items, components, and scores (e.g., dimensionality, factor structure, invariance) and their consistency with the construct. Evidence based on relations to other variables examines expected convergent, discriminant, criterion, and group-difference relations. Evidence based on consequences of testing considers intended and unintended outcomes, including fairness, to the extent that consequences provide information about score meaning and use. Together, these sources justify interpretations and identify limitations.

In this technical report, four evidential sources are used to demonstrate validity for using easyCBM® early reading (K-2) measures and later measures in reading for students in Grades 3-8: (a) criterion related to demonstrate the reading constructs are related to other acceptable measures of reading. (b) regression analyses to support predictions of performance and accounted variance in other acceptable measures of reading, (c) internal structures to reflect uniform constructs in the measures, and (d) classification accuracy to use the measures for indicating risk of potential learning problems. are reported for the. These sources of evidence are applied to the following measures.

Table 1. Grade-Level Benchmark Measures by Assessment Period

Grade Level	Fall Benchmark	Winter Benchmark	Spring Benchmark
K	Letter Names Letter Sounds Phoneme Segmenting	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Phoneme Segmenting Word Reading Fluency
1	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency
2	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading

In Grades 3-8, the measures are Vocabulary, Passage Reading Fluency, and Comprehension (Basic and Proficient Reading).

Validity, as defined in the Standards for Educational and Psychological Testing (2014), refers to the degree to which evidence and theory support the interpretations of test scores for their intended uses. The unified framework emphasizes multiple sources of evidence, including evidence based on **relations to other variables** (criterion-related), internal structure, response processes, and consequences of testing. Across the easyCBM® Reading Technical Reports, a coherent and cumulative validity argument emerges grounded primarily in relations to external criteria, construct coherence, and classification accuracy within MTSS frameworks.

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Evidence based on **relations to other variables** is consistently strong. Across grades K–2, early literacy measures (Letter Names, Letter Sounds, Phoneme Segmenting, Word Reading Fluency) demonstrate moderate to strong correlations with external benchmarks such as DIBELS, TOWRE, and SAT-10. Regression models frequently account for substantial variance in standardized outcomes, with combined predictor models explaining between 35% and over 70% of variance depending on grade and season. **Predictive validity** strengthens when criterion measures closely align with the easyCBM® construct (e.g., WRF predicting SAT-10 Word Reading). In Grades 3–8, Passage Reading Fluency (PRF), Multiple-Choice Reading Comprehension (MCRC), and Vocabulary show moderate predictive relationships with state assessments (e.g., OAKS, MSP, SBAS), with correlations often ranging from .57 to .68 for PRF and strong concurrent relations (.88–.95) with DIBELS ORF.

Evidence based on **internal structure** further supports construct validity. Confirmatory factor analyses generally support multi-factor structures distinguishing early literacy, fluency, and comprehension in lower grades, and coherent reading constructs in upper grades. Factor loadings for fluency and foundational skills are consistently strong, while comprehension indicators demonstrate moderate associations, occasionally influenced by ceiling effects. Rasch modeling across reports confirms item functioning, alternate-form equivalence, and appropriate coverage of the ability continuum.

Classification accuracy analyses provide consequential validity evidence related to screening decisions. ROC analyses demonstrate acceptable to strong AUC values (often .75–.95), with sensitivity prioritized to minimize false negatives in RTI contexts. Cross-validation studies in Oregon and Washington show stable cut scores across split samples, reinforcing defensibility of benchmark decisions within state-specific contexts.

Longitudinal and growth evidence supports the interpretation of reading fluency as a robust general indicator of reading competence, particularly in elementary grades. Growth modeling demonstrates sensitivity to instructional effects, and updated ORF norms confirm stability and modest national gains over time.

The following conclusions can be made from the validity studies with easyCBM® measures.

1. Criterion validity is present with other external measures, when using correlational analyses among single measures. Furthermore, the correlations among the various measures themselves are quite high.
2. When using regression analyses, the various measures work well together in predicting performance on other, external measures, far better than when used as single measures.
3. Although the National Center on Intensive Intervention (NCII) requires these relations to be documented for individual measures, this use is contrary to effective practices, where multiple measures are typically recommended.
4. The classification analyses indicate that most measures have sufficient area under the curve (AUC) and sensitivity. However, specificity is lower, indicating overidentification of students needing service. Three caveats need to be considered: (a) the analyses were conducted with the 40th Percentile Rank (PR) as the cut-off, which may be too high; (b) the cost of making false-positive and false-negative decisions need to be considered; (c) identification of risk on a singular measure reflects poor practice.
5. In another section of this technical report, multiple screening measures are used to identify students at risk of problems in learning to read; this mosaic of performance can then also be to identify effective progress monitoring measures.

Collectively, the integrated findings across technical reports align with the Standards' unified view of validity. Evidence from criterion relations, internal structure, classification accuracy, and growth modeling converges to support the interpretation of easyCBM® Reading scores for screening, progress monitoring, and instructional decision-making within MTSS systems. While no single statistic is sufficient, the cumulative evidence forms a coherent and technically defensible validity argument across Grades K–8.

Note 1: Two important caveats need to be considered in this and other summary reports on easyCM®: (a) Many reports reflect studies with multiple indicators of technical adequacy with both reliability and criterion validity investigated. In a similar manner, both criterion validity and classification accuracy may be reported in the same TR. (b) The reports typically include students from multiple grades with overlap across the TRs. We have attempted, however, to address both caveats by reporting on the main type of technical adequacy so that the suite of these reports are grouped as documents focusing on test development, alignment, reliability, and validity. Furthermore, the general sequence is to report results from earlier to later grades.

Note 2: All tables and figures in this summary are examples of those presented in full within the individual Technical Reports (TR) but are not exhaustive, just illustrative.

Summary of Technical Report 1003: Technical Adequacy of the easyCBM® Primary-Level Reading Measures (Grades K–1), 2009–2010 Version (Lai et al., 2010).

Methods

This technical adequacy report documents validity evidence for easyCBM® primary-level reading measures used in Kindergarten and Grade 1. Measures include foundational early literacy tasks (e.g., Letter Names, Letter Sounds, Phoneme Segmenting), and early reading fluency indicators (e.g., Word Reading Fluency; and grade-appropriate fluency measures). Validity evidence includes relations between easyCBM® measures and external achievement outcomes, notably SAT-10/SESAT Word Reading. Regression models are used to evaluate predictive validity (both single-predictor and multi-predictor), reporting R and R^2 to quantify the strength of association with the standardized criterion.

Results (quotes from pages 11-14 of the Technical Report)

Concurrent validity evidence. The spring model for Grade K was significant, $F(3,179) = 158.80$, $p < .05$, and accounted for 73% of the variance in SAT-10. The spring Word Reading measure had the highest coefficient, $b = 1.98$, $t(189) = 11.49$, $p < .05$, and uniquely explained 20.16% of the variance in the regression model. The descriptive statistics and regression model summaries for Grade K are presented in Tables 5-8. The spring model for Grade 1 was significant, $F(1,178) = 234.35$, $p < .05$, accounting for 57.8% of the variance in SAT-10. The spring Passage Reading Fluency benchmark had a coefficient, $b = .71$, $t(185) = 15.31$, $p < .05$, and uniquely explained 56.78% of the variance. The descriptive statistics and regression model summaries for Grade 1 are presented in Tables 9-12.

Predictive validity evidence. The fall model for Grade K was significant, $F(3,158) = 69.72$, $p < .05$, and accounted for 57% of the variance in SAT-10. The fall LS benchmark had the highest coefficient, $b = 2.48$, $t(2013) = 6.09$, $p < .05$, and uniquely explained 10.11% of the variance. The winter model for Grade K was significant, $F(3,169) = 112.11$, $p < .05$, and accounted for 66.6% of the variance in SAT-10. The winter WRF benchmark had the highest coefficient, $b = 2.48$, $t(1972) = 8.27$, $p < .05$, and uniquely explained 13.54% of the variance.

The fall and winter model for Grade K was significant, $F(6,155) = 86.18$, $p < .05$, and accounted for 76.9% of the variance in SAT-10. The winter WRF benchmark had the highest coefficient, $b = 4.02$, $t(1972) = 8.09$, $p < .05$, and uniquely explained 9.73% of the variance. Descriptive statistics and regression model summaries for Grade K are presented in Tables 13-22. The fall model for Grade 1 was significant, $F(3,155) = 48.18$, $p < .05$, accounting for 48.3% of the variance in SAT-10. The fall WRF benchmark had the highest coefficient, $b = .82$, $t(2179) = 2.54$, $p < .05$, and uniquely explained 2.16% of the variance. The winter model for Grade 1 was significant, $F(3,173) = 73.64$, $p < .05$, and accounted for 56.1% of the variance in SAT-10. The winter LS benchmark had the highest coefficient, $b = .92$, $t(2195) = 5.16$, $p < .05$, and uniquely explained 6.76% of the variance. The model in which we included both fall and winter results for Grade 1 was significant, $F(6,152) = 33.87$, $p < .05$, and accounted for 57.2% of the variance in SAT-10. The winter LS benchmark had the highest coefficient, $b = 1.03$, $t(2195) = 4.47$, $p < .05$, and uniquely explained 5.62% of the variance. Descriptive statistics and regression model summaries for Grade 1 are presented in Tables 23-32.

Table 2. Example Regression Predictors in Reading

Table 6

Kindergarten Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.853[a]	.727	.722	26.704	.727	158.798	3	179	.000

a. Predictors: (Constant), Spr10WRF, Spr10Seg, Spr10LS

Explanation of Model Summaries for Regression Analyses (see Table 3 above)

The regression "Model Summary" shows how well the predictor in the footnote predicts the dependent variable used in the analysis (the outcome name is not displayed in the snippet). The row labeled "1" is Model 1, which includes three predictor (easyCBM® = Spr10WRF, Spr10Seg and Spr10LS) and a constant (intercept).

R = .853. R is the multiple correlation between observed outcome scores and the scores predicted by the regression equation. With only one predictor, it is essentially the absolute value of the Pearson correlation between Spr10PRF and the outcome. An R of .853 indicates a strong positive linear relationship: students with higher scores on the measures (Spr10WRF, Spr10Seg and Spr10LS) tend to have higher predicted outcome scores.

R Square = .727. R^2 (the coefficient of determination) is the proportion of variance in the outcome explained by the model. Here, the three measures (Spr10WRF, Spr10Seg and Spr10LS) accounts for 72.7% of the variability in the Kindergarten outcome—substantial explanatory power for a single predictor.

Adjusted R Square = .722. Adjusted R^2 corrects R^2 for sample size and number of predictors to reduce optimism from fitting a model to a particular dataset. The adjustment is a function of the number of predictors and the sample size. The adjustment is small (.727 → .722). This suggests the explained variance is not driven by overfitting and would likely be similar in comparable samples.

Std. Error of the Estimate = 26.704. This is the standard deviation of the residuals (prediction errors) in the outcome's units. It represents the typical distance between a student's observed outcome score and the model's predicted score. Interpreting its magnitude depends on the outcome scale, but as a rule of thumb, smaller values indicate more precise prediction.

R Square Change = .727. The "Change Statistics" section is produced for hierarchical regression. R^2 Change is the increment in explained variance when the predictor block is entered. Because the measures are the first and only block, the change equals the total R^2.

F Change = 158.798; df1 = 3; df2 = 179. This F test evaluates whether adding the three measures improve model fit beyond an intercept-only model (i.e., whether the R^2 increase is greater than 0). df1 equals the number of predictors entered at this step (1). df2 is the residual degrees of freedom (N − predictors − 1). With df2 = 179, the implied sample size is about N = 180. The very large F value indicates the improvement in fit is enormous relative to residual error.

Sig. F Change = .000. This is the p-value for the F test (SPSS prints .000 when $p < .001$). It indicates the R^2 increase from adding the three measures is statistically significant. In practical terms, the three measures are highly reliable predictor of the Kindergarten outcome in this sample, and the combination of high R and high R^2 indicates both statistical and practical importance.

Overall, the three measures (Spr10WRF, Spr10Seg and Spr10LS) capture many differences on the outcome, but prediction is not perfect: unexplained variance remains ($1 - R^2 = .273$), reflecting other skills, instructional factors, and error.

Table 3. Example Variance Explained Output with easyCBM® Early Reading Measures

Table 10

Grade 1 Model Summary

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.754[a]	.568	.566	27.202	.568	234.347	1	178	.000

a. Predictors: (Constant), Spr10PRF

Table 14

Kindergarten Model Summary (Fall easyCBM and SAT-10)

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.755[a]	.570	.561	33.557	.570	69.715	3	158	.000

a. Predictors: (Constant), Fall09WRF, Fall09Seg, Fall09LS

Table 17

Kindergarten Model Summary (Winter easyCBM and SAT-10)

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.816[a]	.666	.660	29.564	.666	112.114	3	169	.000

a. Predictors: (Constant), Wint10WRF, Wint10Seg, Wint10LS

Table 20

Kindergarten Model Summary (Fall + Winter easyCBM and SAT-10)

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.877[a]	.769	.760	24.802	.769	86.178	6	155	.000

a. Predictors: (Constant), Wint10WRF, Wint10Seg, Fall09Seg, Wint10LS, Fall09LS, Fall09WRF

Table 24

Grade 1 Model Summary (Fall easyCBM and SAT-10)

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.695[a]	.483	.473	29.985	.483	48.184	3	155	.000

a. Predictors: (Constant), Fall09PRF, Fall09LS, Fall09WRF

Table 27

Grade 1 Model Summary (Winter easyCBM and SAT-10)

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.749[a]	.561	.553	27.597	.561	73.641	3	173	.000

a. Predictors: (Constant), Wint10PRF, Wint10LS, Wint10WRF

Table 30

Grade 1 Model Summary (Fall + Winter easyCBM and SAT-10)

Model					Change Statistics				
	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.756[a]	.572	.555	27.536	.572	33.865	6	152	.000

a. Predictors: (Constant), Wint10PRF, Wint10LS, Fall09LS, Fall09PRF, Wint10WRF, Fall09WRF

Construct validity evidence. For grade K, the model fit for the hypothesized 1-factor model for all seasons was good, with CFI indices ranging from .997 to .999, TLI indices ranging from .992 to .997 and RMSEA indices ranging from .028 to .047. Descriptive statistics and full CFA model fit indices are presented in Tables 33-34. For grade 1, CFI and TLI indices suggested good model fit for the hypothesized 1-factor model for all seasons with CFI indices ranging from .978 to .988 and TLI indices ranging from .979 to .982. RMSEA, on the other hand, while indicating a fair model fit for the fall model suggested poor model fit for winter and spring models, with RMSEA indices ranging from .106 to .136. Descriptive statistics and full CFA model fit indices are presented in Tables 35-36.

Figure 1. Example Construct Validity Structure

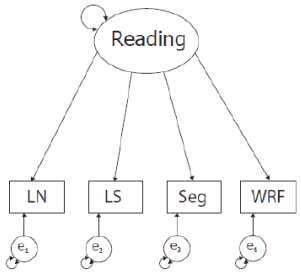


Figure 1. Hypothesized 1-Factor Model for easyCBM Reading Kindergarten

Note. LN = Letter Naming; LS = Letter Sounds; Seg = Phoneme Segmenting; WRF = Word Reading Fluency

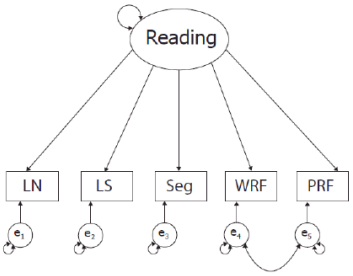


Figure 2. Hypothesized 1-Factor Model for easyCBM Reading Grade 1

Note. LN = Letter Naming; LS = Letter Sounds; Seg = Phoneme Segmenting; WRF = Word Reading Fluency; PRF = Passage Reading Fluency

Reference

Lai, C.-F., Nese, J. F. T., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). *Technical A\adequacy of the easyCBM® primary-level reading measures (Grades K–1), 2009–2010 Version (Technical Report #1003).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1004: Technical Adequacy of the easyCBM® Grade 2 Reading Measures (Jamgochian et al., 2010).

Methods

This report summarizes technical adequacy for Grade 2 easyCBM® reading measures, including criterion-related validity evidence. Measures include grade-appropriate reading indicators used for screening and progress monitoring. Criterion-related evidence is examined by relating easyCBM® scores to SAT-10 Word Reading outcomes using regression model summaries (R and R^2), evaluating how well earlier-season easyCBM® measures predict spring standardized reading performance.

Results

Concurrent validity. None of the grade 2 easyCBM® measures were significantly correlated with the SAT-10. Correlations ranged from .036 - .052 (Table 8). Regression analyses were conducted for each of the measures separately and combined (easyCBM® model). Individually, the measures were weak predictors of SAT-10 scores. In the combined model, MCRC was the best predictor.

Predictive validity. Fall WRF and PRF were significantly correlated with SAT-10 (.154 [$p < .05$] and .194 [$p < .01$], respectively). Winter WRF and PRF were also significantly correlated with SAT-10 (.151 [$p < .05$] and .221 [$p < .01$], respectively). Neither Fall nor Winter MCRC was significantly correlated with SAT-10. Regression analyses were conducted for each of the measures separately and combined by season. For Fall, the individual measures were weak predictors of SAT-10. In the combined model, WRF and MCRC were better predictors than PRF. For Winter, individually, MCRC was the best predictor of SAT-10. In the combined model, MCRC was also the best predictor, WRF was moderate, and PRF was negative. Descriptive statistics are presented in Table 7 and regression model summaries are presented in Tables 17 –24 (Fall) and 25 - 32 (Winter).

Table 4. Correlations Among easyCBMs® in Reading and Criterion Measures

Table 8

Grade 2 Correlation Matrix (easyCBM® N = 3675; SAT-10 N = 205)

		Fall 2009			Winter 2010			Spring 2010			
		WRF	PRF	MCRC	WRF	PRF	MCRC	WRF	PRF	MCRC	SAT-10
Fall 2009											
WRF	Pearson Correlation		.999**	.936**	.913**	.912**	.907**	.801**	.836**	.880**	.154*
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.028
PRF	Pearson Correlation			.934**	.912**	.912**	.906**	.799**	.835**	.877**	.194**
	Sig. (2-tailed)			.000	.000	.000	.000	.000	.000	.000	.005
MCRC	Pearson Correlation				.916**	.914**	.917**	.803**	.838**	.866**	.070
	Sig. (2-tailed)				.000	.000	.000	.000	.000	.000	.318
Winter 2010											
WRF	Pearson Correlation					.998**	.971**	.848**	.891**	.905**	.151*
	Sig. (2-tailed)					.000	.000	.000	.000	.000	.031
PRF	Pearson Correlation						.971**	.848**	.892**	.904**	.221**
	Sig. (2-tailed)						.000	.000	.000	.000	.001
MCRC	Pearson Correlation							.834**	.872**	.901**	-.081
	Sig. (2-tailed)							.000	.000	.000	.251
Spring 2010											
WRF	Pearson Correlation								.885**	.889**	.047
	Sig. (2-tailed)								.000	.000	.502
PRF	Pearson Correlation									.925**	.052
	Sig. (2-tailed)									.000	.458
MCRC	Pearson Correlation										.036
	Sig. (2-tailed)										.608

**Correlation is significant at the 0.01 level. *Correlation is significant at the 0.05 level.

Reference

Jamgochian, E., Park, B. J., Nese, J. F. T., Lai, C.-F., Sáez, L., Anderson, D., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM® grade 2 reading measures (Technical Report #1004*). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1005: Technical Adequacy of the easyCBM® Reading Measures (Grades 3–7), 2009–2010 Version (Sáez et al., 2010).

Methods

This report examines the technical adequacy of the easyCBM® reading assessments across grades 3–8. Measures included Passage Reading Fluency (PRF), Multiple-Choice Reading Comprehension (MCRC), and Vocabulary. Data were drawn from large-scale district administrations during the 2009–2010 school year. Reliability analyses included internal consistency for MCRC and Vocabulary, alternate form reliability, and growth slope reliability estimated using hierarchical linear modeling. Validity analyses included concurrent and predictive correlations with state reading assessments, regression modeling by season, and confirmatory factor analyses evaluating the structural coherence of constructs across grades. Descriptive statistics and quartile-based analyses allowed examination of differential performance patterns. Growth reliability was analyzed conditional on baseline performance.

Results

Internal consistency coefficients for MCRC and Vocabulary were generally moderate to strong. Growth slope reliability varied across grades and quartiles, with higher reliability in lower-performing groups. PRF demonstrated consistent predictive validity with state outcomes, particularly in fall and winter administrations. **Criterion validity** using regression analyses indicated that seasonal models explained meaningful variance in state test performance, with spring models often showing the strongest concurrent validity. Confirmatory factor analyses supported a multidimensional structure separating fluency, comprehension, and vocabulary constructs.
Overall, the measures demonstrated acceptable reliability and **strong criterion-related validity across grades**, supporting their use for RTI-based screening and monitoring.

Table 5. Illustrative Table from Technical Report 1005

Table 386
Grade 6 Fall Easy CBM Scores Predicting Spring OAKS Reading Performance

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.764[a]	.584	.582	6.098	.584	434.920	3	931	.000

a. Predictors: (Constant), Fall09Voc, Fall09MCRC, Fall09PRF

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	199.301	.884		225.565	.000	197.567	201.035					
	Fall09PRF	.074	.006	.315	11.475	.000	.061	.086	.645	.352	.243	.594	1.684
	Fall09MCRC	.401	.069	.154	5.820	.000	.266	.536	.551	.187	.123	.638	1.568
	Fall09Voc	.886	.057	.427	15.473	.000	.773	.998	.693	.452	.327	.589	1.699

a. Dependent Variable: OAKSRdgTot

Reference

Sáez, L., Park, B. J., Nese, J. F. T., Jamgochian, E., Lai, C. F., Anderson, D., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM reading measures (Grades 3–8), 2009–2010 version (Technical Report No. 1005)*. University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1309: easyCBM® Reading Criterion-Related Validity Evidence: Grades K–1 (Lai et al., 2013b).

TR1309 reports a criterion-related validity study examining how easyCBM® early reading measures relate to other well-established literacy assessments for Kindergarten and Grade 1. Participants were recruited from a convenience sample of three elementary schools in one Oregon district that used easyCBM® as part of an RTI model. The study included 214 Kindergarten students and 227 Grade 1 students. Data were collected over three weeks in February 2013. Fifteen University of Oregon undergraduate/graduate data collectors were trained by a licensed school psychologist in standardized administration and scoring; fidelity checks included supervision, double-check scoring, and random verification.

Methods

easyCBM® measures included Phoneme Segmenting (PS), Letter Names (LN), Letter Sounds (LS), and Word Reading Fluency (WRF), each administered individually for 60 seconds using standardized protocols. Comparator measures were selected to target similar constructs and included DIBELS (Kindergarten: Initial Sound Fluency, Phoneme Segmenting Fluency, Letter Naming Fluency; Grade 1: Phoneme Segmenting Fluency, Letter Naming Fluency, Nonsense Word Fluency), the CTOPP Elision subtest (phonological awareness), and the TOWRE-2 Sight Word Efficiency subtest (word reading). To reduce order and form effects, easyCBM® and comparator measures were assembled into a booklet and administered in one of two counterbalanced rotations. Data preparation coded reasons for missingness, and correlation analyses were conducted. Because linearity/normality assumptions were often not met, the primary statistic was Spearman's rho.

Results

Overall, easyCBM® measures showed a wide range of relations with comparator measures, with the strongest evidence when content, format, and administration procedures were closely aligned. In Kindergarten, descriptive statistics indicated broad score ranges on each measure; only 188 students completed the TOWRE due to scheduling constraints. Correlations showed a high relation between easyCBM® Phoneme Segmenting and DIBELS Phoneme Segmenting Fluency (rs = .85), and a high relation between easyCBM® Letter Names and DIBELS Letter Naming Fluency (rs = .86). easyCBM® Letter Sounds had a moderate relation with DIBELS Initial Sound Fluency (rs = .55). easyCBM® Word Reading Fluency was moderately high with TOWRE Sight Word Efficiency (rs = .79). In contrast, easyCBM® Phoneme Segmenting showed a low relation with CTOPP Elision (rs = .39).

In Grade 1, all measures were administered to 214 students except TOWRE (n = 210 in the table). easyCBM® Phoneme Segmenting was moderately to highly related to DIBELS Phoneme Segmenting Fluency (rs = .75), and easyCBM® Letter Names was highly related to DIBELS Letter Naming Fluency (rs = .80). easyCBM® Letter Sounds showed a moderate relation with DIBELS Nonsense Word Fluency (rs = .58). easyCBM® Word Reading Fluency showed a very high relation with TOWRE Sight Word Efficiency (rs = .95). However, the association between easyCBM® Phoneme Segmenting and CTOPP Elision was essentially zero and non-significant in Grade 1 (rs = .05).

The discussion emphasizes that the strongest correlations likely reflect close correspondence in content and procedures (e.g., easyCBM® PS with DIBELS PSF; easyCBM® LN with DIBELS LNF). Moderate relations for Letter Sounds are interpreted as consistent with related but not identical tasks (e.g., picture-based initial sound identification or decoding unfamiliar letter strings). The weak relations with CTOPP Elision were unexpected and are attributed to using only a single CTOPP subtest rather than the recommended phonological awareness composite; the authors recommend including additional CTOPP subtests in future studies. Overall, the report

concludes that easyCBM® PS and LN show strong criterion evidence with DIBELS, WRF shows strong criterion evidence with TOWRE, and LS shows moderate criterion evidence with DIBELS measures.

Summary. In the screen shot below, for Kindergarten and Grade 1:

1. Phoneme Segmenting and Letter Names from easyCBM® correlate highly with their corresponding measures from Dibels.
2. Word Reading from easyCBM® correlates highly with a similar measure from TOWRE.

Table 6. Correlations Among easyCBMs® Reading with DIBELS, CTOPP, and TOWRE

Table 9

Spearman's Rho Rank Correlation Results – Kindergarten

easyCBM® Measures		Comparator Measures				
		DIBELS Phoneme Segmenting Fluency	DIBELS Letter Naming Fluency	DIBELS Initial Sound Fluency	CTOPP Elision	TOWRE Sight Word Efficiency
Phoneme Segmenting	r_s	.85**	-	-	.39**	-
	n	227	-	-	222	-
Letter Names	r_s	-	.86**	-	-	-
	n	-	227	-	-	-
Letter Sounds	r_s	-	-	.55**	-	-
	n	-	-	225	-	-
Word Reading Fluency	r_s	-	-	-	-	.79**
	n	-	-	-	-	188

Note. ISF = Initial Sound Fluency, PSF = Phoneme Segmenting Fluency, LNF = Letter Naming Fluency, PS = phoneme segmenting, LN = Letter names, LS = Letter sounds, WRF = Word reading fluency; **. Correlation is significant at the 0.01 level (2-tailed).

Table 10

Spearman's Rho Rank Correlation Results – Grade 1

easyCBM® Measures		Comparator Measures				
		DIBELS Phoneme Segmenting Fluency	DIBELS Letter Naming Fluency	DIBELS Nonsense Word Fluency	CTOPP Elision	TOWRE Sight Word Efficiency
Phoneme Segmenting	r_s	.75**	-	-	0.05	-
	n	214	-	-	214	-
Letter Names	r_s	-	.80**	-	-	-
	n	-	214	-	-	-
Letter Sounds	r_s	-	-	.58**	-	-
	n	-	-	214	-	-
Word Reading Fluency	r_s	-	-	-	-	.95**
	n	-	-	-	-	210

**. Correlation is significant at the 0.01 level (2-tailed).

Note. PSF = Phoneme Segmenting Fluency, LNF = Letter Naming Fluency, NWF = Nonsense Word Fluency, PS = phoneme segmenting, LN = Letter names, LS = Letter sounds, WRF = Word reading fluency; **. Correlation is significant at the 0.01 level (2-tailed).

Reference

Lai, C.-F., Alonzo, J., & Tindal, G. (2013). *easyCBM® reading criterion-related validity evidence: Grades K–1 (Technical Report #1309).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0922: Technical Adequacy of the easyCBM® Reading Measures (Grades 3–8), 2009–2010 Version (Tindal et al., 2009b)

This technical report investigates how student characteristics influence growth in passage reading fluency (PRF) across grades 3 through 8 using data from the easyCBM® assessment system. The primary goal of the study was to analyze patterns of reading fluency development over time and determine how demographic and educational factors relate to both initial reading performance and growth rates.

Methods

The study analyzed districtwide data collected during the 2009 school year. Students in grades 3 through 8 were assessed on passage reading fluency during three benchmark periods: fall, winter, and spring. Passage reading fluency was measured as the number of words read correctly per minute during timed oral reading passages administered through the easyCBM® system. Sample sizes were large at each grade level, typically including approximately one thousand students per grade. Student characteristics examined in the analysis included gender, race or ethnicity, economic disadvantage, special education status, Title I participation, and identification as historically low achieving.

To analyze reading growth across the school year, the researchers used hierarchical linear modeling (HLM). This statistical approach allowed the analysis to account for the nested structure of the data, where multiple reading assessments were collected for each student across time. In the model, time served as the first level of analysis, representing the repeated fall, winter, and spring assessments, while student characteristics were included at the second level. The modeling approach made it possible to examine both the intercept, representing initial reading fluency levels, and the slope, representing the rate of reading growth over time. By including student characteristics in the model, the analysis estimated how these variables influenced both starting performance and growth patterns.

Results

Results indicated that student characteristics significantly influenced both initial reading fluency scores and the rate of fluency growth across the school year. Several demographic variables were associated with lower starting levels of reading fluency, including being male, belonging to a racial or ethnic minority group, being economically disadvantaged, and receiving either special education or Title I services. These characteristics also influenced growth rates in some grade levels, though the strength and direction of the relationships varied across grades.

The hierarchical models showed that the relationship between initial reading performance and subsequent growth differed by grade level. In some grades, students who began with higher fluency scores showed slower growth rates, while in other grades the pattern differed. Overall, the results demonstrate that reading fluency growth is influenced by multiple student characteristics and that these relationships vary across grade levels.

Table 7. Illustrative Table from Technical Report 0922

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	185.379007	1.658585	111.769	1212	0.000
GENDER_N, G01	9.875223	1.961743	5.034	1212	0.000
ETHNICIT, G02	-9.950520	2.321850	-4.286	1212	0.000
ECONDIS, G03	-12.662480	2.193945	-5.772	1212	0.000
SPECED, G04	-36.176189	3.304185	-10.949	1212	0.000
For ORDER slope, B1					
INTRCPT2, G10	-0.726933	0.115489	-6.294	1212	0.000
GENDER_N, G11	-0.275166	0.144826	-1.900	1212	0.057
ETHNICIT, G12	0.280367	0.187889	1.492	1212	0.136
ECONDIS, G13	0.064711	0.160073	0.404	1212	0.686
SPECED, G14	-0.300663	0.248768	-1.209	1212	0.227

Reference

Tindal, G., Nese, J. F., & Alonzo, J. (2009b). *Hierarchical linear modeling of passage reading fluency growth as a function of student characteristics (Technical Report # 0922)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1801: Classification Accuracy of the easyCBM® Kindergarten – Grade 2 Reading Measures (Alonzo & Anderson, 2018a).

Methods

TR1801 reports classification-accuracy analyses designed to identify screening cut scores for easyCBM® literacy measures in Kindergarten through Grade 2. Data came from one small Pacific Northwest school district and included all enrolled students who were present during benchmark windows from fall 2016 through spring 2017 (Kindergarten n=288; Grade 1 n=315; Grade 2 n=326). Assessments were administered by district personnel (teachers or instructional assistants) using district benchmark protocols. Prior to testing, staff completed easyCBM's online training modules and demonstrated proficiency for each measure they administered.

Measures varied by grade and season. Kindergarten included Phoneme Segmenting (SEG), Letter Names (LN; fall only), Letter Sounds (LS), and Word Reading Fluency (WRF; winter/spring). Grade 1 included SEG (fall), LS, WRF, and Passage Reading Fluency (PRF; winter/spring). Grade 2 included PRF plus group-administered Multiple-Choice Reading Comprehension (MCRC) and Vocabulary (VOC) in fall, winter, and spring. Classification accuracy was evaluated with ROC analyses reporting Area Under the Curve (AUC) and candidate thresholds. Threshold selection followed a decision rule: keep sensitivity ≥0.80 and maximize specificity; if specificity exceeded 0.90, then keep sensitivity ≥0.90 and maximize specificity; ties were resolved by selecting the cut maximizing sensitivity. Outcomes were defined using spring 'benchmark' criteria anchored to national norms: Kindergarten outcome = spring LS 50th percentile raw score (32); Grade 1 outcome = spring WRF 50th percentile raw score (49); Grade 2 outcome = spring PRF 50th percentile raw score (101).

Results

Table 8 summarizes ROC results (AUC, selected threshold, sensitivity, specificity) by grade, measure, and season. Except for Kindergarten fall SEG, all measures showed acceptable-to-strong diagnostic accuracy (AUCs generally in the mid-.70s to high-.90s), and the selected thresholds met the report's sensitivity-first decision rules.

Table 8. Example Classification Accuracy Output for Early Reading Measures

Measure	Grade	Season	AUC	Threshold	Sensitivity	Specificity
LN	K	fall	0.76	6.5	0.82	0.66
LS	K	fall	0.68	1.5	0.8	0.49
SEG*	K	fall	0.37	---	1	0
LS	K	win	0.87	18.5	0.82	0.81
SEG	K	win	0.75	40.5	0.81	0.45
SEG	K	spr	0.85	49.5	0.80	0.77
SEG	1	fall	0.73	46.5	0.81	0.36
LS	1	fall	0.84	33.5	0.81	0.75
WRF	1	fall	0.95	11.5	0.84	0.89
LS	1	win	0.81	45.5	0.81	0.63
WRF	1	win	0.98	22.5	0.9	0.93
MCRC	2	fall	0.76	7.5	0.83	0.57
VOC	2	fall	0.82	10.5	0.85	0.64
PRF	2	fall	0.92	54.5	0.81	0.85
MCRC	2	spr	0.81	10.5	0.84	0.61
VOC	2	win	0.82	11.5	0.82	0.59
MCRC	2	win	0.81	9.5	0.81	0.63
PRF	2	win	0.94	84.5	0.9	0.74
VOC	2	spr	0.85	11.5	0.85	0.64

Overall, the report concludes that easyCBM® K–2 literacy measures support strong screening decisions when cut scores are selected to prioritize sensitivity while maintaining adequate specificity, with the notable exception of fall Kindergarten SEG.

Reference

Alonzo, J., & Anderson, D. (2018). *Classification accuracy of the easyCBM® Kindergarten – Grade 2 Reading Measures (Summary of Technical Report # 1801).* Eugene. OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1805: Supplementary Report on easyCBM® Grade K–2 Literacy Measures: Classification Accuracy and Predictive Validity (Follow-Up to TR 1403) (Alonzo & Anderson, 2018a).

Methods

This supplementary report reanalyzed a year-long K–1 dataset to provide additional indices and confidence intervals for classification accuracy and predictive validity. The study used fall easyCBM® alternate forms of Letter Names, Letter Sounds, Phoneme Segmenting, and Word Reading Fluency. External criteria were spring Stanford Achievement Test (SESAT/SAT-10) subtests appropriate to grade (e.g., kindergarten Sounds and Letters; Kindergarten Word Reading; Grade 1 word-related outcomes). Analyses relied on ROC methodology and area under the curve (AUC) to evaluate how well fall easyCBM® scores classified students relative to criterion benchmark cut points.

Results

For Kindergarten, AUC values varied across measures and alternate forms, but many fell in ranges commonly interpreted as acceptable to good for screening. **Letter Names** forms showed AUCs spanning roughly .69 to .85 (average ~.77) for predicting SAT-10 Sounds and Letters performance. **Letter Sounds** forms showed a narrower band (about .71 to .78; average ~.76). **Phoneme Segmenting** generally produced lower AUCs (roughly .62 to .71; average ~.66), consistent with a more specific construct and/or greater distance from some criteria. **Word Reading**

Fluency forms in kindergarten also showed acceptable AUCs (about .71 to .78; average ~.76). Interpreting these values: AUC near .70 is often treated as minimally acceptable discrimination for universal screening; around .80 is good; values closer to .60 indicate limited discrimination for that pairing. The report also highlights variability across forms and criterion pairings, supporting the practical recommendation to use multiple indicators (and professional judgment) when decisions have higher stakes than routine screening.

See Table 9 on the next page for all coefficients displayed by Grade (Kindergarten, Grade 1, and Grade 2).

easyCBM® Reading Measures and SAT-10 Form: Sounds and Letters

- Area Under the Curve (AUC)
- Sensitivity
- Specificity

Kindergarten Measures:

- Letter Names
- Letter Sounds
- Phoneme Segmenting
- Word Reading

Grade 1 Measures

- Phoneme Segmenting
- Word Reading

Grade 2 Measures

- Passage Reading

Note that the **SAT 10** criterion measure changes with the easyCBM® grade level measures and includes

- Sounds & Letters
- Word Reading
- Vocabulary

Table 9. AUC with 40th PR on Test Forms of SAT-10

1.Kindergarten Letter Names –> SAT-10 Sounds & Letters										
SAT-10 Form: Sounds and Letters	8	9	10	11	12	13	14	15	16	17
Area Under the Curve (AUC)	0.77	0.69	0.85	0.77	0.69	0.85	0.77	0.69	0.85	0.77
Sensitivity	0.74	0.8	0.8	0.8	0.84	0.71	0.76	0.82	0.79	0.84
Specificity	0.63	0.56	0.56	0.59	0.61	0.57	0.5	0.55	0.57	0.61

2. Kindergarten Letter Sounds –> SAT-10 Sounds & Letters										
SAT-10 Form: Sounds and Letters	8	9	10	11	12	13	14	15	16	17
Area Under the Curve (AUC)	0.76	0.78	0.75	0.77	0.76	0.71	0.75	0.77	0.74	0.76
Sensitivity	0.76	0.84	0.82	0.78	0.8	0.69	0.82	0.84	0.74	0.86
Specificity	0.66	0.5	0.52	0.57	0.53	0.62	0.51	0.52	0.57	0.57
3. Kindergarten Phoneme Segmenting –> SAT-10 Sounds & Letters										
SAT-10 Form: Sounds and Letters	5	7	7	8	10	11	12	13	14	15
Area Under the Curve (AUC)	0.7	0.64	0.69	0.65	0.65	0.64	0.62	0.71	0.67	0.66
Sensitivity	0.67	0.76	0.83	0.74	0.77	0.65	0.77	0.88	0.81	0.74
Specificity	0.57	0.47	0.48	0.5	0.49	0.59	0.47	0.5	0.52	0.48

4. Kindergarten Phoneme Segmenting –> SAT-10 Word Reading									
SAT-10 Form: Word Reading	5	7	8	10	11	12	13	14	15
Area Under the Curve (AUC)	0.71	0.6	0.66	0.66	0.6	0.72	0.64	0.68	0.66
Sensitivity	0.85	0.83	0.74	0.81	0.9	0.92	0.96	0.8	0.77
Specificity	0.43	0.31	0.33	0.31	0.44	0.32	0.33	0.36	0.3

5. Kindergarten Word Reading Fluency –> SAT-10 Sounds & Letters										
SAT-10 Form: Sounds and Letters	8	9	10	11	12	13	14	15	16	17
Area Under the Curve (AUC)	0.76	0.78	0.75	0.77	0.76	0.71	0.75	0.77	0.74	0.76
Sensitivity	0.85	0.83	0.74	0.81	0.9	0.92	0.96	0.8	0.77	0.77
Specificity	0.43	0.31	0.33	0.31	0.44	0.32	0.33	0.36	0.3	0.58

6. Grade 1 Phoneme Segmenting –> SAT-10 Sounds & Letters and Word Reading

SAT-10 Form: Sounds Letters (SL) Word Reading (WR)	5 SL	5WR	12WR	12WR	15WR	15WR	16WR	16WR	17WR	17WR
Area Under the Curve (AUC)	0.57	0.59	0.6	0.47	0.57	0.54	0.63	0.53	0.7	0.63
Sensitivity	0.6	0.65	0.8	0.65	0.38	0.53	0.75	0.53	0.53	0.53
Specificity	0.33	0.56	0.67	0.51	0.46	0.49	0.5	0.49	0.49	0.49

7. Grade 1 Word Reading Fluency –> SAT-10 Sounds & Letters

SAT-10 Form: Sounds and Letters	10	16
Area Under the Curve (AUC)	0.6	0.77
Sensitivity	0.8	0.75
Specificity	0.67	0.5

8. Grade 2 Passage Reading Fluency –> SAT-10 Vocabulary

SAT-10 Form: Vocabulary	5	9	10	15	17
Area Under the Curve (AUC)	0.83	0.86	0.86	0.85	0.86
Sensitivity	0.92	0.96	0.96	0.94	0.97
Specificity	0.43	0.37	0.38	0.33	0.39

Interpreting these values: AUC near .70 is often treated as minimally acceptable discrimination for universal screening; around .80 is good; values closer to .60 indicate limited discrimination for that pairing. The report also highlights variability across forms and criterion pairings, supporting the practical recommendation to use multiple indicators (and professional judgment) when decisions have higher stakes than routine screening.

Reference

Alonzo, J., & Anderson, D. (2018). *Supplementary report on easyCBM® grade K–2 literacy measures: classification accuracy and predictive validity, a follow-up to Technical Report 1403 (Technical Report #1805).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1806: Supplementary Report on easyCBM® PRF Measures: A Follow-Up to Previous Technical Reports (Alonzo & Anderson, 2018c).

Methods

This supplementary technical report re-analyzed previously published easyCBM® Passage Reading Fluency (PRF) data in response to requests for additional statistical detail, particularly confidence intervals around classification, reliability, and validity estimates. The report is organized into three primary evidence categories: classification accuracy, reliability, and criterion-related validity.

Criterion validity was investigated in addition to **classification accuracy** analyses, using the Smarter Balanced English Language Arts (SBAS ELA) assessment served as the independent criterion measure. Students in Grades 3–8 from Pacific Northwest districts were included (school year 2014–2015). The 40th percentile of the easyCBM® National Norms was used as the benchmark cut score, reflecting updated accountability expectations under SBAS. Receiver Operating Characteristic (ROC) analyses were conducted in R to compute sensitivity, specificity, false positive and false negative rates, positive and negative predictive power, overall classification rate, Area Under the Curve (AUC), and associated 95% confidence intervals.

Reliability analyses included test-retest reliability (same form, one week apart), alternate-form reliability (different passages), and Generalizability Theory (G-Theory) studies. Samples were drawn from three public elementary schools (Grades 1–5). G-Theory analyses used a fully crossed two-facet design (persons × forms × occasions) to estimate variance components and compute G-coefficients and phi coefficients.

Criterion-related validity evidence was examined using two studies. Study 1 evaluated predictive validity via linear regression between easyCBM® PRF and SBAS ELA scores. Study 2 examined concurrent validity via bivariate correlations between easyCBM® PRF and DIBELS ORF.

Results

Predictive validity coefficients between PRF and SBAS ELA ranged from .57 to .68 across Grades 3–8, indicating moderate positive relationships. Concurrent validity correlations between PRF and DIBELS ORF were very strong (.88–.95), confirming that the two measures assess the same underlying oral reading fluency construct.
This supplementary analysis strengthens the technical argument for easyCBM® PRF by providing confidence intervals and expanded reliability evidence. The findings indicate:

- Strong alternate-form and test-retest reliability
- High generalizability across forms and occasions
- Moderate predictive validity with SBAS ELA
- Very strong concurrent validity with DIBELS ORF

Together, these results provide robust technical evidence supporting the use of easyCBM® PRF for screening and progress monitoring within accountability-aligned MTSS frameworks.

Table 10. Illustrative Table from Technical Report 1806

Table 7
Criterion-Related Validity Evidence

Type of Validity	Grade	Criterion	n	Coefficient	95% Confidence Interval*: Lower Bound	95% Confidence Interval*: Upper Bound
Predictive	3	SBAS English Language Arts	1303	0.67	0.63	0.71
Predictive	4	SBAS English Language Arts	1520	0.64	0.60	0.68
Predictive	5	SBAS English Language Arts	1539	0.68	0.64	0.71
Predictive	6	SBAS English Language Arts	1467	0.61	0.57	0.65
Predictive	7	SBAS English Language Arts	1415	0.62	0.58	0.66
Predictive	8	SBAS English Language Arts	1475	0.57	0.53	0.61
Predictive	3	SBAS English Language Arts	1280	0.67	0.63	0.71
Predictive	4	SBAS English Language Arts	1489	0.63	0.59	0.67
Predictive	5	SBAS English Language Arts	1575	0.68	0.64	0.71
Predictive	6	SBAS English Language Arts	1494	0.63	0.59	0.67

Reference

Alonzo, J., & Anderson, D. (2018). *Supplementary report on easyCBM® PRF measures: A follow-up to previous technical reports (Technical Report No. 1806).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1807: Supplementary Report on easyCBM® MCRC Measures: A Follow-up to Previous Technical Reports (Alonzo & Anderson, 2018b).

Methods

This technical report presents supplementary analyses of the easyCBM® Multiple Choice Reading Comprehension (MCRC) measures to provide additional evidence regarding classification accuracy, reliability, and criterion-related validity. The samples consisted of large-scale datasets of students in Grades 2–8 drawn primarily from school districts in the Pacific Northwest. Classification accuracy analyses used statewide Smarter Balanced English Language Arts (SBAS ELA) scores as the independent criterion measure. The analytic samples for classification accuracy included more than 26,000 to 34,000 students per grade for Grades 3–8 in the 2014–2015 school year. For reliability analyses, fall and winter benchmark data from the 2013–2014 school year were used, including over 20,000 students per grade in many cases. The easyCBM® MCRC assessments were administered in benchmark windows: fall, winter, and spring.

Classification analyses used a cut score corresponding to the 40th percentile from easyCBM® national norms to identify students at risk for reading difficulties. Statistical analyses were conducted in R and included calculation of sensitivity, specificity, positive and negative predictive power, overall classification accuracy, and area under the ROC curve (AUC) with confidence intervals. Reliability was examined using Cronbach's alpha and split-half reliability. Criterion-related validity was evaluated through bivariate correlations and linear regression analyses relating easyCBM® MCRC scores to SBAS ELA performance.

Findings

The classification accuracy analyses demonstrated that easyCBM® MCRC measures provide useful screening information for predicting student performance on the Smarter Balanced ELA assessment. In fall benchmark assessments, overall classification accuracy ranged from approximately .70 to .75 across Grades 3–8. Sensitivity values ranged from about .55 to .62, indicating that the assessment identified most students who would later

perform below proficiency on the criterion assessment. Specificity values were consistently high, ranging from approximately .83 to .91 across grades. Area Under the Curve (AUC) estimates ranged from approximately .79 to .84, indicating strong classification performance for a universal screening tool. Winter benchmark results produced similar findings, with classification accuracy between .63 and .78 and AUC values between approximately .75 and .84.

Spring benchmark analyses showed slightly improved predictive performance for several grades. AUC estimates ranged from approximately .80 to .85, while specificity values were often above .80 and reached as high as .93 in some grades. Overall classification rates remained stable around .70 to .75. These results suggest that the MCRC measures provide increasingly accurate information about student reading comprehension performance as the academic year progresses.

Internal consistency reliability estimates were generally strong across grades. Cronbach's alpha values ranged from approximately .66 to .79 across Grades 2–8. Split-half reliability coefficients ranged from approximately .60 to .76, indicating moderate to strong internal consistency for the timed comprehension assessments. Reliability analyses conducted across student subgroups—including general education students, students receiving special education services, and demographic subgroups—showed similarly strong reliability coefficients.

Criterion-related validity analyses showed moderate to strong correlations between easyCBM® MCRC scores and SBAS ELA scores. Predictive validity correlations for fall MCRC scores predicting spring SBAS scores were approximately .62 to .63 in Grade 3. Concurrent validity correlations between spring MCRC and SBAS scores across Grades 3–8 generally ranged from approximately .54 to .68. These findings indicate that the easyCBM® MCRC assessments capture important aspects of reading comprehension and meaningfully predict performance on an independent statewide assessment of English Language Arts.

Table 11. Main Findings Summary for Technical Report 1807

Evidence Type	Key Results	Interpretation
Classification Accuracy	AUC ≈ .79–.85; classification rate ≈ .70–.75	MCRC effectively identifies students at risk for reading difficulty
Reliability	Cronbach's α ≈ .66–.79; split-half ≈ .60–.76	Moderate to strong internal consistency across grades
Criterion Validity	Correlations with SBAS ELA ≈ .54–.68	MCRC scores strongly relate to external measures of reading achievement

Reference

Alonzo, J., & Anderson, D. (2018). *Supplementary report on easyCBM® MCRC measures: A follow-up to previous technical reports (Technical Report No. 1807)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1403: easyCBM® Kindergarten Beginning Reading Measures: Alternate Form Reliability and Criterion Validity With the SAT-10 (Wray et al., 2014).

Methods

This report evaluated alternate form reliability (addressed in another technical report) and criterion validity for Kindergarten easyCBM® beginning reading measures administered as brief, individually administered fluency assessments. Measures included Letter Names Fluency, Letter Sounds Fluency, Phoneme Segmenting Fluency, and Word Reading Fluency. For criterion validity, easyCBM® scores were compared to spring Stanford Achievement Test (SAT-10/SESAT) subtests aligned to early literacy outcomes (e.g., Sounds and Letters; Word Reading). Analyses

emphasized correlations and regression models predicting SAT-10 outcomes from fall easyCBM® measures, providing evidence of how well early easyCBM® scores relate to later standardized achievement.

Results

Criterion-related validity was examined using regression models relating easyCBM® measures to SESAT/SAT-10 outcomes. In kindergarten, easyCBM® LN, LS, and PS jointly explained about 35%–40% of the variance in SAT-10 Sounds & Letters across time points, and about 48%–58% of the variance in SAT-10 Word Reading. LN was frequently a significant predictor, LS was often significant for SAT-10 Word Reading, and PS contributed inconsistently. In Grade 1, easyCBM® LN, LS, PS, and WRF explained about 14%–32% of SAT-10 Word Study Skills, with WRF consistently the strongest predictor (unique variance commonly in the teens to low 20s). When the criterion was restricted to word reading (easyCBM® WRF later in the year), the model explained about 49%–56% of the variance. The report concludes that the battery provides meaningful evidence of alternate-form reliability and criterion validity, with stronger prediction when the outcome is closely aligned to word reading.

Quotes from Tech Report (page 7 from the Technical Report)

Kindergarten. Results for the criterion validity of the kindergarten easyCBM® measures can be found in tables 18 and 19. The first model regressed easyCBM® LN, LS, and PS from time 1 to time 4 on SAT-10 Sounds and Letters. The overall variance explained by the model across time ranged from 35%---40%. For all time points, LN scores were a significant predictor of Sounds and Letters performance. LS scores were significant predictors only at Time 3 (February) and PS was a significant predictor for all time points except Time 3. The unique variance in performance explained by LN ranged from .16%---.64%. LS and PS both explained .03% of the unique variance across time. The second model regressed the same independent variables on the SAT-10 Word Reading measure. The overall model explained 48%---58% of the variance of the dependent variable. LN was a significant predictor for all time points except Time 2 (December). LS scores were significant for all time points, while PS was a non-significant predictor across all the time points. The variance of SAT-10 Word Reading uniquely explained by LN ranged from .01%---.81% and the variance uniquely explained by LS ranged from 16% – 64%.

Grade 1. Results for the criterion validity study of the grade 1 easyCBM® measures can be found in tables 20 and 21. The first model regressed easyCBM® LN, LS, PS, and WRF from time 1 to time 4 upon SAT---10 WSS. The variance explained by the whole model at each time point varied from 14%---32%. WRF was a significant predictor of SAT---10 WSS for all time points, and PS was significant for Time 4 (April) only. LN and LS were not significant predictors of SAT---10 WSS at any time point. The variance uniquely explained by PS at Time 4 was 2.0%, and the unique variance explained by WRF ranged from 13.0%---21.2%. The second model used the same independent variables, but substituted easyCBM® WRF at Time 4 for SAT---10 WSS. The variance explained by the model across time ranged from 49%---56%. LN performance was a significant predictor of easyCBM® WRF performance at Time 4 for all time points, and LS was significant for Times 2 (December) and 3 (February). PS was not a significant predictor of WRF performance at Time 4 at any time point. The variance in easyCBM® WRF score at Time 4 uniquely explained by LN ranged from 14.4%---41.0%, and the variance in easyCBM® WRF performance at Time 4 uniquely explained by LS performance ranged from 1.4 % – 2.3%.

Table 12. Example Regression Values with easyCBM® in Early Reading Measures

Table 18

Grade Kindergarten Regression Models Predicting SAT-10 Sounds and Letters

	Time 1			Time 2			Time 3			Time 4		
easyCBM © measures	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part
Letter names	0.88*	0.34*	.04	0.98*	0.35*	.04	0.90*	0.33*	.04	1.13**	0.42**	.08
Letter sounds	0.61	0.16	.01	0.74	0.20	.01	0.86*	0.28*	.03	0.47	0.16	.01
Phoneme segmenting	0.54*	0.20*	.03	0.52*	0.19*	.03	0.33	0.11	.01	0.59*	0.17*	.03
R^2		.35			.39			.39			.40	
N		124			141			151			160	

Note. *p < .05, **p < .001, Part = Semipartial correlations.

Table Explanation. The tables report four concurrent regression models (Times 1–4) predicting SAT-10 Sounds and Letters from various easyCBM® measures. Standardized betas (β) show the relative strength of each predictor after all measures are in the model; larger |β| indicates a stronger unique association with the outcome, holding the other predictors constant. For example, in Table 18, across times, Letter Names is the most consistent predictor (β ≈ .33–.42) and is statistically significant at every time point, with its strongest effect at Time 4 (β = .42, p < .001). Letter Sounds is weaker early (β ≈ .16–.20, ns) but becomes a meaningful unique predictor at Time 3 (β = .28, p < .05). Phoneme Segmentation shows modest, significant effects at Times 1, 2, and 4 (β ≈ .17–.20), but not at Time 3.

The Part column gives the semi partial correlation for each predictor—i.e., the correlation between the outcome and the portion of that predictor not shared with the other predictors. Squaring Part yields the unique variance explained. For example in Table 18, Letter Names at Time 4 has Part = .08, so it uniquely explains about .006 (0.6%) of outcome variance beyond the other measures, despite a comparatively large β. Model R^2 values (.35–.40) indicate that the three easyCBM® measures together explain 35–40% of SAT-10 variance, with sample size increasing from N = 124 to 160.

Table 13. Example Regression Model Predictors between easyCBM® and SAT-10

Table 19

Grade Kindergarten Regression Models Predicting SAT-10 Word Reading

	Time 1			Time 2			Time 3			Time 4		
easyCBM © measures	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part
Letter names	1.62**	0.49**	.09	0.64	0.19	.01	1.18**	0.36**	.05	0.96*	0.30*	.04
Letter sounds	1.46*	0.30*	.04	2.33**	0.52**	.08	1.44**	0.39**	.06	1.45**	0.42**	.08
Phoneme segmenting	0.02	0.01	.00	0.14	0.05	.00	0.19	0.05	.00	0.26	0.06	.00
R^2		.58			.49			.53			.48	
N		123			138			150			156	

Note. *p < .05, **p < .001, Part = Semipartial correlations.

Table 20

Grade 1 Regression Models Predicting SAT-10 Word Study Skills

	Time 1			Time 2			Time 3			Time 4		
easyCBM © measures	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part
Letter names	0.09	0.03	.16	-0.09	-0.03	-.02	-0.37	-0.13	-.08	-0.33	-0.12	-.08
Letter sounds	0.34	0.09	.08	0.61	0.18	.12	0.15	0.05	.03	0.04	0.01	.01
Phoneme segmenting	0.10	0.03	.07	0.01	0.00	.00	0.64	0.14	.13	0.64*	0.15*	.14
Word reading fluency	1.04**	0.43**	.36	1.01**	0.45**	.38	1.15**	0.54**	.45	1.23**	0.61**	.46
R^2		.14			.28			.28			.32	
N		148			153			149			161	

Note. *$p < .05$, **$p < .001$, Part = Semipartial Correlation.

Table 21

Grade 1 Regression Models Predicting the easyCBM© Word Reading Fluency measure at Time 5

	Time 1			Time 2			Time 3			Time 4		
easyCBM © measures	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part	*B*	*β*	Part
Letter names	0.86**	0.63**	.42	0.80**	0.61**	.40	0.73**	0.53**	.38	1.01**	0.78**	.64
Letter sounds	0.23	0.13	.09	0.31*	0.19*	.12	0.34*	0.22*	.15	-0.14	-0.09	-.07
Phoneme segmenting	0.13	0.08	.07	-0.20	-0.10	-.09	-0.06	0.13	-.03	-0.15	-0.07	-.06
R^2		.56			.54			.49			.53	
N		170			175			171			184	

Note. *$p < .05$, **$p < .001$, Part = Semipartial Correlation.

Reference

Wray, K. A., Lai, C.-F., Sáez, L., Alonzo, J., & Tindal, G. (2013). *easyCBM® kindergarten beginning reading measures: Alternate form reliability and criterion validity with the SAT-10 (Technical Report #1403).* Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 2401: Criterion validity and classification accuracy of easyCBM® Grades 3-8 (Tindal & Nese, 2024).

Methods

TR2401 examines criterion validity and classification accuracy for easyCBM® reading benchmarks in Grades 3–8, using the Smarter Balanced (SB) English Language Arts (ELA) test as the external criterion. The sample includes well over 8,000 anonymized student records from two states and four school districts. Students completed easyCBM® benchmarks in fall 2023, winter 2024, and spring 2024 on three reading measures: Proficient Reading (20 items), Vocabulary (20 items), and Oral Reading Fluency (ORF; words correct per minute). SB ELA was administered in spring.

For predictive/concurrent validity, the report presents grade-specific correlation matrices (Tables 19–24 of the Technical Report) linking each easyCBM® reading measure to SB ELA in fall, winter, and spring. Because the reading measures are on different scales, the authors also compute a reading composite by converting each reading score to a grade-specific z score and averaging within grade; the composite is then correlated with SB ELA by season.

For classification accuracy, SB ELA serves as the criterion and "risk" is primarily defined at the 20th percentile. For each season and reading measure, the report provides (a) an evidence table with the SB cut score, the corresponding easyCBM® cut score, confusion-matrix counts (true/false positives/negatives), and AUC with a 95% confidence interval; and (b) a summary table reporting sample size, base rate, overall classification rate, sensitivity, specificity, false positive/negative rates, and positive/negative predictive power.

Results

Criterion validity evidence indicates that all three easyCBM® reading measures relate meaningfully to SB ELA, with strength varying by grade, season, and measure. For Proficient Reading, correlations with SB ELA range from .53 (Grade 8 fall) to .70 (Grade 4 fall), with most values in the high .50s to high .60s across grades and seasons (e.g., Grade 3: .63 fall, .60 winter, .69 spring; Grade 5: .62–.66). Vocabulary shows similar or slightly stronger relations in several grades, ranging from .55 (Grade 7 winter) to .69 (Grade 3 winter), with many values around .61–.67 (e.g., Grade 4: .67 fall/winter/spring; Grade 6: .62 fall, .58 winter, .66 spring). ORF correlations are more variable and generally lower in upper grades, ranging from .48 (Grade 6 fall) and .49 (Grade 7 winter) to .66 (Grade 3 winter) and .65 (Grade 3 fall/spring). These patterns support the interpretation that easyCBM® reading benchmarks provide moderate-to-strong predictive/concurrent evidence with SB ELA, and that alignment between the easyCBM® measure and the SB construct is strongest for Proficient Reading and Vocabulary in most grades.

The reading composite reinforces the overall relationship with SB ELA: the composite correlation is .64 in fall, .57 in winter, and .66 in spring. Taken together, these coefficients show that easyCBM® reading performance months before SB (fall/winter) is meaningfully associated with end-of-year SB ELA.

Classification accuracy results (risk defined at the SB 20th percentile) show that the reading measures can screen for risk with generally strong AUCs and favorable operating characteristics. For Proficient Reading, AUCs are strong across seasons (Fall: .80–.85; Winter: .76–.85; Spring: .80–.84). Corresponding sensitivity and specificity are typically in the .70–.80+ range (e.g., fall sensitivity .68–.83; spring specificity up to .88). Base rates cluster around .15–.24. Positive predictive power is consistently very high (often .91–.96), while negative predictive power is lower (often ~.33–.56), reflecting base rates and the sensitivity/specificity balance.

ORF also shows strong screening accuracy. AUCs are .75–.85 in fall, .77–.86 in winter, and .80–.84 in spring. Sensitivity is often strong in Grades 3–6 (e.g., winter Grade 3 = .83; spring Grade 4 = .82) but weaker in some upper-grade cells (e.g., fall Grade 7 = .62). Specificity ranges from about .59 to .85 depending on grade and season. Vocabulary yields similarly strong AUCs (Fall: .81–.84; Winter: .80–.89; Spring: .82–.87) with generally balanced sensitivity and specificity and high positive predictive power. Across measures, TR2401 concludes that easyCBM® reading benchmarks provide credible criterion validity with SB ELA and strong classification accuracy for identifying students at risk when using season- and grade-specific cut scores.

NCII-aligned summaries show overall classification rates typically around .73–.83 for Proficient Reading (e.g., fall Grade 6 = .80; spring Grade 5 = .83), about .70–.81 for ORF (with a lower value in fall Grade 7 = .65), and about .73–.83 for Vocabulary. Thus, roughly three-quarters of students are correctly classified in most grade/season combinations relative to the SB 20th-percentile criterion.

Reference

Tindal, G. & Nese, J. F. T., & (2024). *Criterion validity and classification accuracy of easyCBM®: Grades 3-8. (Technical Report # 2401).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0910: Criterion-related Evidence Using easyCBM® Reading Measures and Student Demographics to Predict State Test Performance in Grades 3-8 (Tindal et al., 2009a).

Methods

This technical report examines the relationship between easyCBM® reading benchmark assessments, student demographic characteristics, and performance on the Oregon Assessment of Knowledge and Skills (OAKS) reading test. The study was designed to evaluate how well easyCBM® measures and demographic variables predict student performance on a statewide standardized reading assessment.

The study used student data from two Oregon school districts and included students in grades 3 through 8. Sample sizes varied by grade and district but generally ranged from several hundred to over one thousand students per grade. The second district included a higher proportion of economically disadvantaged students and students receiving special education services.

Three easyCBM® reading benchmark measures served as predictors: passage reading fluency, vocabulary, and multiple-choice reading comprehension. Passage reading fluency was measured as the number of words read correctly per minute during individually administered reading tasks. Vocabulary and comprehension measures were computer-based, group-administered assessments.

The outcome variable was performance on the OAKS reading test, a computer-adaptive statewide assessment with vertically scaled scores. Students completed the easyCBM® benchmarks during fall, winter, and spring testing periods. Demographic variables included gender, ethnicity, economic disadvantage, Title I status, and special education status. Data from the easyCBM® database were merged with district demographic and testing records. Analyses included correlation analyses and multiple regression models to determine how well the benchmark measures and demographic factors predicted state reading test scores.

Results

Results showed consistently strong relationships between easyCBM® reading measures and the OAKS reading test across grades and benchmark periods. Correlations between the benchmark assessments and the state test were typically above .60. Regression analyses indicated that the combined predictors explained moderate to substantial portions of variance in state reading scores, with R^2 values generally ranging from about .50 to .65. The easyCBM® reading measures consistently accounted for the largest share of explained variance. Demographic variables such as economic disadvantage and special education status were sometimes significant predictors but showed smaller and less consistent effects. Overall, the findings indicate that easyCBM® reading benchmarks provide meaningful predictive evidence for statewide reading assessment performance

Table 14. Illustrative Table from Technical Report 0910

Correlations

		Passage Reading Fluency score, fall	Vocabulary score, fall	Multiple Choice Reading Comprehension score, fall	OAKS reading score, 2008-2009
Passage Reading Fluency	Pearson Correlation	1	.715**	.629**	.628**
	Sig. (2-tailed)		.000	.000	.000
	N	338	314	315	338
Vocabulary score	Pearson Correlation	.715**	1	.640**	.676**
	Sig. (2-tailed)	.000		.000	.000
	N	314	314	314	314
Multiple Choice Reading Comprehension	Pearson Correlation	.629**	.640**	1	.624**
	Sig. (2-tailed)	.000	.000		.000
	N	315	314	315	315
OAKS reading score, 2008-2009	Pearson Correlation	.628**	.676**	.624**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	338	314	315	797

**. Correlation is significant at the 0.01 level (2-tailed).

Reference

Tindal, G., Nese, J. F., & Alonzo, J. (2009). *Criterion-related evidence using easyCBM® reading measures and student demographics to predict state test performance in Grades 3-8 (Technical Report # 0910)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1106: Diagnostic Efficiency of easyCBM® Reading: Oregon (Park, Anderson, et al., 2011).

Methods

This report examined the diagnostic efficiency of easyCBM® reading benchmark assessments for predicting performance on the Oregon Assessment of Knowledge and Skills (OAKS). Data were collected during the 2009–2010 academic year from three Oregon districts. The sample included approximately 2,000 students per grade in grades 3–8.

Measures included Passage Reading Fluency (PRF), Multiple-Choice Reading Comprehension (MCRC), and Vocabulary (VOC). Fluency and comprehension were administered fall, winter, and spring; vocabulary was administered fall and spring.

Receiver Operating Characteristic (ROC) analyses were conducted to evaluate classification accuracy. Sensitivity, specificity, and area under the curve (AUC) statistics were calculated for each measure and benchmark occasion. Cut score selection emphasized maximizing sensitivity while maintaining specificity above .70, consistent with RTI priorities aimed at minimizing false negatives. Demographic subgroup comparisons were conducted to ensure comparability across participating districts.

Results

Across grades and seasons, AUC values generally fell within acceptable to strong ranges, indicating meaningful discrimination between students who passed and did not pass OAKS. PRF demonstrated particularly strong diagnostic utility in winter and spring administrations. MCRC and VOC also demonstrated adequate discrimination, with comprehension measures often showing stronger predictive utility at upper grades.

Sensitivity was consistently prioritized in cut score selection, often exceeding .80 while maintaining acceptable specificity. Results indicated that easyCBM® benchmark measures provided strong criterion-related classification evidence when aligned with Oregon's state proficiency standards.
However, performance varied slightly by grade and season, reinforcing the importance of grade-specific cut score interpretation. Overall findings support the use of easyCBM® reading benchmarks as effective RTI screening tools in Oregon.

Table 15. Illustrative Table from Technical Report 1106

Table 30
Resulting Statistics for Each Chosen Cut Score: ELL, PRF

Grd	Season	Meets Score	*n*	Failure Base Rate	False Positive Rate	False Negative Rate	Sensitivity	Specificity	Positive Predictive Power	Negative Predictive Power	Overall Correct Classification	AUC
3	Fall	62	97	0.25	0.32	0.17	0.83	0.68	0.47	0.93	0.72	0.81
	Winter	90	97	0.26	0.39	0.20	0.80	0.61	0.42	0.90	0.66	0.82
	Spring	93	93	0.22	0.32	0.25	0.75	0.68	0.39	0.91	0.70	0.76
4	Fall	96	100	0.39	0.33	0.18	0.82	0.67	0.62	0.85	0.73	0.82
	Winter	111	77	0.38	0.25	0.14	0.86	0.75	0.68	0.90	0.79	0.87
	Spring	126	81	0.37	0.37	0.10	0.90	0.63	0.59	0.91	0.73	0.88
5	Fall	138	95	0.49	0.54	0.11	0.89	0.46	0.62	0.81	0.67	0.75
	Winter	142	75	0.44	0.50	0.15	0.85	0.50	0.57	0.81	0.65	0.76
	Spring	160	90	0.46	0.49	0.17	0.83	0.51	0.59	0.78	0.66	0.79
6	Fall	135	48	-	-	-	-	-	-	-	-	-
	Winter	149	25	-	-	-	-	-	-	-	-	-
	Spring	154	35	-	-	-	-	-	-	-	-	-
7	Fall	144	145	0.77	0.76	0.12	0.88	0.24	0.79	0.38	0.73	0.66
	Winter	160	140	0.76	0.79	0.08	0.92	0.21	0.79	0.44	0.75	0.75
	Spring	152	154	0.77	0.67	0.12	0.88	0.33	0.81	0.46	0.75	0.69
8	Fall	166	149	0.85	0.70	0.06	0.94	0.30	0.88	0.50	0.85	0.78
	Winter	157	150	0.85	0.57	0.07	0.93	0.43	0.90	0.53	0.85	0.80
	Spring	160	156	0.83	0.50	0.05	0.95	0.50	0.90	0.65	0.87	0.81

Note. AUC = Area Under the ROC Curve

Reference

Park, B. J., Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). *Diagnostic efficiency of easyCBM® reading: Oregon (Technical Report No. 1106)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1107: Diagnostic Efficiency of easyCBM® Reading: Washington State (Anderson, Park, et al., 2011).

Methods

This study evaluated the diagnostic efficiency of easyCBM® reading benchmark assessments in predicting performance on Washington's Measurements of Student Progress (MSP). Data were collected during the 2009–2010 academic year from two Washington districts. The sample included approximately 900 students per grade in grades 3–8. Measures included PRF, MCRC, and VOC administered across fall, winter, and spring benchmark windows. ROC curve analyses were conducted for each measure and grade. AUC statistics were calculated to evaluate overall discrimination, and sensitivity and specificity were examined for all potential cut scores. Cut score decisions emphasized minimizing false negatives, consistent with RTI implementation priorities. Demographic comparisons confirmed that subgroup representation did not meaningfully bias classification outcomes.

Results

AUC statistics demonstrated acceptable to strong classification accuracy across grades and measures. PRF generally yielded strong discrimination, particularly in winter and spring. MCRC and VOC demonstrated moderate to strong diagnostic utility depending on grade level. Sensitivity values were typically emphasized above specificity in final cut score decisions, reflecting the goal of identifying students at risk of failing MSP. Classification accuracy was strongest in lower and middle grades and showed some variability in upper grades. Overall, findings support the use of easyCBM® reading benchmarks for predicting MSP outcomes and informing RTI tier placement decisions in Washington.

Table 16. Illustrative Table from Technical Report 1107

Table 5
Resulting Statistics for Each Chosen Cut Score: Full Sample, PRF

Grd	Season	Meets Score	*n*	Failure Base Rate	False Positive Rate	False Negative Rate	Sensitivity	Specificity	Positive Predictive Power	Negative Predictive Power	Overall Correct Classification	AUC
	Fall	79	253	0.36	0.29	0.15	0.85	0.71	0.62	0.89	0.76	0.84
3	Winter	106	285	0.35	0.29	0.22	0.78	0.71	0.59	0.86	0.73	0.81
	Spring	101	310	0.35	0.30	0.21	0.79	0.70	0.59	0.86	0.73	0.82
	Fall	109	131	0.44	0.49	0.26	0.74	0.51	0.54	0.71	0.61	0.66
4	Winter	124	245	0.38	0.38	0.26	0.74	0.62	0.55	0.79	0.67	0.71
	Spring	132	259	0.40	0.40	0.24	0.76	0.60	0.56	0.79	0.66	0.76
	Fall	130	146	0.42	0.30	0.18	0.82	0.70	0.67	0.84	0.75	0.84
5	Winter	140	248	0.44	0.30	0.29	0.71	0.70	0.66	0.75	0.71	0.80
	Spring	161	228	0.43	0.40	0.19	0.81	0.60	0.60	0.81	0.69	0.78
	Fall	145	231	0.44	0.39	0.30	0.70	0.61	0.59	0.72	0.65	0.74
6	Winter	176	241	0.45	0.38	0.25	0.75	0.62	0.62	0.75	0.68	0.77
	Spring	177	61	0.57	0.38	0.29	0.71	0.62	0.71	0.62	0.67	0.75
	Fall	143	198	0.36	0.30	0.24	0.76	0.70	0.59	0.84	0.72	0.78
7	Winter	165	428	0.41	0.30	0.30	0.70	0.70	0.62	0.77	0.70	0.76
	Spring	152	259	0.38	0.40	0.29	0.71	0.60	0.52	0.78	0.64	0.72
	Fall	144	486	0.39	0.30	0.22	0.78	0.70	0.62	0.83	0.73	0.82
8	Winter	153	508	0.37	0.29	0.26	0.74	0.71	0.60	0.82	0.72	0.81
	Spring	174	510	0.36	0.30	0.19	0.81	0.70	0.61	0.86	0.74	0.80

Note. AUC = Area Under the ROC Curve

Reference

Anderson, D., Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). *Diagnostic efficiency of easyCBM® reading: Washington State (Technical Report No. 1107)*. University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1108: Cross-validation of easyCBM® Reading Cut Scores in Oregon: 2009–2010 (Grades 3 – 8) (Park, Irvin, et al., 2011).

Methods

This report extended prior diagnostic efficiency work by conducting a cross-validation study of easyCBM® reading benchmark cut scores in Oregon. Data from approximately 2,000 students per grade (3–8) were randomly split into two demographically comparable groups. Measures included PRF, MCRC, and VOC administered across seasonal benchmarks. ROC analyses were conducted separately for each split sample. AUC statistics and sensitivity/specificity values were computed for all possible cut scores. Cut score selection followed modified Silberglitt and Hintze decision rules, emphasizing sensitivity thresholds above .70 and preferably above .80 while maintaining acceptable specificity. Stability of optimal cut scores was evaluated by comparing differences between groups and examining overlap of 95% confidence intervals for AUC statistics.

Results

Optimal cut scores demonstrated strong stability across randomly split groups. The average difference between PRF cut scores across groups was approximately 1–2 CWPM across grade-level comparisons. Differences for MCRC and VOC were typically less than one raw score point. AUC confidence intervals overlapped across groups for all grades and benchmark occasions, indicating non-significant differences in classification accuracy. Sensitivity and specificity patterns remained consistent across groups. These findings provide strong cross-validation evidence that previously identified Oregon cut scores are statistically stable and robust. However, authors caution that cut scores are criterion-dependent and should not be generalized across states without validation.

Table 17. Illustrative Table from Technical Report 1108

Grade 8
Spring PRF Benchmark

Cut score	Group 1		Group 2	
	Sensitivity	Specificity	Sensitivity	Specificity
14.00	.000	1.000	-	-
17.00	.002	1.000	-	-
21.00	-	-	.000	1.000
22.50	.005	1.000	-	-
25.00	-	-	.002	1.000
30.50	-	-	.005	1.000
31.50	.007	1.000	-	-
35.50	-	-	.007	1.000
40.00	.010	1.000	-	-
41.50	-	-	.009	1.000
48.00	.012	1.000	.012	1.000
53.50	.015	1.000	-	-
54.00	-	-	.014	1.000
54.50	.017	1.000	-	-
57.00	.020	1.000	-	-
57.50	-	-	.016	1.000
59.00	-	-	.019	1.000
60.50	.025	1.000	-	-
61.00	-	-	.021	1.000
62.50	.027	1.000	.026	1.000
63.50	.029	1.000	.028	1.000
64.50	-	-	.031	1.000
65.50	-	-	.035	1.000
66.00	.032	1.000	-	-
67.50	-	-	.040	1.000
69.50	-	-	.042	1.000
70.50	.034	1.000	.045	1.000
71.50	-	-	.047	.999
72.50	-	-	.049	.999
73.50	-	-	.052	.999
74.00	.037	1.000	-	-
74.50	-	-	.056	.999
75.50	.039	1.000	.061	.999
77.00	-	-	.066	.999
77.50	.042	1.000	-	-
78.50	-	-	.071	.999
79.50	.044	1.000	.073	.999

Reference

Park, B. J., Irvin, P. S., Anderson, D., Alonzo, J., & Tindal, G. (2011). *Cross-validation of easyCBM® reading cut scores in Oregon: 2009–2010 (Technical Report No. 1108)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1109: Cross-validation of easyCBM® Reading Cut Scores in Washington: 2009–2010 (Irvin et al., 2011).

Methods

This report conducted a cross-validation study of easyCBM® reading benchmark cut scores aligned with Washington's MSP assessment. Approximately 900 students per grade (3–8) were randomly divided into two comparable samples. Measures included PRF, MCRC, and VOC administered across fall, winter, and spring. ROC analyses were conducted separately within each group. AUC statistics were compared using overlapping 95% confidence intervals. Cut scores were selected using sensitivity-prioritized decision rules designed to minimize false negatives in RTI implementation. Demographic comparability between groups was examined via t-tests across ethnicity, SPED, ELL, and FRL categories.

Results

Optimal cut scores were reasonably stable across randomly split groups, though slightly more variable than observed in Oregon. Average PRF cut score differences were approximately 7 CWPM across grade-level comparisons. Differences for MCRC and VOC were generally under one point. AUC confidence intervals overlapped for all measures and grades, indicating statistically equivalent classification accuracy across groups. Sensitivity values were typically maintained at or above .80 while specificity remained acceptable.
Overall, findings provide strong evidence that identified Washington cut scores are stable and defensible for MSP classification prediction. Authors note that cut scores should be interpreted within state-specific accountability contexts.

Table 18. Illustrative Table from Technical Report 1109

Grade 4
Fall MCRC Benchmark

Cut score	Group 1		Group 2	
	Sensitivity	Specificity	Sensitivity	Specificity
-1.00	.000	1.000	.000	1.000
1.50	-	-	.000	.973
2.00	.037	.972	-	-
3.50	-	-	.032	.973
4.50	.111	.972	.065	.946
5.50	.185	.972	.097	.919
6.50	.259	.917	.097	.865
7.50	.370	.889	.355	.838
8.50	.444	.806	.516	.784
9.50	.630	.778	**.710**	**.757**
10.50	.667	.722	.774	.649
11.50	**.741**	**.639**	.871	.622
12.50	.815	.528	.935	.486
13.50	.852	.361	.968	.297
14.50	.926	.361	1.000	.297
15.50	1.000	.250	1.000	.216
16.50	1.000	.167	1.000	.108
17.50	1.000	.139	1.000	.081
18.50	1.000	.028	1.000	.027
20.00	1.000	.000	1.000	.000

Reference

Irvin, P. S., Park, B. J., Anderson, D., Alonzo, J., & Tindal, G. (2011). *Cross-validation of easyCBM® reading cut scores in Washington: 2009–2010 (Technical Report No. 1109)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1302: Examining the Internal Structure of the easyCBM® Reading Measures, K–5 (Alonzo et al., 2013c).

Methods

This technical report investigated the **internal structure** (construct validity evidence) of the easyCBM® CCSS reading measures for Grades K–5. The study used student assessment data collected during a defined administration window and evaluated whether the measures' observed score patterns matched the hypothesized constructs of early literacy, fluency, and comprehension across grade levels. The authors applied confirmatory factor analysis (CFA) to test competing measurement models, including one-factor versus multi-factor structures and versions that accounted for "testlet effects" (shared variance among grouped items, such as comprehension questions tied to the same passage). Analyses were organized by grade bands (notably K–2 versus 3–5) to reflect developmental differences in reading skills and the structure of the assessment system. The report examined parameter estimates such as factor loadings (how strongly each measure related to its intended construct), correlations among latent constructs, and variance explained in observed indicators. Attention was given to data features that can distort CFA results, including non-normal score distributions and ceiling effects on some components, and the authors interpreted models with those constraints in mind. Fit was evaluated using standard indices and comparative evidence across models rather than relying on a single statistic.

Therefore, this report also provides important construct validity evidence for easyCBM® reading measures using confirmatory factor analysis (CFA). Students in grades K–5 completed a battery of easyCBM® reading assessments. For grades K–2, measures reflect early literacy (Phoneme Segmenting, Letter Names, Letter Sounds), fluency (two

Word Reading Fluency forms and Passage Reading Fluency), and comprehension (CCSS-style comprehension and MCRC). Competing CFA models (e.g., one-factor vs multi-factor; with/without testlet effects) are compared using model fit indices, and standardized factor loadings are examined to evaluate alignment of measures with hypothesized constructs.

Results

Findings generally supported a multi-component reading structure that aligns with how the easyCBM® CCSS system is intended to function, while also flagging some grade-band-specific complications. For Grades K–2, the early literacy indicators (e.g., letter knowledge and phonological measures) and fluency indicators tended to show strong relationships to their intended factors, while some comprehension components demonstrated weaker associations under certain model specifications. Where comprehension performance showed heavy skew (including strong ceiling effects), model estimates for comprehension paths were less stable and appeared to influence overall fit and covariance patterns. When alternative model specifications were examined, the comprehension indicators typically showed improved associations with the comprehension construct, and the overall model became more interpretable. Across analyses, the latent constructs showed meaningful relationships with one another (generally moderate associations), consistent with the idea that early literacy, fluency, and comprehension are related but distinguishable skills. Overall, the results provided supportive evidence for the construct validity of the K–5 easyCBM® CCSS reading measures, while also identifying places where components (especially early comprehension subparts in lower grades) may require careful interpretation due to distributional issues.

For the K–2 sample, CFA results support a multi-dimensional reading structure and report that a three-factor model (Early Literacy, Fluency, Comprehension) provides the best fit among the tested models. Standardized factor loadings (lambda coefficients) are strong for most early literacy and fluency indicators: Phoneme Segmenting loads moderately (.56), Letter Names strongly (.89), and Letter Sounds strongly (~.82) on the Early Literacy factor; fluency indicators load strongly (e.g., WRF1 .88–.89; WRF2 .80–.81; PRF .84–.85). Comprehension indicators vary: one CCSS comprehension subcomponent shows a near-zero loading (.03), consistent with weak association/ceiling effects, whereas MCRC comprehension shows moderate loadings (.55–.67). Factor correlations show Early Literacy and Fluency are moderately related (.61–.62), while Fluency and Comprehension show a moderate negative relation in one model (about −.33), and some early literacy–comprehension relations are small/non-significant in the reported solutions. Overall, the findings support interpreting easyCBM® reading measures as assessing theoretically coherent but distinct constructs at K–2, strengthening the construct validity argument for these measures.

Table 19. Example Construct Validity Output

Table 3
CFA Model Fit Summary for Grade K-2 Sample

	Grades K-2			
	1st Model	2nd Model	**3rd Model**	4th Model
CFI	0.56	0.54	**0.77**	0.73
TLI	0.38	0.41	**0.62**	0.62
RMSEA	0.31	0.25	**0.24**	0.20
SRMR	0.15	0.13	**0.08**	0.08
AIC	8254.31	10764.87	**8136.82**	10650.39
BIC	8322.76	10850.43	**8213.82**	10744.50
ABIC	8246.86	10755.56	**8128.44**	10640.14

1st Model: One factor CFA model without testlet effect
2nd Model: One factor CFA model with testlet effect
3rd Model: Three factor CFA model without testlet effect
4th Model: Three factor CFA model with testlet effect

Table 4
Maximum Likelihood Parameter Estimates of Three Factor Confirmatory Factor Analysis Model with Testlet Effects for Grades K-2

Parameters	Early Literacy		Fluency		Comprehension	
	3rd	4th	3rd	4th	3rd	4th
	Lambda coefficients					
Phoneme segmenting	0.56	0.54				
Letter names	0.89	0.89				
Letter sounds	0.82	0.82				
WRF1			0.88	0.89		
WRF2			0.81	0.80		
PRF			0.85	0.84		
Comprehension ↔ Part1A						0.73
Comprehension ↔ Part1B					0.78	0.61
Comprehension ↔ Part1C						0.03
Comprehension ↔ Part2					0.67	0.55
	Factor correlation					
	Model 3			Model 4		
Early Literacy ↔ Fluency	0.61			0.62		
Early Literacy ↔ Comprehension	0.11 NS			-0.12 NS		
Fluency ↔ Comprehension	-0.33			-0.54		
	Variance explained by the model					
	Model 3			Model 4		
PS	0.31			0.29		
LN	0.79			0.79		
LS	0.67			0.68		
WRF1	0.77			0.79		
WRF2	0.66			0.64		
PRF	0.71			0.70		
Part1A				0.53		
Part1B	0.60			0.38		
Part1C				<0.01		
Part2	0.45			0.30		

Note. Parameter estimates reported in this table are standardized coefficients.

Table 20. Illustrative Table from Technical Report 1302

Table 12
Maximum Likelihood Parameter Estimates of Two Factor Confirmatory Factor Analysis Model with Testlet Effects for Grades 3-5

Parameter	Grade 3		Grade 4		Grade 5	
	Fluency	Comp	Fluency	Comp	Fluency	Comp
	Lambda coefficients					
Fluency ↔ WRF	0.82		0.91		0.90	
Fluency ↔ PRF1	0.97		0.97		0.99	
Fluency ↔ PRF2	0.95		0.98		0.96	
Comprehension ↔ Part1A		0.42		0.44		0.96
Comprehension ↔ Part1B		0.88		0.20		0.97
Comprehension ↔ Part1C		0.74		0.72		0.70
Comprehension ↔ Part2		0.34		0.81		-0.37
	Factor correlation					
Fluency ↔ Comprehension	0.47		0.04 NS		-0.56	
	Variance explained by the model					
WRF	0.64		0.83		0.81	
PRF1	0.95		0.94		0.99	
PRF2	0.91		0.95		0.92	
Part1A	0.18		0.20		0.91	
Part1B	0.78		0.04		0.94	
Part1C	0.55		0.51		0.49	
Part2	0.12		0.66		0.14	

Note. Parameter estimates reported in this table are standardized coefficients.

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2013). *Examining the internal structure of the easyCBM® reading measures, grades K–5* (Technical Report #1302). Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1304: An Examination of the Internal Structures of the Gr. 6–8 easyCBM CCSS Reading Measures: A Construct Validity Study (Alonzo et al., 2013a).

Methods

This study examined the internal structure of the Grades 6–8 easyCBM® CCSS Reading assessments using confirmatory factor analysis. The purpose was to determine whether the assessments functioned as unidimensional measures of literal reading comprehension or whether distinct factors emerged based on passage genre. Participants included 482 students in Grades 3–5 drawn from multiple states; the design was later extended to middle grades. Each test form contained 45 items organized into nine testlets representing three genres: Read to Perform a Task, Informational Text, and Short Literary Text. Each testlet included five selected-response items. Using maximum likelihood estimation, researchers tested one-factor, two-factor, and three-factor models. Model fit indices including CFI, TLI, RMSEA, and chi-square statistics were evaluated. Factor loadings, variance estimates, and squared multiple correlations were examined to determine dimensional coherence.

Results

Across grades, results consistently supported a unidimensional interpretation of the CCSS Reading measures. Although certain two-factor models demonstrated comparable fit, factor correlations were high, indicating that genre distinctions did not meaningfully separate constructs. Three-factor models were not supported. The one-factor model provided acceptable fit across grades, supporting the interpretation that the assessments measure a broad construct of literal reading comprehension. The findings reinforce the design intention that items drawn from different genres contribute to a unified comprehension construct rather than distinct subskills.

Table 21. Illustrative Table from Technical Report 1304

Table 6
Grade 3 Regression Weights, Two-Factor Model (Read to Perform a Task & Informational Text/Short Literary Text)

	Unstandardized				Standardized
	Estimate	S.E.	C.R.	P	
T1<--- F1	1.00				.69
T2<--- F1	0.73	.12	6.35	***	.64
T3<--- F1	1.12	.14	8.13	***	.83
T4<--- F2	1.00				.83
T5<--- F2	1.06	.07	11.09	***	.85
T6<--- F2	1.03	.09	11.31	***	.87
T7<--- F2	1.00	.10	10.15	***	.81
T8<--- F2	0.98	.11	9.06	***	.75
T9<--- F2	0.97	.12	8.04	***	.67

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2013). *An examination of the internal structures of the easyCBM CCSS reading measures (Technical Report # 1304).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1305: An Examination of the Internal Structures of the Gr. K–5 easyCBM® CCSS Reading Measures: A Construct Validity Study (Alonzo et al., 2013b).

This study examined the internal structure and construct validity of the easyCBM® K–5 reading measures using confirmatory factor analysis (CFA). The purpose was to determine whether the assessments measured a single general reading construct or multiple related but distinct constructs, and whether genre or measure type altered the factor structure.

Methods

The study was conducted in winter 2012 with 22 teachers and 519 students across kindergarten through Grade 5. Participants were recruited from multiple states and included students with disabilities and English learners. Teachers administered the assessments using standardized protocols.

Different measures were administered by grade band. In Grades K–2, students completed early literacy measures (phoneme segmenting, letter names, letter sounds), word reading fluency, passage reading fluency, and two comprehension measures (CCSS and MCRC). Fluency measures were individually administered, while comprehension measures were group administered. In Grades 3–5, students completed word reading fluency, two passage reading fluency measures, and both CCSS and MCRC comprehension assessments.

Four CFA models were tested for each grade band using maximum likelihood estimation in Mplus. For Grades K–2, models compared one-factor versus three-factor solutions, with and without testlet effects. For Grades 3–5, models compared one-factor and two-factor solutions, again with and without testlet effects. Model fit was evaluated using CFI, TLI, RMSEA, SRMR, and information criteria (AIC, BIC, ABIC). Due to negative skew in CCSS comprehension scores, collapsed scoring procedures were implemented to address convergence issues.

Results

For Grades 3–5, results varied somewhat by grade. Both one-factor and two-factor models showed acceptable fit, but the two-factor model separating fluency and comprehension generally demonstrated slightly better theoretical and statistical support. Fluency indicators consistently showed strong loadings. Comprehension measures also loaded meaningfully, despite differences in passage length and genre between CCSS and MCRC measures. Importantly, both comprehension types loaded on a single comprehension factor, suggesting that shorter, more accessible CCSS measures assess the same underlying construct as longer MCRC measures.

Overall, the findings support the construct validity of the easyCBM® reading measures. In early grades, reading appears best represented by multiple interrelated constructs, while in upper elementary grades, fluency and comprehension remain distinguishable but more closely integrated. The results provide evidence that the system's measures align with their intended theoretical constructs and can be used confidently for screening and progress monitoring purposes.

Table 22. Illustrative Table from Technical Report 1305

Table 11
CFA Model Fit Summary for Grades 3-5

Grade 3					Grade 4				Grade 5			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
CFI	0.96	0.83	-	**0.84**	0.94	0.86	-	**0.98**	0.97	0.53	1.00	**0.96**
TLI	0.93	0.74	-	**0.74**	0.88	0.80	-	**0.97**	0.95	0.29	1.01	**0.94**
RMSEA	0.17	0.24	-	**0.24**	0.23	0.21	-	**0.08**	0.16	0.49	0.00	**0.15**
SRMR	0.06	0.12	-	**0.15**	0.11	0.14	-	**0.07**	0.07	0.18	0.01	**0.09**
AIC	3529.52	4191.76	-	**4187.88**	3983.52	4263.06	-	**4207.04**	2485.10	3144.17	2475.93	**2933.84**
BIC	3568.45	4246.26	-	**4244.97**	4023.61	4319.19	-	**4265.84**	2518.39	3190.78	2511.44	**2982.67**
ABIC	3521.08	4179.94	-	**4175.49**	3976.22	4252.84	-	**4196.33**	2471.16	3124.65	2461.06	**2913.39**

1st Model: One-factor CFA model without testlet effect
2nd Model: One-factor CFA model with testlet effect
3rd Model: Two-factor CFA model without testlet effect
4th Model: Two-factor CFA model with testlet effect
* - Model did not converge or indicated significant errors, thus the results are not interpreted

Reference

Alonzo, J., Park, B. J., & Tindal, G. (2013). *An examination of the internal structures of the Gr. K–5 easyCBM CCSS reading measures: A construct validity study (Technical Report No. 1305).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0025: Analysis of Reading Fluency and Comprehension Measures for First Grade Students (Alonzo & Tindal, 2004a).

This technical report (Technical Report 0025) presents a comprehensive evaluation of reading assessment measures designed for universal screening, benchmarking, and progress monitoring.

Methods

The study draws upon large and diverse student samples representing multiple grade levels. Standardized administration protocols were implemented to ensure consistency across testing sites. Analyses included classical test theory indices such as internal consistency reliability, item-total correlations, and standard error of measurement. Item response theory modeling was applied where appropriate to examine item difficulty calibration, discrimination parameters, and overall model fit. Dimensionality analyses were conducted to evaluate whether the measures captured a unidimensional construct of reading proficiency or reflected multiple latent traits. Additional procedures included scaling analyses to ensure vertical alignment across grade levels, subgroup analyses to explore differential item functioning, and growth modeling to assess sensitivity to instructional change. Classification accuracy statistics were calculated to evaluate screening cut scores and decision consistency. Together, these analytic procedures were designed to establish both technical adequacy and practical utility for educational decision-making.

Results

Results indicate that the reading measures demonstrate acceptable to strong reliability across grades and administration periods. Reliability coefficients consistently met or exceeded commonly accepted benchmarks for screening and progress monitoring applications. Item analyses revealed that most items functioned within recommended statistical thresholds for difficulty and discrimination, with only a small number flagged for further review due to misfit or instability. Dimensionality findings generally supported a dominant reading proficiency construct, while scaling analyses confirmed appropriate score interpretation across grade levels. Classification accuracy indices demonstrated adequate sensitivity and specificity for identifying students at risk. **Predictive and concurrent validity** evidence showed meaningful relationships between these measures and external indicators of reading performance. Growth analyses suggested that the assessments were sensitive to developmental change over time, supporting their use for longitudinal progress monitoring. Overall, findings support the technical soundness of these reading assessment measures and their suitability for informing instructional planning, intervention decisions, and system-level educational evaluation.

Reference

Alonzo, J., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for first grade students (Technical Report No. 0025).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0022: Analysis of Reading Fluency and Comprehension Measures (3rd Grade) (Ketterlin-Geller & Tindal, 2004).

Methods

The study analyzed third-grade student data collected during a district-wide reading assessment administration. Measures included timed oral reading fluency passages and structured comprehension probes aligned with grade-level standards. Oral reading fluency was operationalized as words read correctly per minute. Comprehension performance was evaluated using passage-based response items. Statistical analyses included correlation analysis between fluency and comprehension scores, internal consistency reliability estimates, and score distribution evaluations. Subgroup analyses examined patterns across achievement levels to determine the extent to which fluency functioned as a proxy for comprehension performance.

Results

Results indicated a strong relationship between oral reading fluency and comprehension performance in third grade. Students with higher fluency rates tended to demonstrate stronger comprehension outcomes. Reliability estimates for both measures were within acceptable ranges for screening and progress monitoring .Fluency measures effectively identified students at risk for reading difficulties, while comprehension measures added depth to instructional interpretation. Overall findings supported the technical adequacy of the third grade reading assessment system and confirmed fluency as a robust indicator of reading proficiency at this grade level.

Table 23. Illustrative Table from Technical Report 0022

Table 7

Regression Summary for the Composite State Score in Reading.

	Unstandardized Coefficients		Standardized Coefficients		95% Confidence Interval for B	
Independent Variables	B	Std. Error	Beta	t	Lower Bound	Upper Bound
ORF	0.09	0.02	.19	5.65*	0.06	0.13
Reading Comprehension	0.66	0.44	.04	1.48	-0.21	1.52
Vocabulary	2.18	0.22	.34	9.78*	1.75	2.62
Constant	150.93	4.56		33.07*	141.98	156.89

* $p < .01$

Reference

Ketterlin-Geller, L. R., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for 3rd-grade students (Technical Report # 22).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0027: Analysis of Reading Fluency and Comprehension Measures for Fourth Grade Students (Alonzo & Tindal, 2004b).

Methods

This report documents the development and evaluation of fourth grade reading fluency and comprehension measures intended for screening and progress monitoring. Student samples were drawn from multiple instructional contexts to ensure variability in performance. Standardized administration procedures were implemented. Passage construction and alternate-form development followed structured technical specifications to ensure comparability across administrations. Classical test theory indices were calculated to evaluate internal consistency and item discrimination. Rasch modeling procedures were used to calibrate item difficulty and examine fit statistics. Dimensionality analyses examined the coherence of reading constructs, and classification accuracy statistics were computed to evaluate screening thresholds.

Results

Results indicate acceptable to strong reliability for screening and progress monitoring. Reliability coefficients met commonly accepted benchmarks for educational decision-making. Most items demonstrated appropriate difficulty and discrimination levels. Rasch modeling supported stable calibration across forms and student samples. Dimensionality analyses supported interpretation of the measures as coherent reading constructs. Alternate forms demonstrated sufficient equivalency to support growth interpretation over time. **Classification accuracy indices showed adequate sensitivity and specificity** for identifying at-risk students. Growth analyses confirmed responsiveness to developmental change. Overall findings support the technical adequacy and instructional utility of the fourth grade reading fluency and comprehension measures.

Table 24. Illustrative Table from Technical Report 0027

Table 12

Correlations Between the Grade 4 Measures, Form A of CR

		District ORF	District Voc.	District SR Rdg	District CR Rdg	State Rdg
District ORF	Pearson Correlation	1	.363**	.444**	.332**	.605**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	n	585	583	583	583	502
District Voc.	Pearson Correlation		1	.362**	.301**	.442**
	Sig. (2-tailed)		.	.000	.000	.000
	n		594	590	590	510
District SR Reading, Form A	Pearson Correlation			1	.195**	.513**
	Sig. (2-tailed)			.	.000	.000
	n			591	591	506
District CR Reading, Form A	Pearson Correlation				1	.344**
	Sig. (2-tailed)				.	.000
	n				591	506
State Reading	Pearson Correlation					1
	Sig. (2-tailed)					.
	n					511

**. Correlation is significant at the .01 level (2-tailed).

Reference

Alonzo, J., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for fourth grade students (Technical Report No. 0027).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0028: Analysis of Reading Fluency and Comprehension Measures for Fifth Grade Students (Beghetto & Tindal, 2004).

Methods

This technical report analyzes the fifth-grade District Reading Assessment, which included three primary subtests: Oral Reading Fluency (ORF), District Vocabulary, and District Reading Comprehension. The study summarized spring 2003 achievement data from 1,443 fifth-grade students across 29 schools in a mid-sized Pacific Northwest district. Three data sources were merged: District Reading Comprehension, District Vocabulary, and state reading achievement scores.

Data cleaning procedures included verification of scoring formulas, cross-checking student identifiers, and resolving duplicate records. Independent variables examined included gender, ethnicity, Special Education (SPED) status, English Language Learner (ELL) status, and school income level (coded into two regional groupings).

Form comparability analyses were conducted for ORF and Reading Comprehension alternate forms. Descriptive statistics were computed for each measure, including means and standard deviations. Item-level analyses examined item difficulty and discrimination indices. The purpose of the analysis was threefold: evaluate cross-form comparability, examine demographic performance differences, and explore abbreviation of longer forms to improve efficiency.

Results

Results indicated strong comparability between ORF alternate forms, with no statistically significant differences in form difficulty. In contrast, the Reading Comprehension forms demonstrated significant differences in difficulty between Form A and Form B, raising concerns about longitudinal interpretability across forms.

Demographic analyses revealed statistically significant differences across most subgroup comparisons. ORF scores differed significantly across gender, ethnicity, SPED, ELL status, and income level. Vocabulary results showed significant differences across all demographic variables except gender. For Reading Comprehension, Form A demonstrated significant group differences across all demographic categories except ethnicity, while Form B showed differences across all variables except school income level.
Item analyses revealed a range of difficulty levels and acceptable discrimination indices for most items. However, differences in alternate form difficulty suggested the need for further calibration to improve cross-form equivalence. Overall, the measures demonstrated adequate technical characteristics but required refinement to ensure fairness and comparability in district-level decision making.

Table 25. Illustrative Table from Technical Report 0028

Table 9
Item Analysis: CRs

Form	Item	Difficulty	Differentiation (Item-Total Correlation)
A	1	66%	.65
A	2	63%	.71
A	3	60%	.76
A	4	55%	.76
B	1	69%	.65
B	2	59%	.75
B	3	74%	.58
B	4	56%	.73

Reference

Beghetto, R., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for fifth grade students (Technical Report No. 0028).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0024: Analysis of Reading Fluency and Comprehension Measures for Sixth Grade Students (Alonzo & Tindal, 2004d).

Methods

This study examined sixth grade reading data collected during a district-wide administration of oral reading fluency and comprehension measures. Students completed timed passage readings scored as words read correctly per minute, followed by comprehension probes aligned with grade-level instructional standards. The purpose was to evaluate the technical adequacy of both measures for screening and progress monitoring in upper elementary settings. Analyses included descriptive statistics, correlation analyses between fluency and comprehension scores, and estimates of internal consistency reliability. The researchers also evaluated score distributions to identify possible ceiling or floor effects and to determine how effectively each measure differentiated among students across varying performance levels. Comparisons were conducted to determine whether fluency alone sufficiently represented overall reading proficiency at this grade level.

Results

Results indicated a meaningful but reduced relationship between oral reading fluency and comprehension compared to lower elementary grades. While students with higher fluency rates generally demonstrated stronger comprehension performance, the correlation was moderate rather than strong.

Reliability estimates for both fluency and comprehension measures were acceptable for instructional decision-making. Distributional analyses suggested that fluency measures began to show less variability among higher-performing students, whereas comprehension probes captured a broader range of performance differences.

The findings suggest that although oral reading fluency remains a useful indicator in sixth grade, comprehension measures provide additional explanatory power as texts become more complex. Together, the measures offered complementary information that strengthened the overall validity of the district's reading assessment system.

Table 26. Illustrative Table from Technical Report 0024

Table 13

Correlations Between the Grade 6 Measures

		District ORF	District Voc.	District SR Rdg	District CR Rdg	State Rdg
District ORF	Pearson Correlation	1	.29**	.47**	.40**	.54**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	n	202	202	202	202	169
District Voc.	Pearson Correlation		1	.27**	.38**	.21**
	Sig. (2-tailed)		.	.000	.000	.005
	n		221	221	221	185
District SR Reading	Pearson Correlation			1	.47**	.64**
	Sig. (2-tailed)			.	.000	.000
	n			221	221	185
District CR Reading	Pearson Correlation				1	.33**
	Sig. (2-tailed)					.000
	n					185
State Reading	Pearson Correlation					1
	Sig. (2-tailed)					.
	n					185

**. Correlation is significant at the .01 level (2-tailed).

Reference

Alonzo, J., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for sixth grade students (Technical Report No. 0024)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0023: Analysis of Reading Fluency and Comprehension Measures for Seventh Grade Students (Alonzo & Tindal, 2004c).

Methods

The study examined seventh-grade assessment data from a district-wide administration of reading measures. Students completed timed oral reading fluency passages and grade-level comprehension assessments. Fluency was measured as words read correctly per minute, while comprehension was assessed through structured response items linked to narrative and informational texts.

Analyses included correlation estimates between fluency and comprehension, internal consistency reliability calculations, and evaluation of score distributions. Performance patterns were examined across achievement levels to assess whether fluency maintained predictive utility in middle school contexts.

Results

Findings demonstrated a moderate relationship between fluency and comprehension at the seventh-grade level. Although fluency remained associated with comprehension outcomes, the relationship was weaker than typically observed in elementary grades. Reliability estimates were acceptable for educational decision-making.

Comprehension measures captured additional variance beyond fluency, particularly among higher-performing students. The results suggest that while fluency remains useful, comprehension measures play an increasingly important role in evaluating reading competence at the secondary level.

Table 27. Illustrative Table from Technical Report 0023

Table 17

Correlation Matrix for the Different District Tests

		District ORF	District Voc.	District SR Rdg	District CR Rdg
District ORF	Pearson Correlation	1	.43**	.27**	.43**
	Sig. (2-tailed)	.	.000	.000	.000
	n	141	141	141	141
District Voc.	Pearson Correlation		1	.26**	.44**
	Sig. (2-tailed)		.	.002	.000
	n		146	146	146
District SR Reading	Pearson Correlation			1	.30**
	Sig. (2-tailed)			.	.000
	n			146	146
District CR Reading	Pearson Correlation				1
	Sig. (2-tailed)				.
	n				146

**. Correlation is significant at the .01 level (2-tailed).

Reference

Alonzo, J., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for seventh grade students (Technical Report No. 0023).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1310: easyCBM® Reading Criterion Related Validity Evidence: Grades 2–5 (Lai et al., 2013a).

Methods

This technical report examined criterion-related validity evidence for the easyCBM® reading measures administered in Grades 2 through 5. The purpose of the study was to evaluate the extent to which easyCBM® vocabulary, comprehension, and fluency measures relate to established, standardized reading assessments with documented reliability and validity evidence. Participants were drawn from a convenience sample of ten schools within one Oregon district implementing easyCBM® as part of an RTI framework. Data collection occurred in January 2013 and extended to approximately six weeks due to high absenteeism. Initial recruitment included 1,017 students; final sample sizes were 204 (Grade 2), 288 (Grade 3), 184 (Grade 4), and 206 (Grade 5). No demographic information was collected. easyCBM® measures included Vocabulary, Common Core State Standards (CCSS) Reading (Grades 3–5 only), Multiple-Choice Reading Comprehension (MCRC), and Passage Reading Fluency (PRF). Comparator measures included the Gates-MacGinitie Reading Tests (Word Knowledge and Reading Comprehension subtests) and DIBELS Oral Reading Fluency (ORF). For logistical reasons, paper-pencil versions of the easyCBM® progress monitoring measures were used. Prior to analysis, assumptions of normality and linearity were evaluated. Because most measures did not meet normality assumptions, Spearman's rank-order correlations (rs) were used for vocabulary and comprehension analyses. Pearson's correlations (r) were used for fluency measures, as assumptions were met. Correlation analyses were conducted separately by grade level to examine the strength and direction of relations between easyCBM® and comparator assessments.

Results

Across grades, correlations between easyCBM® vocabulary and comprehension measures and the Gates-MacGinitie assessments ranged from low to moderate. Vocabulary correlations generally fell in the .30 to .70 range depending on grade. For example, Grade 2 easyCBM® Vocabulary correlated .76 with Gates-MacGinitie Word Knowledge, while Grade 3 correlations were more modest. CCSS and MCRC comprehension measures demonstrated correlations with Gates-MacGinitie Reading Comprehension that ranged from the .40s to the .70s across Grades 3–5, indicating

moderate criterion-related validity. In contrast, the strongest validity evidence was observed for fluency measures. Across Grades 2–5, easyCBM® Passage Reading Fluency demonstrated consistently high correlations with DIBELS ORF, with coefficients ranging from the .80s to mid-.90s. These findings indicate strong convergence between the two fluency-based measures. The report notes variability in missing data for the Gates-MacGinitie measures, particularly in Grades 3–5, due to technological and logistical constraints. Despite this limitation, overall findings suggest moderate criterion-related validity evidence for vocabulary and comprehension measures and strong validity evidence for fluency measures. The results support the use of easyCBM® reading assessments for screening and progress monitoring within Grades 2–5, particularly for oral reading fluency.

Table 28. Illustrative Table from Technical Report 1310

Table 3
Correlation Results – Grade 2

Measures		easyCBM® Vocabulary	easyCBM® MCRC	Gates-MacGinitie WK	Gates-MacGinitie RC	DIBELS ORF
easyCBM® Vocabulary	r_s	1	.56**	**.76****	.58**	-
	n	233	233	194	176	-
easyCBM® MCRC	r_s	-	1	.61**	**.66****	-
	n	-	233	194	176	-
Gates-MacGinitie WK	r_s	-	-	1	.68**	-
	n	-	-	199	181	-
Gates-MacGinitie RC	r_s	-	-	-	1	-
	n	-	-	-	181	-
easyCBM® PRF	*r*	-	-	-	-	**.95****
	n	-	-	-	-	229

Note. r_s = Spearman's rho rank correlation coefficient. *r* = Pearson's correlation coefficient. **. Correlation is significant at the 0.01 level (2-tailed). easyCBM®-comparator measure coefficients in bold-red fonts. MCRC = Multiple Choice Reading Comprehension, WK = Word Knowledge, RC = Reading Comprehension, PRF = Passage Reading Fluency.

Reference

Lai, C.-F., Alonzo, J., & Tindal, G. (2013). *easyCBM® reading criterion related validity evidence: Grades 2–5 (Technical Report No. 1310)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1103: easyCBM® Reading Criterion-Related Validity Evidence: Oregon State Test 2009–2010 (Anderson, Alonzo, et al., 2011b).

Methods

This report evaluates the criterion-related validity of easyCBM® reading assessments against Oregon's OAKS assessment for grades 3–7. Data were drawn from three Oregon districts. Measures included PRF, MCRC, Vocabulary, and grade 3 WRF. Predictive validity analyses used fall and winter easyCBM® administrations. Concurrent validity analyses used spring administrations. Multiple regression models were conducted, including full-year models and seasonal models. Correlation analyses were conducted by district and full sample. Scatterplots examined classification accuracy relative to OAKS cut scores.

Results

Full annual models explained between 67% and 73% of variance in OAKS scores, exceeding variance explained in the Washington study. At grade 3, the full model accounted for 71% of variance. Removing WRF reduced explained variance by 5%. Vocabulary and MCRC measures frequently served as the strongest unique predictors. Fall and winter models demonstrated strong predictive validity (up to 64% variance explained), while spring models showed strong concurrent validity (up to 66%). Scatterplots indicated that students below the 20th percentile on easyCBM® rarely achieved OAKS proficiency. Classification was more accurate for identifying non-proficient students than proficient students. Overall, findings demonstrate strong criterion-related validity and robust predictive relationships between easyCBM® reading measures and OAKS outcomes.

Table 29. Illustrative Table from Technical Report 1103

District 2 Correlations – Grade 3

		OAKS Rdg	Fall WRF	Fall PRF	Fall MCRC	Fall Voc	Wint WRF	Wint PRF	Wint MCRC	Spr WRF	Spr PRF	Spr MCRC	Spr Voc
OAKS Rdg	Pearson Corr.	1	.651**	.671**	.600**	.708**	.618**	.694**	.601**	.601**	.687**	.682**	.728**
	N	825	775	776	765	762	797	797	789	819	818	814	813
Fall WRF	Pearson Corr.	.651**	1	.919**	.644**	.740**	.916**	.887**	.517**	.875**	.873**	.511**	.688**
	N	775	802	801	784	786	798	798	783	796	795	784	787
Fall PRF	Pearson Corr.	.671**	.919**	1	.652**	.718**	.859**	.906**	.517**	.820**	.898**	.533**	.657**
	N	776	801	802	784	786	798	798	783	796	795	784	787
Fall MCRC	Pearson Corr.	.600**	.644**	.652**	1	.630**	.573**	.647**	.526**	.548**	.626**	.553**	.522**
	N	765	784	784	787	778	782	782	777	783	783	779	781
Fall Voc	Pearson Corr.	.708**	.740**	.718**	.630**	1	.703**	.707**	.559**	.668**	.705**	.581**	.712**
	N	762	786	786	778	787	783	783	774	783	783	774	776
Wint WRF	Pearson Corr.	.618**	.916**	.859**	.573**	.703**	1	.898**	.470**	.905**	.863**	.482**	.679**
	N	797	798	798	782	783	826	826	808	819	818	807	810
Wint PRF	Pearson Corr.	.694**	.887**	.906**	.647**	.707**	.898**	1	.522**	.850**	.913**	.549**	.677**
	N	797	798	798	782	783	826	827	808	820	819	807	810
Wint MCRC	Pearson Corr.	.601**	.517**	.517**	.526**	.559**	.470**	.522**	1	.458**	.515**	.555**	.494**
	N	789	783	783	777	774	808	808	809	808	808	802	802
Spr WRF	Pearson Corr.	.601**	.875**	.820**	.548**	.668**	.905**	.850**	.458**	1	.888**	.479**	.677**
	N	819	796	796	783	783	819	820	808	852	851	837	839
Spr PRF	Pearson Corr.	.687**	.873**	.898**	.626**	.705**	.863**	.913**	.515**	.888**	1	.567**	.670**
	N	818	795	795	783	783	818	819	808	851	851	837	839
Spr MCRC	Pearson Corr.	.682**	.511**	.533**	.553**	.581**	.482**	.549**	.555**	.479**	.567**	1	.599**
	N	814	784	784	779	774	807	807	802	837	837	839	837
Spr Voc	Pearson Corr.	.728**	.688**	.657**	.522**	.712**	.679**	.677**	.494**	.677**	.670**	.599**	1
	N	813	787	787	781	776	810	810	802	839	839	837	841

a. Cannot be computed because at least one of the variables is constant. **Significant at the 0.01 level. *Significant at the 0.05 level.

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2011). *easyCBM reading criterion-related validity evidence: Oregon state test 2009–2010 (Technical Report No. 1103)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1101: easyCBM® Reading Criterion Related Validity Evidence: Washington State Test 2009-2010 (Anderson, Alonzo, et al., 2011a).

This study examined criterion-related validity of easyCBM® reading measures using a large statewide sample of students in Washington across grades 3 through 8. Participants included thousands of students per grade who completed easyCBM assessments (Vocabulary, Multiple Choice Reading Comprehension, and Passage Reading Fluency) administered during fall, winter, and spring benchmark periods. These measures were aligned with reading constructs of comprehension, fluency, and vocabulary knowledge. Data collection procedures involved routine benchmark administrations within school settings, with scores linked to performance on the Washington Measurement of Student Progress (MSP) reading assessment.

Statistical analyses included Pearson correlations to estimate concurrent and predictive validity between easyCBM measures and MSP reading scores. Multiple regression models were also conducted to evaluate the combined predictive power of multiple easyCBM® measures within and across seasons. Model fit indices (R and R^2) and regression coefficients (both unstandardized and standardized) were reported by grade and season, allowing detailed examination of unique and shared contributions of each measure to predicting state test outcomes.

The findings provide strong evidence for the criterion-related validity of easyCBM® reading measures in predicting performance on the Washington MSP reading assessment. Across grades 3 through 8, individual easyCBM® measures demonstrated moderate to strong correlations with MSP scores, typically ranging from approximately .57 to .68 depending on grade, season, and measure. Passage Reading Fluency (PRF) and Multiple Choice Reading Comprehension (MCRC) consistently showed the strongest relationships with MSP outcomes, reflecting their close

alignment with reading comprehension constructs.
Regression analyses further indicated that combining measures improved predictive accuracy. Full seasonal models, which included multiple easyCBM® measures within a given benchmark period, produced R^2 values generally in the range of approximately .40 to .55, indicating that a substantial proportion of variance in MSP reading scores could be explained by easyCBM® performance. Models that incorporated data across seasons showed even stronger predictive utility, demonstrating the cumulative value of repeated measurement over time.

Standardized regression coefficients revealed that different measures contributed uniquely to prediction. PRF often had strong contributions in earlier grades, while comprehension measures (MCRC and Vocabulary) became increasingly important in upper grades. This developmental pattern is consistent with theoretical expectations regarding reading development, where fluency supports comprehension in early stages and comprehension processes become more dominant later.

The consistency of findings across grades and seasons supports the stability of the relationships between easyCBM® and MSP outcomes. Additionally, the relatively narrow confidence intervals around correlation coefficients indicate precision in estimates due to large sample sizes. Overall, the results demonstrate that easyCBM® measures provide valid and educationally meaningful indicators of reading proficiency as assessed by statewide accountability tests, supporting their use in screening, progress monitoring, and instructional decision-making contexts.

Table 30. Summary of Key Findings from Technical Report 1101

Grade Range	Measure	Validity Coefficients	Key Findings
Grades 3–8	PRF	.60–.68	Strong predictor of MSP reading, especially in lower grades
Grades 3–8	MCRC	.62–.68	Consistently strong relationship with comprehension outcomes
Grades 3–8	Vocabulary	.57–.65	Moderate to strong predictor; increases in importance in upper grades
Grades 3–8	Full Models	$R^2 \approx .40–.55$	Combined measures substantially improve prediction accuracy

Reference

Alonzo, J., Park, B. J., Tindal, G., & colleagues. (2011). *Technical Report 1101: easyCBM® Reading Criterion-Related Validity Evidence—Washington State Test 2009–2010*. Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1408: Technical Manual: easyCBM® (Anderson et al., 2014).

Methods

Technical evidence for easyCBM® reading measures was gathered from diverse samples spanning Grades K–8. Early literacy samples (Grades K–1) included approximately 2,000 students in the Pacific Northwest, with additional national samples from 76 schools across 26 states. Criterion validity studies for upper-grade reading measures included several hundred to several thousand students per grade, depending on the study. The 2013–2014 norming sample was drawn using a stratified random sampling procedure targeting 500 students per demographic cell (region, race-ethnicity, gender) to reflect national enrollment proportions according to the National Center for Education Statistics Common Core of Data.

The easyCBM® reading assessment system spans multiple measure types: early literacy measures (Letter Names, Letter Sounds, and Phoneme Segmenting, Grades K–1); fluency measures (Word Reading Fluency and Passage Reading Fluency, Grades K–8); and comprehension measures (Vocabulary, Multiple Choice Reading Comprehension [MCRC], and CCSS Reading, Grades 2–8). All measures include three seasonal benchmarks and up to 20–30 progress monitoring forms per grade. CCSS Reading measures assess literal comprehension across three genres: Read to Perform a Task, Informational Text, and Short Literary Text.

Measures were developed through iterative item-writing and review processes involving trained teacher panels and University of Oregon researchers. For fluency measures, piloting occurred across Pacific Northwest schools; CCSS Reading measures were piloted using online national samples. Alignment studies employed teacher raters who independently judged item-standard correspondence, with inter-rater reliability evaluated across multiple sessions. Criterion validity data were collected by administering easyCBM® benchmarks and external criterion tests (e.g., DIBELS, Stanford Achievement Test-10 [SAT-10], state tests, Oregon Assessment of Knowledge and Skills [OAKS]) within the same school year.

Rasch modeling was applied to all discrete-item measures to calibrate item difficulty and evaluate fit (Mean Square Outfit; acceptable range 0.50–1.50). Because Passage Reading Fluency measures lack discrete items, Analysis of Variance (ANOVA) was used to evaluate test form equivalence, supplemented by Flesch-Kincaid readability indices. Reliability was assessed via Pearson's r (alternate form and test-retest), Cronbach's alpha (internal consistency), and generalizability theory (G-coefficients). Criterion validity analyses used Pearson's r, Spearman's rho (for non-normal distributions), simple and multiple linear regression, and diagnostic efficiency statistics (sensitivity, specificity, AUC). Construct validity was examined through confirmatory factor analysis (CFA) using one-, two-, and three-factor models.

Results

Rasch-based item development was employed across all three early literacy measures to ensure test form equivalence. Alternate form reliability was consistently strong: for Letter Names, same-day alternate form correlations ranged from .87 to .90 for Grade K and .85 to .90 for Grade 1; test-retest correlations ranged from .79 to .82. Letter Sounds alternate form reliability ranged from .76 to .92, and generalizability theory G-coefficients ranged from .87 to .95, indicating high measurement consistency. Phoneme Segmenting showed similarly robust alternate form reliability, with G-coefficients exceeding .80 across forms.

Criterion validity for early literacy measures was demonstrated through multiple studies. Letter Names showed Spearman's rho correlations of .86 (Grade K) and .80 (Grade 1) with DIBELS Letter Naming Fluency. Combined early literacy measures (Letter Names, Letter Sounds, Phoneme Segmenting) predicted 35–58% of variance in Kindergarten SAT-10 scores and 49–56% of variance in Grade 1 SAT-10 Word Reading Fluency performance. Letter Sounds showed moderate concurrent correlations with DIBELS measures (.55 with ISF at Grade K; .58 with NWF at Grade 1). Confirmatory factor analyses supported a single latent "Reading" factor at both grade levels, with Letter Names factor loadings ranging from the .80s to .90s (Kindergarten) and .90s (Grade 1), confirming strong construct validity.

Word Reading Fluency (WRF) and Passage Reading Fluency (PRF) measures were developed with ANOVA-based equating to ensure form equivalence, supplemented by Rasch modeling for WRF. Both measures demonstrated strong alternate form and test-retest reliability, generally exceeding .80 across grades. PRF generalizability studies indicated that the passage facet (i.e., differences across passages) was the largest source of measurement variance after the student facet, which was addressed through the ANOVA-based equating process.

Criterion validity for fluency measures was strong and consistent. Predictive and concurrent correlations with external criteria (state tests, DIBELS Oral Reading Fluency, SAT-10) were generally in the .60–.85 range across grades. Fluency measures consistently accounted for substantial proportions of variance in year-end reading outcomes. Construct validity analyses placed WRF and PRF within a unidimensional reading factor at Grades K–2, and as part of a fluency factor distinct from comprehension at Grade 5, reflecting expected developmental differentiation. CFA factor loadings were consistently in the moderate-to-strong range (.50–.85).

The Multiple-Choice Reading Comprehension (MCRC) and CCSS Reading measures were developed using Rasch modeling to ensure form equivalence across all 20 progress monitoring and 3 benchmark forms per grade. Internal consistency (Cronbach's alpha) for CCSS Reading was investigated by Guerreiro et al. (2014) with strong results reported across Grades 3–8. Criterion validity studies showed that easyCBM® reading measures at Grades 2–5 had significant predictive and concurrent correlations with criterion measures (Lai, Alonzo, & Tindal, 2013), and the CCSS Reading measures demonstrated strong alignment with grade-level literacy standards.

Construct validity analyses for CCSS Reading (Grades K–5) used CFA to evaluate competing factor structures. For Grades K–2, the three-factor model (phonological awareness, fluency, comprehension) fit the data best, suggesting

that these three components of reading are distinguishable at early grades. For Grades 3 and 4, the one-factor model without testlet effects provided the best fit, indicating a general reading factor. For Grade 5, the two-factor model without a testlet effect fit best, with factors representing (a) fluency (WRF and PRF) and (b) comprehension (CCSS and MCRC). For Grades 3–5 CCSS-only models, the one-factor model also showed the best fit, with item loadings on the latent reading comprehension factor ranging from .60–.80 (Grade 3), .50–.70 (Grade 4), and .10–.70 with a median of .58 (Grade 5).

Across all reading measure types and grade levels, the easyCBM® reading assessments demonstrated strong technical adequacy. Reliability evidence was consistently acceptable to strong, with alternate form correlations, test-retest coefficients, and Cronbach's alpha values typically meeting or exceeding the .80 threshold. Criterion validity evidence was robust, with moderate-to-strong correlations across multiple external benchmarks and state assessments. Construct validity evidence supported a unidimensional reading factor at most grade levels, with developmental differentiation into subfactors (fluency vs. comprehension) emerging in upper elementary grades, consistent with theoretical models of reading development. The alignment of CCSS Reading items to literacy standards was generally strong, providing content validity support for their intended use as RTI screening and progress monitoring tools.

Table 31. Summary of Results from Technical Report 1408

Measure	Analysis Type	Metric	Result
Letter Names (K–1)	Alternate Form Reliability	Pearson's r (same-day)	.87–.90 (K); .85–.90 (Gr. 1)
Letter Names (K–1)	Criterion Validity	Spearman's ρ vs. DIBELS LNF	.86 (K); .80 (Gr. 1)
Letter Sounds (K–1)	Generalizability	G-coefficient	.87–.95
Letter Sounds (K–1)	Concurrent Validity	ρ vs. DIBELS ISF/NWF	.55 (K); .58 (Gr. 1)
Early Literacy Battery (K–1)	Predictive Validity	R^2 vs. SAT-10	.35–.58 (K); .49–.56 (Gr. 1)
Early Literacy (K–1)	Construct Validity (CFA)	Factor Loadings on Reading	.80s–.90s (K); .90s (Gr. 1)
Fluency Measures (K–8)	Reliability	Alternate Form / Test-retest r	Generally ≥.80 across grades
Fluency Measures (K–8)	Criterion Validity	r vs. State Tests / DIBELS ORF	.60–.85 across grades
CCSS Reading (K–2)	Construct Validity (CFA)	Best-fit Model	Three-factor (phonological, fluency, comprehension)
CCSS Reading (Gr. 3–4)	Construct Validity (CFA)	Best-fit Model	One-factor (general reading)
CCSS Reading (Gr. 5)	Construct Validity (CFA)	Best-fit Model	Two-factor (fluency; comprehension)
CCSS Reading (Gr. 3–5)	Construct Validity (CFA)	Item Loadings (CCSS-only model)	.60–.80 (Gr. 3); .50–.70 (Gr. 4); Mdn .58 (Gr. 5)
CCSS Reading (Gr. 3–8)	Internal Consistency	Cronbach's α	Strong across grades

Reference

Anderson, D., Alonzo, J., Tindal, G., Farley, D., Irvin, P. S., Lai, C. F., Saven, J. L., & Wray, K. A. (2014). *Technical manual: easyCBM (Technical Report # 1408).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1702: An Update to Compiled ORF Norms (Hasbrouck & Tindal, 2017).

Methods

This technical report presents an updated compilation of oral reading fluency (ORF) norms for students in Grades 1 through 6. The purpose of the study was to revise previously published compiled norms from 1992 and 2006 to reflect more current reading performance data and assessment practices. Unlike early ORF norm development, which relied on locally created passages and district-level benchmarks, this update draws upon large-scale datasets from widely used, commercially available curriculum-based measurement (CBM) systems.

Data were obtained from three ORF assessments: DIBELS 6th Edition, DIBELS Next, and easyCBM®. Rather than collecting raw student-level data directly from schools, the authors extracted percentile rank information from technical reports produced by the publishers of each assessment. These reports included norms derived from very large national samples. The total number of ORF scores represented across the three assessments exceeded 6.8 million. DIBELS datasets represented complete population samples for their respective years, while easyCBM® norming used stratified random sampling to ensure demographic representation by region, gender, and ethnicity.

A master dataset was constructed that included grade level, testing season (fall, winter, spring), percentile rank, assessment source, and words correct per minute (WCPM). For each grade, season, and percentile rank, the corresponding WCPM values from the three assessments were averaged to produce compiled norms. Consistent with earlier reports, only the 10th, 25th, 50th, 75th, and 90th percentiles were reported. The resulting norms were then compared to the 2006 compiled norms to evaluate change over time.

Results

The updated 2017 compiled ORF norms generally show modest increases in WCPM scores compared to the 2006 norms, particularly in Grades 1 through 5. In these grade levels, most percentile benchmarks increased across fall, winter, and spring administrations. Average gains across percentiles ranged from approximately 6 to 12 WCPM depending on grade, with the largest increases occurring in Grade 3. Lower percentile levels, especially the 10th and 25th percentiles, tended to show the greatest improvements, suggesting stronger performance growth among lower-performing students relative to earlier benchmarks.
Grade 6 demonstrated a different pattern. Although many percentile scores increased slightly, some remained unchanged and a few decreased marginally. Overall average gains in Grade 6 were smaller than in earlier grades. Across all six grades and percentile levels combined, the average increase from 2006 to 2017 was approximately five WCPM, a change that falls within expected developmental growth ranges reported in prior research.

The updated norms provide evidence of general stability in ORF performance across nearly three decades of reading instruction, while also reflecting incremental improvements in reading fluency outcomes. The authors conclude that compiled norms remain valuable for screening and progress monitoring because they reduce the risk of setting benchmarks too low in underperforming local contexts. These updated benchmarks allow educators to interpret student performance relative to broad, aggregated national data rather than relying solely on local comparisons.

Table 32. Illustrative Table from Technical Report 1702

%iles	Grade 1	F	W	S	Grade 2	F	W	S
90	2017		97	116	2017	111	131	148
90	2006		81	111	2006	106	125	142
	Difference		16	5	Difference	5	6	6
75	2017		59	91	2017	84	109	124
75	2006		47	82	2006	79	100	117
	Difference		12	9	Difference	5	9	7
50	2017		29	60	2017	50	84	100
50	2006		23	53	2006	51	72	89
	Difference		6	7	Difference	-1	12	11
25	2017		16	34	2017	36	59	72
25	2006		12	28	2006	25	42	61
	Difference		4	6	Difference	11	17	11
10	2017		9	18	2017	23	35	43
10	2006		6	15	2006	11	18	31
	Difference		3	3	Difference	12	17	12

%iles	Grade 3	F	W	S	Grade 4	F	W	S
90	2017	134	161	166	2017	153	168	184
90	2006	128	146	162	2006	145	166	180
	Difference	6	15	4	Difference	8	2	4
75	2017	104	137	139	2017	125	143	160
75	2006	99	120	137	2006	119	139	152
	Difference	5	17	2		6	4	8
50	2017	83	97	112	2017	94	120	133
50	2006	71	92	107	2006	94	112	123
	Difference	12	5	5	Difference	0	8	10
25	2017	59	79	91	2017	75	95	105
25	2006	44	62	78	2006	68	87	98
	Difference	15	17	13		7	8	7
10	2017	40	62	63	2017	60	71	83
10	2006	21	36	48	2006	45	61	72
	Difference	19	26	15	Difference	15	10	11

Reference

Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms (Technical Report No. 1702)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1701: Oral Reading Fluency: Outcomes from 30 Years of Research (Tindal, 2017).

Methods

This technical report synthesizes approximately 30 years of research on Oral Reading Fluency (ORF) and its role within curriculum-based measurement (CBM) and large-scale accountability systems. Rather than presenting new experimental data, the report uses a research synthesis approach, drawing on empirical studies, meta-analytic findings, and technical adequacy research conducted across multiple decades. The purpose was to examine the evolution of ORF as a measurement tool and to evaluate the evidence base supporting its reliability, validity, and instructional utility.

The report reviews foundational CBM studies from the 1980s and 1990s that established ORF as a sensitive indicator of reading competence. It then integrates subsequent research examining alternate-form reliability, growth modeling, decision rules, and criterion-related validity with high-stakes state assessments. Attention is given to methodological variations across studies, including differences in passage selection, administration time, scoring conventions (words correct per minute), and equating procedures. The synthesis also addresses criticisms of ORF, including concerns about construct representation, overemphasis on speed, and whether ORF sufficiently captures comprehension. The report evaluates empirical evidence related to these concerns by examining correlations between ORF and comprehension measures, growth trajectories across grade levels, and predictive relationships with reading outcomes.

Results

Across three decades of research, findings consistently demonstrate strong technical adequacy for ORF as a screening and progress monitoring measure. Studies reviewed in the report indicate high alternate-form reliability

and strong correlations between ORF and broader reading proficiency measures, including comprehension-based assessments. ORF scores have repeatedly shown substantial predictive relationships with state reading tests, particularly in elementary grades.

The report highlights that ORF functions as a robust general indicator of reading competence, especially in early and middle elementary grades where decoding and fluency are tightly linked to comprehension development. Growth analyses across studies indicate that ORF is sensitive to instructional change and responsive to intervention effects, making it particularly useful in multi-tiered systems of support (MTSS).
While acknowledging that ORF does not directly measure deep comprehension processes, the report concludes that it serves as a highly efficient and empirically supported proxy for overall reading proficiency. Limitations become more apparent in upper grades, where comprehension demands diversify and the strength of ORF–comprehension correlations may attenuate. Overall, the synthesis supports continued use of ORF as a technically sound, efficient, and instructionally meaningful measure within screening and progress monitoring systems.

Table 33. Illustrative Table from Technical Report 1701

Table 1

Comparison of Studies on Research Variables Referenced in Studying Growth of Reading Fluency

Authors (date)	Grades	Students	Measures	N-Measures	Slope Calculation	Averages and Growth – WCPM			
Marston, Lowry, Deno, & Mirkin (1981)	Grades (n): 1 (13), 2 (9), 3 (10), 4 (7), 5 (7), 6 (9)	Students from a small Midwest city	Third grade basal reading series: Allyn-Bacon, Ginn 720, Houghton Mifflin	Three administrations: Fall, winter, and spring	Plot of raw scores and percentage of change; significance tests of differences	Grades	Fall	Winter	Spring
						1	18.1	31.1	45.7
						2	73.2	101.1	127.8
						3	108.3	123.6	136.2
						4	125.4	131.7	155.3
						5	125.7	147.3	161.1
						6	142.9	176.7	182.8
Tindal, Germann, Marston, & Deno (1983)	Grades (n): 1 (5), 2 (13), 3 (17), 4 (22), 5 (18), 6 (21)	Students from six districts referred, assessed, and eligible for special education	Curricula in use in the school district	Three administrations: Fall, winter, and spring	Change in raw score and in discrepancy from general education	Grades	Fall	Winter	Spring
						1	6.8	14.3	9.6
						2	5.5	16.2	23.6
						3	20.5	36.6	41.3
						4	31.0	50.3	52.9
						5	59.1	72.9	79.1
						6	59.5	65.9	66.8
Tindal, Germann, & Deno (1983)	Grades (n): 1 (276) 2 (284) 3 (302) 4 (294) 5 (315) 6 (328)	Students randomly sampled from six districts in Pine County	Two passages sampled from basal reading curriculum	Three administrations: Fall, winter, and spring	Change in raw scores	Grades	Fall	Winter	Spring
						1	5	63	75
						2	35	83	93
						3	67	89	108
						4	98	111	128
						5	121	120	138
						6	123	126	134
Fuchs, Deno, & Mirkin (1984)	Grades 3-5: 64 students in (DBPM) 77 students in no DBPM	All students were 'handicapped'	3rd grade passage reading test from Ginn 720	Pre-post (unknown: sometime between Nov. and May)	Pre-post difference @ 28 weeks	Condition[a]		Pre	Post
						Experimental*		41.6	70.2
						Contrast		51.5	51.3
						DBPM versus none			

Reference

Tindal, G. (2017). *Oral reading fluency: Outcomes from 30 years of research (Technical Report No. 1701).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0030: District Reading Assessments, Spring 2004 Administration (Alonzo & Tindal, 2004e).

Methods

This report documents district-wide administration of a comprehensive K–8 reading assessment system in spring 2004, with pilot data from grade 9. The assessment system was developed collaboratively between district educators and University of Oregon researchers to meet eight design criteria including psychometric soundness, use of authentic literature, progress monitoring capability, and alignment with state tests.
Subtests varied by grade level and included Letter Naming, Letter Sounds, Phoneme Segmentation, High Frequency Words, Word Reading, Oral Reading Fluency (ORF), Multiple Choice (MC) Comprehension, Vocabulary, and Maze Comprehension. Administration was conducted by trained retired or substitute teachers. Data were captured using optical character recognition scan sheets and analyzed at the university.
Analyses included descriptive statistics (means, SDs, percentile ranks), district norm development, and Rasch-based item response theory (IRT) modeling for multiple choice and other selected response measures. Misfit statistics, item redundancy, ceiling effects, and discrimination patterns were evaluated. Recommendations were generated for item revision and form abbreviation.

Results

District norm tables were produced for each grade level. Early literacy measures showed expected developmental progression, with strong discrimination across percentile bands. ORF measures demonstrated acceptable discrimination and reliable differentiation across ability levels. Vocabulary measures at grades 3 through 6 exhibited ceiling effects, limiting sensitivity among higher-performing students. As a result, the committee recommended increasing item difficulty. MC Comprehension measures generally demonstrated acceptable Rasch model fit. However, redundant items were identified in multiple grade levels, and several poorly fitting items were recommended for removal or revision. Grade 9 measures, piloted for the first time, showed appropriate discrimination but required shortening to align with district testing time constraints. Specific items were flagged for redundancy or misfit. The district accepted recommendations to revise poorly fitting items, increase vocabulary difficulty at several grade levels, rewrite Grade 8 ORF passages to address readability imbalances, and eliminate open-ended comprehension items due to limited incremental information. Overall, the system was deemed technically sound and highly informative for screening, intervention planning, and longitudinal tracking.

Reference

Alonzo, J., & Tindal, G. (2004). *District reading assessments, spring 2004 administration (Technical Report No. 0030).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Appendix A: Technical Report Table and Figure Titles

Table 1. Grade-Level Benchmark Measures by Assessment Period

Table 2. Example Regression Predictors in Reading

Table 3. Example Variance Explained Output with easyCBM® Early Reading Measures

Table 4. Correlations Among easyCBMs® in Reading and Criterion Measures

Table 5. Illustrative Table from Technical Report 1005

Table 6. Correlations Among easyCBMs® Reading with DIBELS, CTOPP, and TOWRE

Table 7. Illustrative Table from Technical Report 0922

Table 8. Example Classification Accuracy Output for Early Reading Measures

Table 9. AUC with 40th PR on Test Forms of SAT-10

Table 10. Illustrative Table from Technical Report 1806

Table 11. Main Findings Summary for Technical Report 1807

Table 12. Example Regression Values with easyCBM® in Early Reading Measures

Table 13. Example Regression Model Predictors between easyCBM® and SAT-10

Table 14. Illustrative Table from Technical Report 0910

Table 15. Illustrative Table from Technical Report 1106

Table 16. Illustrative Table from Technical Report 1107

Table 17. Illustrative Table from Technical Report 1108

Table 18. Illustrative Table from Technical Report 1109

Table 19. Example Construct Validity Output

Table 20. Illustrative Table from Technical Report 1302

Table 21. Illustrative Table from Technical Report 1304

Table 22. Illustrative Table from Technical Report 1305

Table 23. Illustrative Table from Technical Report 0022

Table 24. Illustrative Table from Technical Report 0027

Table 25. Illustrative Table from Technical Report 0028

Table 26. Illustrative Table from Technical Report 0024

Table 27. Illustrative Table from Technical Report 0023

Table 28. Illustrative Table from Technical Report 1310

Table 29. Illustrative Table from Technical Report 1103

Table 30. Summary of Key Findings from Technical Report 1101

Table 31. Summary of Results from Technical Report 1408

Table 32. Illustrative Table from Technical Report 1702

Table 33. Illustrative Table from Technical Report 1701

Figure 1. Example Construct Validity Structure

Appendix B: Guide to Spreadsheet Technical Report Value Displays

See Riverside Insights or BRT to access exact values for TR Summaries
2603-VK8R_ValidityReadingTables.xlsx

- 1003
- 1004
- 1005
- 1309
- 0922
- 1801
- 1403
- 1805
- 1806
- 1807
- 2401
- 0910
- 1106
- 1107
- 1108
- 1109
- 1302
- 1304
- 1305
- 0025
- 0022
- 0027
- 0028
- 0024
- 0023
- 1310
- 1103
- 1101
- 1408…Not summarized in a spreadsheet
- 1702…Not summarized in a spreadsheet
- 1701…Not summarized in a spreadsheet
- 0030…Not summarized in a spreadsheet

Technical Report References

Alonzo, J., & Anderson, D. (2018a). *Classification accuracy of the easyCBM Kindergarten – Grade 2 reading measures (Technical Report # 1801)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Anderson, D. (2018b). *Supplementary report on easyCBM MCRC Measures: A Follow-up to previous technical report (Technical Report # 1807)*. Eugene OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Anderson, D. (2018c). *Supplementary report on easyCBM PRF measures: A Follow-up to previous technical reports (Technical Report #1806)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2013a). *An examination of the internal structures of the easyCBM CCSS reading measures (Technical Report # 1304)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2013b). *An examination of the internal structures of the Grades K-5 easyCBM CCSS reading measures: A Construct validity study (Technical Report # 1305)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2013c). *Examining the internal structure of the easyCBM reading measures, Grades K-5 (Technical Report # 1302)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2004a). *Analysis of reading fluency and comprehension measures for first-grade students (Technical Report # 25)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2004b). *Analysis of reading fluency and comprehension measures for fourth-grade students (Technical Report # 27)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2004c). *Analysis of reading fluency and comprehension measures for seventh-grade students (Technical Report # 23)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2004d). *Analysis of reading fluency and comprehension measures for sixth-grade students (Technical Report # 24)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2004e). *District reading assessments, spring 2004 administration (Technical Report # 30)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

American Educational Research Association, American Psychological Association, & Education, N. C. o. M. i. (2014). *Publication manual* (7th ed.). Author.

Anderson, D., Alonzo, J., & Tindal, G. (2011a). *easyCBM reading criterion related to validity evidence: Washington state test 2009-2010 (Technical Report # 1101)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2011b). *easyCBM reading criterion related validity evidence: Oregon state test 2009-2010 (technical Report # 1103)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., Tindal, G., Farley, D., Irvin, P. S., Lai, C. F., Saven, J. L., & Wray, K. A. (2014). *Technical manual: easyCBM (Technical Report # 1408)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). *Diagnostic efficiency of easyCBM reading: Washington state (Technical Report # 1107)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Beghetto, R., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for fifth grade students (Technical Report # 28)*. Behavioral Research and Teaching, University of Oregon.

Guerreiro, M., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM CCSS reading measures grades K-8 (Technical Report # 1407)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms (Technical Report # 1702)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Anderson, D., Alonzo, J., & Tindal, G. (2011). *A cross-validation of easyCBM reading cut scores in Washington: 2009-2010 (Technical Report # 1109)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Sáez, L., Anderson, D., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM Grade 2 reading measures (Technical Report # 1004)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Ketterlin-Geller, L. R., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for 3rd-grade students (Technical Report # 22)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013a). *easyCBM reading criterion related validity evidence: Grades 2-5 (Technical Report # 1310)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013b). *easyCBM reading criterion related validity evidence: Grades K-1 (Technical Report # 1309)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Nese, J. F. T., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM primary-level reading measures (grades k-1), 2009-2010 version (technical report 1003)*.

Park, B. J., Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). *Diagnostic efficiency of easyCBM reading: Oregon (Technical Report # 1106)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Anderson, D., Alonzo, J., & Tindal, G. (2011). *A cross-validation of easyCBM reading cut scores in Oregon: 2009-2010 (Technical Report # 1108)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Sáez, L., Park, B. J., Nese, J. F. T., Jamgochian, E. M., Lai, C. F., Anderson, D., Kamata, A., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM reading measures (Grades 3-7), 2009-2010 version (Technical Report # 1005)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G. (2017). *Oral reading fluency: Outcomes from 30 years of research (Technical Report # 1701)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., Nese, J. F., & Alonzo, J. (2009a). *Criterion-related evidence using easyCBM® reading measures and student demographics to predict state test performance in Grades 3-8 (Technical Report # 0910)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., Nese, J. F., & Alonzo, J. (2009b). *Hierarchical linear modeling of passage reading fluency growth as a function of student characteristics (Technical Report # 0922)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., & Nese, J. F. T. (2024). *Criterion validity and classification accuracy of easyCBM: Grades 3-8. (Technical Report No. 2401).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Wray, K., Lai, C. F., Sáez. L., Alonzo, J., & Tindal, G. (2014). *easyCBM beginning reading measures: Grades K-1 alternate form reliability and criterion validity with the SAT-10 (Technical Report # 1403)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Conclusions Supporting Claims for Criterion 2.3: Sub-scores

The interim assessment provides for valid inferences about a student's specific areas of strength and need.

This evaluation applies Criterion 2.3 (Indicators 2.3.a–2.3.d) to all sections of the EGCI_2.3 documentation and finds strong evidence of defensible reporting and interpretation for instructional decision-making.

Indicator 2.3.a: The assessment design clearly supports reporting at multiple levels of granularity. easyCBM reading measures are intentionally constructed to sample distinct components of reading (e.g., foundational skills, fluency, vocabulary, and comprehension), allowing educators to identify specific areas of strength and need. Sub-scores correspond directly to theoretically defined constructs and are embedded within a broader CBM framework that emphasizes instructional sensitivity and progress monitoring.

Indicator 2.3.b: Extensive reliability and precision evidence is provided for all reported sub-scores. Reliability estimates are documented in technical manuals and peer-reviewed research using defensible psychometric procedures appropriate for brief measures and repeated administration. Classification accuracy studies and NCII technical ratings further support the precision of sub-scores for screening purposes, demonstrating acceptable trade-offs between sensitivity and specificity across grades and measures.

Indicator 2.3.c: A substantial empirical literature shows that sub-scores represent meaningful sub-domains rather than arbitrary score partitions. Factor-analytic and structural modeling studies indicate both a general reading factor and systematic differences among sub-skills such as fluency, decoding, and comprehension processes. Research demonstrates that students with similar overall scores can exhibit distinct performance profiles, reinforcing the interpretive value of reporting sub-scores for instructional planning.

Indicator 2.3.d: The intended uses of sub-scores are clearly articulated and consistently bounded. Sub-scores are intended to support screening, progress monitoring, and hypothesis generation about instructional needs—not high-stakes diagnostic decisions in isolation. Longitudinal, predictive, and normative studies provide converging evidence that these uses are appropriate and educationally meaningful when sub-scores are interpreted within an MTSS framework and alongside complementary data.

Overall, EGCI_2.3 demonstrates strong alignment between assessment design, psychometric evidence, and intended use, supporting valid and defensible inferences from reported sub-scores.

National Center on Intensive Interventions Results on Screening with easyCBM Reading Measures
https://intensiveintervention.org/tools-charts/overview

- Tools charts display expert ratings on the technical rigor of assessments and interventions.
- Products are reviewed by an external Technical Review Committee of experts.
- The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation from either the TRC or NCII.
- Products are rated against established criteria and not compared to each other or ranked.
- Charts are updated during a call for submissions. The submission process is voluntary and reviews of all eligible submissions are posted on the chart.

Usability

Academic Screening Tools Chart

This tools chart has three tabs that include ratings on the technical rigor of the tools: (1) Classification Accuracy, (2) Technical Standards, and (3) Usability Features.

The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation. All tools that meet the criteria for review are posted on the chart, regardless of results. The chart represents all tools that were reviewed, not those that were "approved."

Last updated: October 2025. Click here for a brief summary of the new and improved tools we've released.

View Chart Resources

Print Current Chart View

Legend: Convincing evidence; Partially convincing evidence; Unconvincing evidence; Data unavailable; d Disaggregated data available

FILTER RESULTS

Subject: Reading, Mathematics

Grade: Pre-K, Elementary (K-5), Middle School (6-8), High School (9-12)

Apply Filters | Show Advanced Filters | Clear Filters

Compare Tools | Reset Chart | Classification Accuracy | Technical Standards | Usability Features

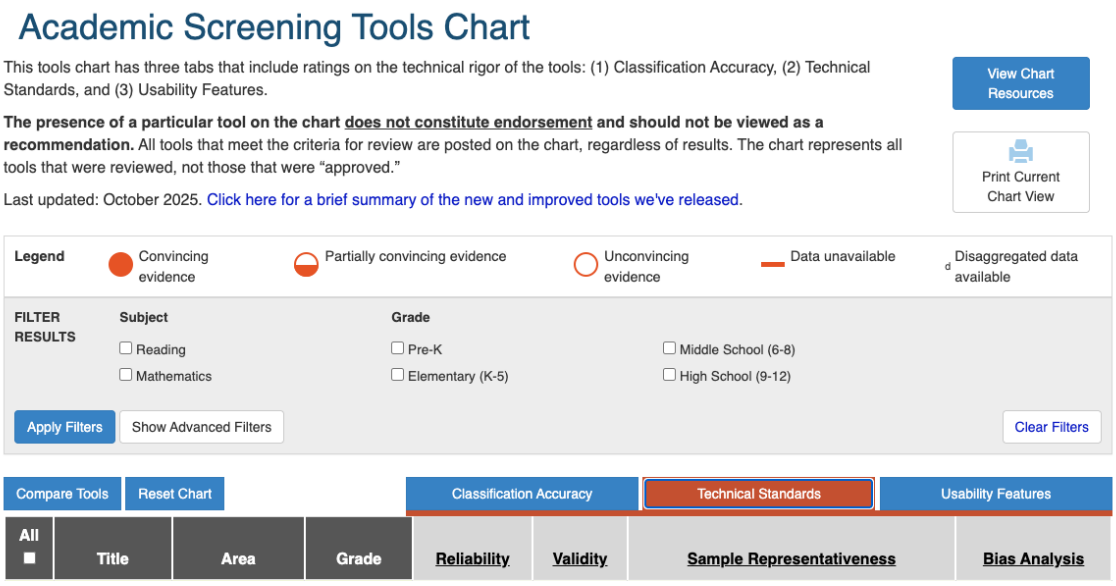
All ■	Title	Area	Grade	Admin Format	Admin & Scoring Time	Scoring Format	Types of Decision Rules	Evidence Available for Multiple Decision Rules
☐	easyCBM	Passage Reading Fluency	Grade 3	Individual	2 minutes	Automatic	Risk Levels	No
☐	easyCBM	Passage Reading Fluency	Grade 4	Individual	2 minutes	Automatic	Risk Levels	No
☐	easyCBM	Passage Reading Fluency	Grade 5	Individual	2 minutes	Automatic	Risk Levels	No
☐	easyCBM	Passage Reading Fluency	Grade 6	Individual	2 minutes	Automatic	Risk Levels	No
☐	easyCBM	Passage Reading Fluency	Grade 7	Individual	2 minutes	Automatic	Risk Levels	No
☐	easyCBM	Passage Reading Fluency	Grade 8	Individual	2 minutes	Automatic	Risk Levels	No

☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 3	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 4	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 5	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 6	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 7	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 8	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No

☐	easyCBM	Reading Composite Score	Grade 3	Individual Group Computer-administered	44 minutes	Automatic	Risk Levels	No
☐	easyCBM	Reading Composite Score	Grade 4	Individual Group Computer-administered	44 minutes	Automatic	Risk Levels	No
☐	easyCBM	Reading Composite Score	Grade 5	Individual Group Computer-administered	44 minutes	Automatic	Risk Levels	No
☐	easyCBM	Reading Composite Score	Grade 6	Individual Group Computer-administered	44 minutes	Automatic	Risk Levels	No
☐	easyCBM	Reading Composite Score	Grade 7	Individual Group Computer-administered	44 minutes	Automatic	Risk Levels	No
☐	easyCBM	Reading Composite Score	Grade 8	Individual Group Computer-administered	44 minutes	Automatic	Risk Levels	No

☐	easyCBM	Vocabulary	Grade 3	Individual Group Computer-administered	12 minutes	Automatic	Risk Levels	No
☐	easyCBM	Vocabulary	Grade 4	Individual Group Computer-administered	12 minutes	Automatic	Risk Levels	No
☐	easyCBM	Vocabulary	Grade 5	Individual Group Computer-administered	12 minutes	Automatic	Risk Levels	No
☐	easyCBM	Vocabulary	Grade 6	Individual Group Computer-administered	12 minutes	Automatic	Risk Levels	No
☐	easyCBM	Vocabulary	Grade 7	Individual Group Computer-administered	12 minutes	Automatic	Risk Levels	No
☐	easyCBM	Vocabulary	Grade 8	Individual Group Computer-administered	12 minutes	Automatic	Risk Levels	No

Technical Standards



☐	easyCBM	Passage Reading Fluency	Grade 3	●	●	Regional without Cross-Validation	Provided
☐	easyCBM	Passage Reading Fluency	Grade 4	●	●	Regional without Cross-Validation	Provided
☐	easyCBM	Passage Reading Fluency	Grade 5	●	●	Regional without Cross-Validation	Provided
☐	easyCBM	Passage Reading Fluency	Grade 6	–	○	Regional without Cross-Validation	Provided
☐	easyCBM	Passage Reading Fluency	Grade 7	–	○	Regional without Cross-Validation	Provided
☐	easyCBM	Passage Reading Fluency	Grade 8	–	○	Regional without Cross-Validation	Provided

☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 3	○[d]	◒	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 4	○[d]	●	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 5	○[d]	◒	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 6	○[d]	◒	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 7	○[d]	◒	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 8	○[d]	◒	Regional without Cross-Validation	Provided

☐	easyCBM	Reading Composite Score	Grade 3	—	●	Regional without Cross-Validation	Provided
☐	easyCBM	Reading Composite Score	Grade 4	—	●	Regional without Cross-Validation	Provided
☐	easyCBM	Reading Composite Score	Grade 5	—	●	Regional without Cross-Validation	Provided
☐	easyCBM	Reading Composite Score	Grade 6	—	●	Regional without Cross-Validation	Provided
☐	easyCBM	Reading Composite Score	Grade 7	—	●	Regional without Cross-Validation	Provided
☐	easyCBM	Reading Composite Score	Grade 8	—	●	Regional without Cross-Validation	Provided

☐	easyCBM	Vocabulary	Grade 3	○[d]	●	Regional without Cross-Validation	Provided
☐	easyCBM	Vocabulary	Grade 4	○[d]	◒	Regional without Cross-Validation	Provided
☐	easyCBM	Vocabulary	Grade 5	○[d]	◒	Regional without Cross-Validation	Provided
☐	easyCBM	Vocabulary	Grade 6	○[d]	◒	Regional without Cross-Validation	Provided
☐	easyCBM	Vocabulary	Grade 7	○[d]	●	Regional without Cross-Validation	Provided
☐	easyCBM	Vocabulary	Grade 8	○[d]	◒	Regional without Cross-Validation	Provided

Classification Accuracy

Academic Screening Tools Chart

This tools chart has three tabs that include ratings on the technical rigor of the tools: (1) Classification Accuracy, (2) Technical Standards, and (3) Usability Features.

The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation. All tools that meet the criteria for review are posted on the chart, regardless of results. The chart represents all tools that were reviewed, not those that were "approved."

Last updated: October 2025. Click here for a brief summary of the new and improved tools we've released.

View Chart Resources

Print Current Chart View

Legend ● Convincing evidence ◒ Partially convincing evidence ○ Unconvincing evidence — Data unavailable [d] Disaggregated data available

FILTER RESULTS

Subject
☐ Reading
☐ Mathematics

Grade
☐ Pre-K
☐ Elementary (K-5)
☐ Middle School (6-8)
☐ High School (9-12)

Apply Filters | Show Advanced Filters | Clear Filters

Compare Tools | Reset Chart | Classification Accuracy | Technical Standards | Usability Features

All ■	Title	Area	Grade	Classification Accuracy Fall	Classification Accuracy Winter	Classification Accuracy Spring

☐	easyCBM	Passage Reading Fluency	Grade 3	◒	◒	◒
☐	easyCBM	Passage Reading Fluency	Grade 4	◒	◒	◒
☐	easyCBM	Passage Reading Fluency	Grade 5	◒	◒	◒
☐	easyCBM	Passage Reading Fluency	Grade 6	◒	◒	◒
☐	easyCBM	Passage Reading Fluency	Grade 7	◒	◒	◒
☐	easyCBM	Passage Reading Fluency	Grade 8	◒	◒	◒

☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 3	◒	◒	◒
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 4	◒	◒	◒
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 5	◒	◒	◒
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 6	◒	◒	◒
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 7	◒	◒	◒
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 8	◒	◒	◒

☐	easyCBM	Reading Composite Score	Grade 3	●	●	●
☐	easyCBM	Reading Composite Score	Grade 4	●	●	●
☐	easyCBM	Reading Composite Score	Grade 5	●	●	●
☐	easyCBM	Reading Composite Score	Grade 6	●	◒	●
☐	easyCBM	Reading Composite Score	Grade 7	◒	●	◒
☐	easyCBM	Reading Composite Score	Grade 8	◒	◒	●

☐	easyCBM	Vocabulary	Grade 3	◒	◒	◒
☐	easyCBM	Vocabulary	Grade 4	◒	◒	◒
☐	easyCBM	Vocabulary	Grade 5	◒	◒	◒
☐	easyCBM	Vocabulary	Grade 6	◒	◒	◒
☐	easyCBM	Vocabulary	Grade 7	◒	◒	◒
☐	easyCBM	Vocabulary	Grade 8	◒	◒	◒

Published Research Literature on Screening with easyCBM Reading Measures

Inclusion criteria. Peer-reviewed journal publications that used easyCBM scores, subtests, passages, or benchmark/progress-monitoring outputs as study measures (screening, outcome, predictor, or analytic dataset).

Exclusion criteria. BRT technical reports, manuals, white papers, dissertations, conference papers, and articles that only cite easyCBM without using it as a measure.

Alonzo, J., Basaraba, D., Tindal, G., & Carriveau, R. S. (2009). They read, but how well do they understand? An empirical look at the nuances of measuring reading comprehension. *Assessment for Effective Intervention, 35(1)*, 34–44. https://doi.org/10.1177/1534508408330082

Abstract (≈100 words). Using school-based data and a set of brief comprehension measures implemented within the easyCBM system, this study explored what "reading comprehension" looks like when measured in different ways. The authors examined patterns of student performance across comprehension item types and levels of demand, highlighting how students with similar overall scores can display meaningfully different profiles. Results emphasize interpretation issues for screening and instructional decision-making—particularly the risk of over-relying on a single score when different item types tap different aspects of understanding. Implications focus on using brief comprehension data to guide instruction to better describe students' strengths and needs.

Nese, J. F. T., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment: Implications for researchers and teachers. *The Elementary School Journal, 111(4),* 608–624. https://doi.org/10.1086/659034

Abstract (≈100 words). This study examined the predictive relations between easyCBM benchmark measures and a statewide high-stakes assessment, with attention to the accuracy of classification decisions. The authors evaluated how well benchmark scores forecast later outcomes and considered how different cut scores influence sensitivity and specificity. Findings support the use of CBM benchmarking for identifying students at risk while showing that cut-score selection meaningfully changes error trade-offs and identification rates. The paper

highlights practical implications for educators using screening data in RTI/MTSS, and methodological implications for researchers studying predictive validity and decision accuracy in applied settings.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26(3)*, 349–379. https://doi.org/10.1007/s11145-012-9372-9

Abstract (≈100 words). This study tested whether reading comprehension items designed to reflect literal, inferential, and evaluative understanding form separable dimensions when used in a school-oriented assessment context associated with easyCBM comprehension measures. The authors compared competing measurement models that included a general comprehension factor with and without specific factors for comprehension type. Results indicated systematic differences in item functioning and provided evidence relevant to whether subdomains can be interpreted meaningfully beyond an overall score. Findings inform comprehension assessment design and score interpretation, especially for educators and researchers who want to use brief comprehension measures for screening or instructional planning.

Sáez, L., Nese, J. F. T., Alonzo, J., & Tindal, G. (2016). Individual differences in kindergarten through grade 2 fluency relations. *Learning and Individual Differences, 49*, 100–109. https://doi.org/10.1016/j.lindif.2016.05.020

Abstract (≈100 words). Using longitudinal easyCBM fluency data from kindergarten through Grade 2, the authors modeled relations among early fluency skills (e.g., letter sounds, word reading) and later passage reading fluency. Students were grouped by end-of-Grade-2 passage reading fluency levels (low, average, high), and the study examined how initial status and growth on precursor skills were associated with later outcomes for each group. Results highlight that the developmental pathways to later passage reading fluency differ across performance groups, underscoring the importance of early screening and targeted intervention. Implications emphasize interpreting early fluency patterns in ways that support differentiated instruction.

Alonzo, J. (2016). The relation between easyCBM and Smarter Balanced reading and mathematics assessments. *Journal of School Administration Research and Development, 1(1)*, 17–35. https://doi.org/10.32674/jsard.v1i1.1906

Abstract (≈100 words). This study investigated relations between easyCBM benchmark assessments (reading and mathematics) and the Smarter Balanced summative assessments in grades 3–8 using district data. Reported correlations indicated strong associations for mathematics and moderate associations for reading across grades and benchmark seasons. Regression analyses showed that easyCBM measures explained substantial variance in Smarter Balanced total scores, supporting predictive validity for screening and instructional decision-making. The paper discusses how benchmark assessment results may be used to anticipate end-of-year performance and to support system-level planning. Findings are presented as evidence for the practical utility of easyCBM within standards-aligned assessment systems.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education (International Scholarly Research Network)*, Volume 2013, Article ID 958530, 29 pages. http://dx.doi.org/10.1155/2013/958530

Abstract (≈100 words). Tindal (2013) traces the historical development of curriculum-based measurement (CBM) from its roots in the 1970s to contemporary applications. The article documents how CBM evolved as a practical, technically sound alternative to traditional norm-referenced testing, emphasizing frequent measurement, standardized administration, and sensitivity to instructional change. Key milestones include advances in oral reading fluency, math computation, and written expression, as well as improvements in reliability, validity, and decision rules. Tindal highlights CBM's role in data-based decision making, progress monitoring, and response-to-intervention frameworks, concluding that CBM has become a foundational assessment approach linking instruction, assessment, and accountability.

Other publications that support defenisble scores:

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (2017). *The Twentieth Mental Measurements Yearbook*. Buros Center for Testing, University of Nebraska.

Anderson, D., Alonzo, J., Tindal, G., Farley, D., Irvin, P. S., Lai, C. F., Saven, J. L., & Wray, K. A. (2014). *Technical manual: easyCBM (Technical Report # 1408)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Research on Oral Reading Fluency and Passage Reading Fluency Normative Performance

Tindal, G. (2017). *Oral reading fluency: Outcomes from 30 Years of research* (Technical Report # 1701). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Study1. Hasbrouck & Tindal (1992). Initial ORF/PRF norms in Grades 2–5. Curriculum-Based Oral Reading Fluency Norms for Students in Grades 2 through 5 published in *Teaching Exceptional Children*. Volume 24, Issue 3

Study2. Technical Report # 33 (2005) is BRT's widely cited seasonal ORF/PRF norm set (Grades 1–8). ERIC. Compares 1992 and 2005.

3. Hasbrouck & Tindal (2006) is the classic peer-reviewed ORF/ORF compilation used in practice (Grades 1–8): *Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers*, *The Reading Teacher*.

4. Technical Report #1409 (2014) provides **easyCBM** ORF/PRF seasonal norms (by grade, with national and regional percentiles). ERIC+

5. Study3_Technical Report # 1702 (2017) **updates** ORF/PRF norms using very large multi-vendor datasets in Grades 1–6. ERIC

6. easyCBM PRF Norms (2020) provide Riverside's published "Detailed Percentile Lookup Tables" for **ALL Measures** with a PDF that also includes seasonal PRF percentiles by grade. lficheweb1.d91.k12.id.us+1

7. Technical Report # 2508 (2025) provides updated normative data for easyCBM **All Measures** with ORF/PRF (and other measures) in Grades K-8, including seasonal benchmark norms. brtprojects.org+1.

8. Oral Reading Fluency Norms (2023) is a blog that provides a summary of publications on norms for ORF/PRF. https://brtapps.com/oral-reading-fluency-norms/

Conclusions Supporting Claims for Criterion 2.4: Student Progress

The interim assessment provides valid information regarding student progress in the content domain.

This evaluation applies Criterion 2.4 (Indicators 2.4.a–2.4.d) to all sections of the EGCI_2.4 documentation and concludes that the assessment system provides strong, defensible evidence for measuring and interpreting student progress in reading.

Indicator 2.4.a: The assessment is explicitly designed to support growth measurement. The easyCBM system employs vertically coherent content, stable constructs across grades, and consistent scaling, enabling meaningful within-year and across-year comparisons. Both benchmark and progress-monitoring designs support repeated measurement, and the reported score scales are appropriate for capturing change over time. Design features reflect long-standing CBM principles emphasizing sensitivity to instructional effects.

Indicator 2.4.b: Reliability and precision of growth estimates are well documented. Peer-reviewed studies demonstrate appropriate procedures for estimating standard errors of growth, including evaluations of slope reliability under varying measurement densities. Evidence shows that growth reliability improves with increased numbers of observations, particularly for students receiving special education services, and that the precision of growth estimates varies systematically across the ability distribution—findings that are explicitly acknowledged and incorporated into guidance for use.

Indicator 2.4.c: The assessment system provides clear documentation of growth-calculation procedures and by empirical evidence validating intended interpretations. Research comparing linear and nonlinear growth models illustrates how analytic choices influence growth estimates and supports the appropriateness of commonly used models for CBM data. When potential discontinuities could arise (e.g., benchmark intervals, design changes), the documentation emphasizes the need for empirical checks, preserving interpretive continuity. Multiple studies demonstrate that observed growth corresponds meaningfully to instructional exposure and intervention effects.

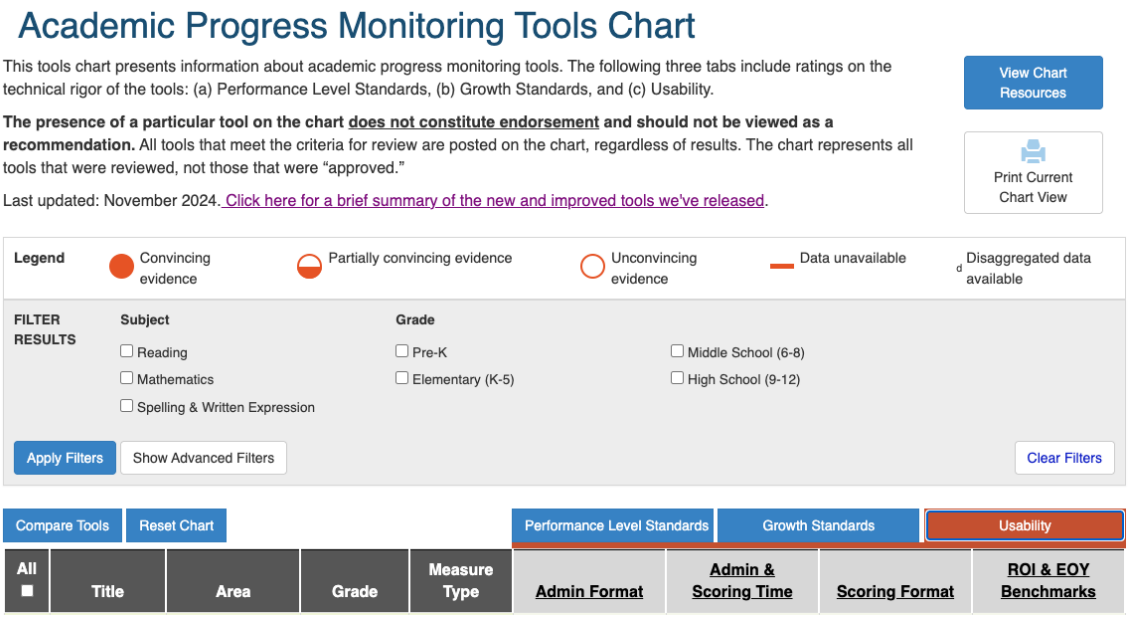
Indicator 2.4.d: The intended uses of growth scores are clearly bounded and consistently articulated: progress monitoring, goal setting, instructional adjustment, and evaluation of intervention response within MTSS frameworks. Growth scores are not positioned for high-stakes decisions in isolation. Converging theoretical and empirical evidence supports these uses across general and special education populations.

Overall, EGCI_2.4 demonstrates strong alignment between assessment design, psychometric evidence, and intended use, supporting valid and reliable inferences about student progress.

National Center on Intensive Interventions Results on Screening with easyCBM Reading Measures
https://intensiveintervention.org/tools-charts/overview

- Tools charts display expert ratings on the technical rigor of assessments and interventions.
- Products are reviewed by an external Technical Review Committee of experts.
- The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation from either the TRC or NCII.
- Products are rated against established criteria and not compared to each other or ranked.
- Charts are updated during a call for submissions. The submission process is voluntary and reviews of all eligible submissions are posted on the chart.

Usability

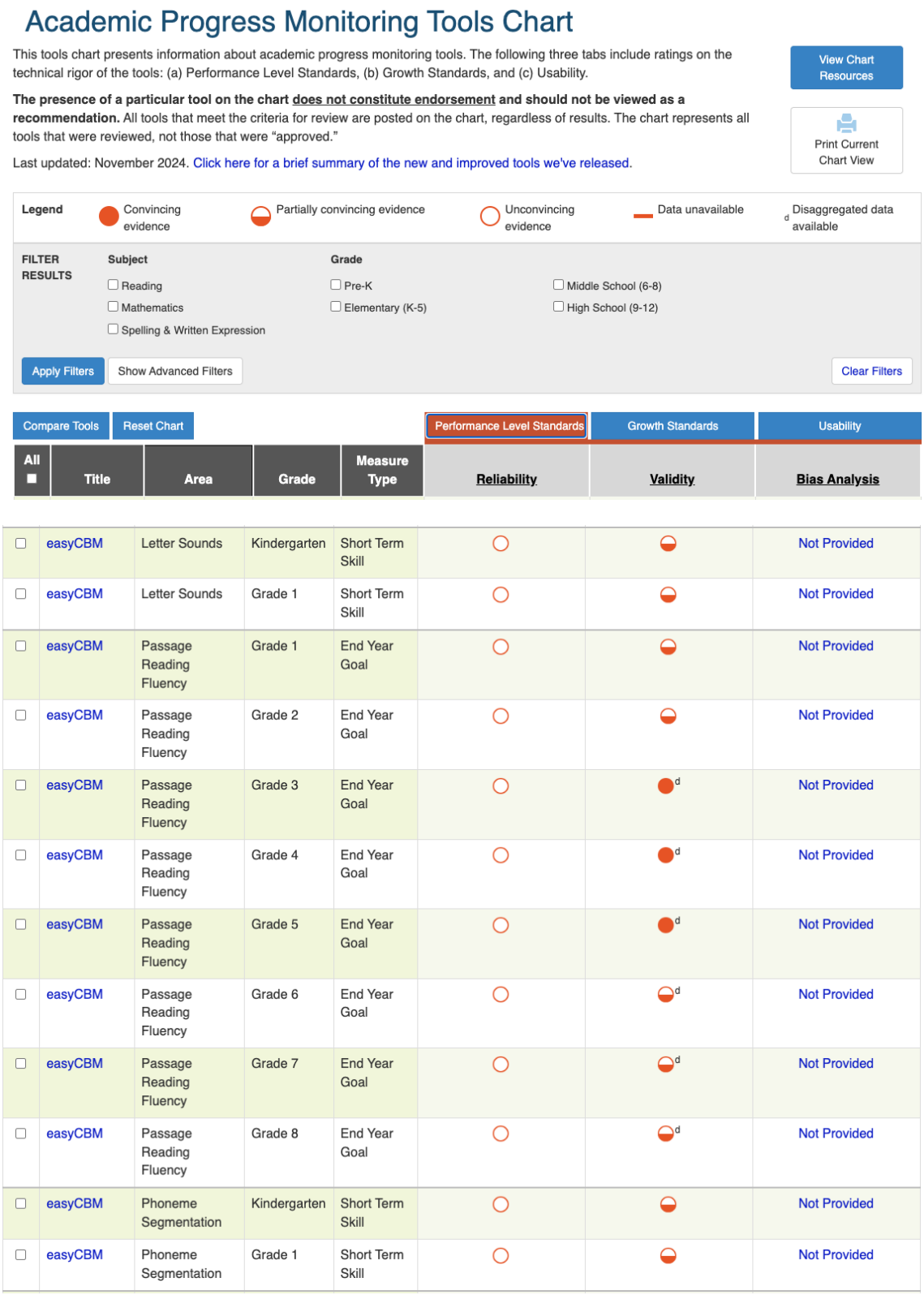


☐	easyCBM	Letter Sounds	Kindergarten	Short Term Skill	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Letter Sounds	Grade 1	Short Term Skill	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 1	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 2	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 3	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 4	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 5	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 6	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 7	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Passage Reading Fluency	Grade 8	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Phoneme Segmentation	Kindergarten	Short Term Skill	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Phoneme Segmentation	Grade 1	Short Term Skill	Individual Other	2 minutes	Manual Automatic	Benchmarks Available

☐	easyCBM	Word Reading Fluency	Grade 1	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Word Reading Fluency	Grade 2	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Word Reading Fluency	Grade 3	End Year Goal	Individual Other	2 minutes	Manual Automatic	Benchmarks Available

☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 3	End Year Goal	Individual Group Computer-administered Other	30 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 4	End Year Goal	Individual Group Computer-administered Other	30 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 5	End Year Goal	Individual Group Computer-administered Other	30 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 6	End Year Goal	Individual Group Computer-administered Other	30 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 7	End Year Goal	Individual Group Computer-administered Other	30 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 8	End Year Goal	Individual Group Computer-administered Other	30 minutes	Manual Automatic	Benchmarks Available
☐	easyCBM	Vocabulary	Grade 3	End Year Goal	Individual Group Computer-administered Other	15 minutes	Automatic	Benchmarks Available
☐	easyCBM	Vocabulary	Grade 4	End Year Goal	Individual Group Computer-administered Other	15 minutes	Automatic	Benchmarks Available
☐	easyCBM	Vocabulary	Grade 5	End Year Goal	Individual Group Computer-administered Other	15 minutes	Automatic	Benchmarks Available
☐	easyCBM	Vocabulary	Grade 6	End Year Goal	Individual Group Computer-administered Other	15 minutes	Automatic	Benchmarks Available
☐	easyCBM	Vocabulary	Grade 7	End Year Goal	Individual Group Computer-administered Other	15 minutes	Automatic	Benchmarks Available
☐	easyCBM	Vocabulary	Grade 8	End Year Goal	Individual Group Computer-administered Other	15 minutes	Automatic	Benchmarks Available

Performance Level Standards



Academic Progress Monitoring Tools Chart

This tools chart presents information about academic progress monitoring tools. The following three tabs include ratings on the technical rigor of the tools: (a) Performance Level Standards, (b) Growth Standards, and (c) Usability.

View Chart Resources

The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation. All tools that meet the criteria for review are posted on the chart, regardless of results. The chart represents all tools that were reviewed, not those that were "approved."

Print Current Chart View

Last updated: November 2024. Click here for a brief summary of the new and improved tools we've released.

Legend Convincing evidence | Partially convincing evidence | Unconvincing evidence | Data unavailable | d Disaggregated data available

FILTER RESULTS

Subject
- Reading
- Mathematics
- Spelling & Written Expression

Grade
- Pre-K
- Elementary (K-5)
- Middle School (6-8)
- High School (9-12)

Apply Filters | Show Advanced Filters | Clear Filters

Compare Tools | Reset Chart | Performance Level Standards | Growth Standards | Usability

All	Title	Area	Grade	Measure Type	Reliability	Validity	Bias Analysis
☐	easyCBM	Letter Sounds	Kindergarten	Short Term Skill	○	◒	Not Provided
☐	easyCBM	Letter Sounds	Grade 1	Short Term Skill	○	◒	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 1	End Year Goal	○	◒	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 2	End Year Goal	○	◒	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 3	End Year Goal	○	●d	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 4	End Year Goal	○	●d	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 5	End Year Goal	○	●d	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 6	End Year Goal	○	◒d	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 7	End Year Goal	○	◒d	Not Provided
☐	easyCBM	Passage Reading Fluency	Grade 8	End Year Goal	○	◒d	Not Provided
☐	easyCBM	Phoneme Segmentation	Kindergarten	Short Term Skill	○	◒	Not Provided
☐	easyCBM	Phoneme Segmentation	Grade 1	Short Term Skill	○	◒	Not Provided

☐	easyCBM	Word Reading Fluency	Grade 1	End Year Goal	○	○	Not Provided
☐	easyCBM	Word Reading Fluency	Grade 2	End Year Goal	○	◒	Not Provided
☐	easyCBM	Word Reading Fluency	Grade 3	End Year Goal	○	◒	Not Provided

☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 3	End Year Goal	○[d]	◒[d]	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 4	End Year Goal	○[d]	●[d]	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 5	End Year Goal	○[d]	●[d]	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 6	End Year Goal	○[d]	◒[d]	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 7	End Year Goal	○[d]	◒[d]	Provided
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 8	End Year Goal	○[d]	◒[d]	Provided

☐	easyCBM	Vocabulary	Grade 3	End Year Goal	○[d]	●	Provided
☐	easyCBM	Vocabulary	Grade 4	End Year Goal	○[d]	◒	Provided
☐	easyCBM	Vocabulary	Grade 5	End Year Goal	○[d]	◒	Provided
☐	easyCBM	Vocabulary	Grade 6	End Year Goal	○[d]	◒	Provided
☐	easyCBM	Vocabulary	Grade 7	End Year Goal	○[d]	●	Provided
☐	easyCBM	Vocabulary	Grade 8	End Year Goal	○[d]	◒	Provided

Growth Standards

Academic Progress Monitoring Tools Chart

This tools chart presents information about academic progress monitoring tools. The following three tabs include ratings on the technical rigor of the tools: (a) Performance Level Standards, (b) Growth Standards, and (c) Usability.

The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation. All tools that meet the criteria for review are posted on the chart, regardless of results. The chart represents all tools that were reviewed, not those that were "approved."

Last updated: November 2024. Click here for a brief summary of the new and improved tools we've released.

View Chart Resources

Print Current Chart View

Legend Convincing evidence | Partially convincing evidence | Unconvincing evidence | Data unavailable | d Disaggregated data available

FILTER RESULTS

Subject: Reading; Mathematics; Spelling & Written Expression

Grade: Pre-K; Elementary (K-5); Middle School (6-8); High School (9-12)

Apply Filters | Show Advanced Filters | Clear Filters

Compare Tools | Reset Chart | Performance Level Standards | Growth Standards | Usability

All ■	Title	Area	Grade	Measure Type	Sensitivity: Reliability of Slope	Sensitivity: Validity of Slope	Alternate Forms	Decision Rules: Setting & Revising Goals	Decision Rules: Changing Instruction
☐	easyCBM	Letter Sounds	Kindergarten	Short Term Skill	○	○	○	—	—
☐	easyCBM	Letter Sounds	Grade 1	Short Term Skill	○	○	○	—	—
☐	easyCBM	Passage Reading Fluency	Grade 1	End Year Goal	○	—	—	—	—
☐	easyCBM	Passage Reading Fluency	Grade 2	End Year Goal	○	—	—	—	—
☐	easyCBM	Passage Reading Fluency	Grade 3	End Year Goal	○^{d}	○^{d}	—	—	—
☐	easyCBM	Passage Reading Fluency	Grade 4	End Year Goal	○^{d}	○^{d}	—	—	—
☐	easyCBM	Passage Reading Fluency	Grade 5	End Year Goal	○^{d}	○^{d}	—	—	—
☐	easyCBM	Passage Reading Fluency	Grade 6	End Year Goal	○^{d}	○^{d}	—	—	—
☐	easyCBM	Passage Reading Fluency	Grade 7	End Year Goal	○^{d}	○^{d}	—	—	—
☐	easyCBM	Passage Reading Fluency	Grade 8	End Year Goal	○^{d}	○^{d}	—	—	—
☐	easyCBM	Phoneme Segmentation	Kindergarten	Short Term Skill	○	○	○	—	—
☐	easyCBM	Phoneme Segmentation	Grade 1	Short Term Skill	—	—	○	—	—

☐	easyCBM	Vocabulary	Grade 3	End Year Goal	○	—	●	—	—
☐	easyCBM	Vocabulary	Grade 4	End Year Goal	○	—	●	—	—
☐	easyCBM	Vocabulary	Grade 5	End Year Goal	○	—	●	—	—
☐	easyCBM	Vocabulary	Grade 6	End Year Goal	○	—	●	—	—
☐	easyCBM	Vocabulary	Grade 7	End Year Goal	○	—	●	—	—
☐	easyCBM	Vocabulary	Grade 8	End Year Goal	○	—	●	—	—
☐	easyCBM	Word Reading Fluency	Grade 1	End Year Goal	○	○	○	—	—
☐	easyCBM	Word Reading Fluency	Grade 2	End Year Goal	○	○	○	—	—
☐	easyCBM	Word Reading Fluency	Grade 3	End Year Goal	○	○[d]	○	—	—

☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 3	End Year Goal	—[d]	○[d]	○	—	—
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 4	End Year Goal	—[d]	●[d]	○	—	—
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 5	End Year Goal	—[d]	●[d]	○	—	—
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 6	End Year Goal	—[d]	●[d]	○	—	—
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 7	End Year Goal	—[d]	●[d]	○	—	—
☐	easyCBM	Proficient Reading (formerly Reading Comprehension)	Grade 8	End Year Goal	—[d]	●	○	—	—

Published Research Literature on Progress Monitoring with easyCBM Reading Measures

Inclusion criteria. Peer-reviewed journal publications that used easyCBM scores, subtests, passages, or benchmark/progress-monitoring outputs as study measures (screening, outcome, predictor, or analytic dataset). **Exclusion criteria.** BRT technical reports, manuals, white papers, dissertations, conference papers, and articles that only cite easyCBM without using it as a measure.

Nese, J. F. T., Biancarosa, G., Anderson, D., Lai, C.-F., & Tindal, G. (2012). Within-year oral reading fluency with CBM: A comparison of models. *Reading and Writing, 25*, 887–915. https://doi.org/10.1007/s11145-011-9304-0

Abstract (≈100 words). Using repeated oral reading fluency (ORF) administrations collected within a school year (CBM context including easyCBM-type procedures), this paper compared alternative growth models used to describe student change. The authors evaluated how model choice (e.g., linear and nonlinear representations) affects estimated growth and the interpretation of individual differences in progress. Results demonstrate that different analytic models can yield meaningfully different conclusions about rate of improvement and student trajectories. The study offers guidance for researchers and practitioners on selecting growth models that align with observed ORF patterns and intended decision uses (e.g., goal setting and progress monitoring).

Kamata, A., Nese, J. F. T., Patarapichayatham, C., & Lai, C.-F. (2013). Modeling nonlinear growth with three data points: Illustration with benchmarking data. *Assessment for Effective Intervention, 38(2),* 105–116. https://doi.org/10.1177/1534508412457872

Abstract (≈100 words). This methods paper demonstrated approaches for modeling nonlinear growth when only three benchmark occasions (fall, winter, spring) are available—an analysis situation common in easyCBM benchmarking datasets. Using passage reading fluency benchmarking data as an illustration, the authors showed how piecewise growth mixture modeling can capture potential nonlinearity and heterogeneity of growth patterns across latent classes. The paper provides a practical technical overview of the limitations of three-time-point growth models, strategies for parameterization, and interpretation of class-based results. Implications focus on improving inferences about student growth and identifying subgroups with distinct trajectories.

Tindal, G., Nese, J. F. T., Stevens, J., & Alonzo, J. (2016). Growth on oral reading fluency measures as a function of special education and measurement sufficiency. *Remedial and Special Education, 1-13*. https://doi.org/10.1177/0741932515590234

Abstract (≈100 words). Tindal, Nese, Stevens, and Alonzo (2016) examined growth in oral reading fluency (ORF) as a function of special education status and measurement sufficiency. Using longitudinal CBM data, the authors compared growth trajectories for students with and without disabilities across multiple administrations. Results showed that students receiving special education demonstrated slower ORF growth on average, but growth estimates were highly sensitive to the number and spacing of measurements. Reliability of slope estimates increased substantially with more data points, particularly for students with disabilities. The study underscores the importance of sufficient progress-monitoring data to support valid instructional and special education decision making.

Nese, J. F. T., Farley, D., & Anderson, D. (2019). Educator-reported instructional characteristics of Grade 1 reading interventions within a CBM assessment system. *Learning Disabilities Research & Practice, 34(2)*, 97–109. https://doi.org/10.1111/ldrp.12191

Abstract (≈100 words). Using data from the 2014–2015 easyCBM assessment system, this study described the characteristics of reading interventions implemented for Grade 1 students within districts using an MTSS framework. The authors summarized the number of interventions, progress-monitoring assessments, intervention start dates, curricula and instructional strategies, tier placement, group size, frequency, dosage, total instructional time, and quantitative intensity. Results provide a field-based portrait of how interventions are actually delivered when linked to a CBM assessment platform, highlighting variability in intensity and implementation patterns. Implications address how schools might use assessment-system data to monitor intervention delivery and align supports with student need.

Coggins, J. V. (2023). Using Readable English leads to reading gains for rural elementary students: An experimental study. *PLOS ONE, 18(7)*, e0288292. https://doi.org/10.1371/journal.pone.0288292

Abstract (≈100 words). In a rural elementary sample, this experimental study evaluated the impact of Readable English instruction on reading outcomes, using easyCBM measures (e.g., benchmark reading comprehension and passage reading fluency) as primary outcomes alongside standardized reading tests for some students. The analysis compared intervention and control groups on gains across the school year. Results reported advantages for the intervention group on easyCBM fluency and comprehension indices, with corresponding improvement on standardized outcomes for students who received additional testing. The paper interprets findings as evidence that structured, fluency-supportive instruction can produce measurable gains in reading and discusses implications for schools serving rural communities.

Elliott, T. C. C., Gadaire, D. M., & McKevitt, B. C. (2023). Using direct and indirect functional assessments to guide the treatment of escape-maintained problem behavior during academic tasks: A case example. *Behavior Modification. Advance online publication*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9979893/

Abstract (≈100 words). This case example used easyCBM proficient reading passages as part of an academic-task context while evaluating assessment-informed treatment for escape-maintained problem behavior. The authors compared direct and indirect functional assessment information to guide intervention design during reading tasks and documented behavioral and instructional outcomes across phases. easyCBM materials provided standardized grade-level reading content to structure and replicate academic demands during assessment and treatment sessions. The report illustrates how standardized academic materials can be incorporated into behavior-analytic intervention planning, while highlighting the practical role of selecting consistent instructional stimuli when evaluating function-based supports.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education (International Scholarly Research Network)*, Volume 2013, Article ID 958530, 29 pages. http://dx.doi.org/10.1155/2013/958530

Abstract (≈100 words). Tindal (2013) traces the historical development of curriculum-based measurement (CBM) from its roots in the 1970s to contemporary applications. The article documents how CBM evolved as a practical, technically sound alternative to traditional norm-referenced testing, emphasizing frequent measurement, standardized administration, and sensitivity to instructional change. Key milestones include advances in oral reading fluency, math computation, and written expression, as well as improvements in reliability, validity, and decision rules. Tindal highlights CBM's role in data-based decision making, progress monitoring, and response-to-intervention frameworks, concluding that CBM has become a foundational assessment approach linking instruction, assessment, and accountability.

Conclusions Supporting Claims for Criterion 3.1: Overall Achievement

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of overall achievement results.

This evaluation applies Criterion 3.1 (Indicators 3.1.a–3.1.c) to all sections of the EGCI_3.1 documentation and finds strong evidence that easyCBM® score reports and supporting materials appropriately facilitate valid interpretations and uses of overall achievement results.

Indicator 3.1.a: The design of score reports is well aligned with intended users and purposes. Reports support educators, administrators, parents, and students through differentiated formats and appropriate levels of detail. Overall achievement is clearly represented through benchmark scores, composites, and distributional displays that align with screening, grouping, and instructional planning decisions. The documentation demonstrates explicit attention to audience needs and includes cautions regarding potential misuses, such as overgeneralizing from a single score or ignoring method effects. Reports and manuals also articulate conditions under which interpretations may be compromised, prompting follow-up analyses when score patterns are divergent or inconsistent.

Indicator 3.1.b: The system provides defensible information about measurement error and score variability. Overall achievement interpretations are grounded in a Multi-Skill, Multi-Method (MS-MM) framework that triangulates performance across measures and response formats, reducing overreliance on any single indicator. Norm-referenced percentile ranks and performance distributions further contextualize variability, while documentation explains how convergent and divergent score patterns should be interpreted cautiously. These features collectively support accurate understanding of uncertainty, even when numeric confidence intervals are not displayed directly.

Indicator 3.1.c: Extensive guidance is provided to support appropriate use of overall achievement results. Interpretive guidance is firmly grounded in CBM theory, national testing standards, and educator-informed practice. Teachers are guided in using percentile ranks, risk cut points, and performance distributions to support decisions across the full performance continuum, from enrichment to intensive intervention. The integration of overall achievement with subsequent sub-score analysis and progress monitoring reinforces coherent, defensible decision making within MTSS frameworks.

Overall, EGCI_3.1 demonstrates strong coherence among report design, interpretive guidance, and intended instructional uses, supporting valid, transparent, and responsible interpretation of overall achievement.

3.1a Design of Score Reports to Document Overall Achievement in Reading, Spanish, and Math – Grades K-8

We have expanded easyCBM® to serve as an **interim assessment**, while maintaining its utility as an integral part of a Response to Intervention (RTI) or Multi-Tiered System of Supports (MTSS) model, with the primary goal of helping to facilitate data-driven instructional decision making through enhanced reporting options. The software platform is an online system that provides benchmark (BM) and progress monitoring (PM) assessments and reports in the areas of reading, Spanish literacy, and math.

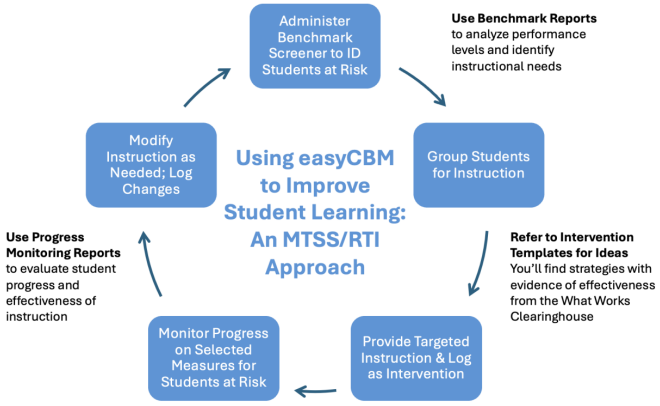
Table 1. Usage and Reporting for easyCBM® (User's Manual, page 4)

Features of easyCBM District Edition

Usage	Reporting
• Usable in an RTI framework or as a formative assessment system. • Offers online administration of all measures, including audio and Spanish language support for math. For individually administered measures, teachers can enter student responses online while administering the test to the student. • Assessments can be taken in computer labs, on laptops, and via tablet devices. • Provides access to students at the district, building, teacher, and group levels. Teachers can create their own instructional groups. • Multiple training options, including free online training on test administration and scoring. • Centralizes uploading of student and staff information. • Single sign-on and student rostering via Clever and ClassLink. • Offers ease-of-use for other key staff who require access to the data needed for data-team meetings, student support team meetings, IEP meetings, or other data-oriented meetings.	• Provides benchmark and progress monitoring reports in reading and mathematics for Kindergarten through Grade 8. • Delivers individual printer-friendly PDF Reports ideal for sharing with parents. • Allows teachers to enter intervention information and individual goals for students, which will then appear on student graphs. • Provides group graphs for whole-class performance. • Organizes sortable student rosters and provides customizable color-coding to indicate 'risk level' after each benchmark assessment. • Provides student Lexile measures based on performance on the easyCBM Proficient Reading measure for Grades 2 to 8. • Benchmark reading assessment calculates a composite score across three different measures to provide a better overall indication of student performance.

As an interim assessment, two primary features are highlighted: (a) measurement at three points-in-time (using scaled benchmark tests and (b) even more frequent follow up testing between these benchmark tests (using equivalently scaled forms of the benchmark tests. The purpose of benchmark testing, or universal screening, is to provide information regarding students' progress toward meeting end-of-year grade-level expectations and to determine which students may require intervention or enrichment. Benchmark tests are given on grade-level material and are administered three times per year: Fall, Winter, and Spring. Benchmark tests are designed to assess students' performance on the highest-priority instructional targets in reading and mathematics for each grade level. When used as an interim assessment, the focus is on systems evaluation and overall accountability with these benchmark tests. When used for identification of students at risk of learning problems (as part of MTSS), further subtests are administered to complement outcomes from the benchmark tests.

Figure 1: Using easyCBM® to Improve Student Learning (User's Manual, page 6)



3.1b Bridging Overall Achievement with Sub Scores and Progress Monitoring (Advance Organizer for 3.3 and 3.4)

Two types of overall achievement can be documented with easyCBM®: The first is a single score based on a total number of items. This type of outcome is consistent with traditional large scale testing programs. With easyCBM®, however, we also consider overall achievement in specific skill areas that allows a more sensitive bridge to interpretations needed to monitor progress and adjust instruction. For this latter interpretation, the construct of overall achievement is based on Campbell and Fiske's (1959) multi-trait, multi-method analysis (MT-MM)[1] which we describe as Multiple Skills-Multiple Methods (MS-MM). By focusing on multiple **skills** with multiple **methods** used for assessing proficiency levels, easyCBM® provides a robust and comprehensive operationalization of the constructs labeled 'Reading' and Spanish'. As a further articulation of this approach, easyCBM® provides teachers information on the convergence and divergence of various skills. In the end, this section serves as an introduction to the subsequent sections on sub scores and progress monitoring.

Overall Achievement Scores

easyCBM® provides teachers, parents, and administrators an overall reading score for students in Grades Kindergarten through 8 and in Spanish (GradeK through 3). Many different reading measures exist over the entire grade span of Kindergarten through Grade 8 using both production and selection responses (assessment methods) to ensure adequate construct representation. Benchmarks [BM] facilitates several educational functions for all students, allowing teachers to organizing students into intervention groups and analyze specific skill areas by charting item responses. Usage follows a systematic process of documenting Benchmark (BM) performance, organizing students (in classrooms and Tiers, developing interventions, monitoring progress, adjusting the interventions as needed, and benchmarking students again (Fall to Winter and Winter to Spring).

Reading: Three overall achievement outcomes are available: Proficient Reading and Basic Reading (used in progress monitoring and making decisions about instructional adjustments). A Composite Reading measure is also available in making screening decisions. The first two measures are separate from the individual measures that feed into the mosaic used to both create a multi-skill and multi method construct of reading; the last one (Composite) is constructed using three of the measures in Grades 3 through 8 (*measures listed in italics in Table 2 are not used for the composite score).*

Proficient Reading measures are available in Grades 3 through 8. In this measure, a rather extensive narrative text (story) of 1,000+ words is presented and students then answer 20 comprehension questions that represent literal and inferential as well as evaluative comprehension.

[1] https://conjointly.com/kb/multitrait-multimethod-matrix/

Basic Reading Measure is also available for teachers to use both as a total score in which the construct has been mapped into three reading components: Literal, Inferential, and Informative comprehension, as well as Locating Information. With these measures, the passages are shorter with questions following immediately after the text. This feature facilitates their use as progress measures for evaluating instructional interventions.

With the Composite Reading Measure, a total achievement score is based on three measures: Oral Reading Fluency, Vocabulary, and Proficient Reading combined in the following manner.

- The mean and SD for each of the three easyCBM® measures are used to calculate three individual z-scores, one for each measure: $(scorestudent - Meanmeasure)/SDmeasure$.
- The easyCBM® measure average is subtracted from the student's score, then divided by the standard deviation. This z-score is a way of calculating how far from the average student's score is, measured in a standard way that allows comparison across different scales.
- The average of those three z-scores is computed.
- This z-score average can then be directly mapped to a percentile by using the shape of the normal distribution. To convert that into a percentile, we use a quantile function (stats::qnorm()in R) to calculate the proportion of the distribution that lies below that z-score.
- For example, a z-score of 0 is right at the middle of the distribution, which corresponds to the 50th percentile. A z-score of +1.0 is one standard deviation above average and corresponds to roughly the 84th percentile, meaning the score is higher than about 84% of all values.

MS-MM Reading Measures (User's Manual, page 13)

Many different reading measures exist over the entire grade span of Kindergarten through Grade 8. Both production and selection responses (assessment methods) are deployed to ensure adequate construct representation.

Table 2. Measures that Map into a Multi-Skill Construct of Reading Achievement (User's Manual, page 9)

Grade-Level Benchmark Measures by Assessment Period

	Grade Level	Fall Benchmark	Winter Benchmark	Spring Benchmark
Reading	K	Letter Names Letter Sounds Phoneme Segmenting	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Phoneme Segmenting Word Reading Fluency
	1	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency
	2	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading
	3–8	Passage Reading Fluency Vocabulary Proficient Reading *Basic Reading**	Passage Reading Fluency Vocabulary Proficient Reading *Basic Reading**	Passage Reading Fluency Vocabulary Proficient Reading *Basic Reading**

These reading assessments reflect the "Big Five" components of reading identified by the National Reading Panel (2000) report and state standards. The subtitle to this publication is "Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction." We introduce the various subtests of reading in a developmental manner, beginning with three measures in Kindergarten and then over the grades, moving through a progression of further subtests.

Table 3. Measures Defining Various Constructs of Reading (User's Manual, page 13).

Measures	Constructs
Phoneme Segmenting	Phonemic Awareness
Letter Names Letter Sounds	Alphabetic Principle and Phonics
Word Reading Fluency Passage Reading Fluency	Fluency
Vocabulary	Vocabulary
Basic Reading Proficient Reading	Comprehension

MS-MM Spanish Literacy Measures

Spanish language literacy measures are available for Kindergarten through Grade 8 for all seasonal Benchmarks (BMs): Fall, Winter, and Spring. These measures are authentic Spanish measures, created by native Spanish speakers, rather than translations of *easyCBM® District Edition User's Manual.* Measures and items were developed to be appropriate for assessing students receiving literacy instruction in Spanish. The specific measures available are based on three years of research at the University of Oregon to identify and develop CBMs addressing the ways in which Spanish literacy develops.

Table 4. Measures Defining the Construct of Early Spanish Literacy Proficiency (User's Manual, page 10)

	Grade	Fall BM	Winter BM	Spring BM
Spanish Literacy	K	Syllable Segmenting Syllable Sounds	Syllable Segmenting Syllable Sounds	Syllable Segmenting Syllable Sounds
	1	Syllable Segmenting Syllable Sounds Word Reading Sentence Reading	Syllable Segmenting Syllable Sounds Word Reading Sentence Reading	Syllable Segmenting Syllable Sounds Word Reading Sentence Reading
	2	Word Reading Sentence Reading Vocabulary	Word Reading Sentence Reading Vocabulary	Word Reading Sentence Reading Vocabulary
	3–8	Vocabulary	Vocabulary	Vocabulary

An Application of MS-MM Analysis of Overall Achievement

Given this array of individual measures (administered with different methods using both selection and production responses, analysis of overall achievement in easyCBM® is not only based on a single score but also a mosaic of different score patterns across the measures at any one grade level. In turn, this mosaic is directly turned into a risk analysis (as addressed in Section 3.3) and progress monitoring (as addressed in Section 3.4). In both cases, normative performance is used as a basis for determining both who is at risk of learning problems and how should they be monitored.

The mosaic of overall achievement provides teachers to an overall achievement index across skill areas that can be convergent or divergent (or discriminate). Clearly, when the reading skills within a grade and benchmark are all high or all low, overall achievement can be viewed as either, respectively. However, when proficiencies across the measures are not consistent, then divergent (or discriminate) information is provided for making judgments of risk that may be skill specific.

For example, if a student is high on early reading measures (letter names and sounds) but low on word or passage reading fluency, teachers can view their overall achievement as marginal in the synthesis of reading skills. If students are dysfluent readers (their rate of word reading correct per minute in either word or passage fluency), and they are low in vocabulary or comprehension, the structural components of overall reading achievement is in question.

In the end, this type of convergent or discriminate patterns of skills allows teachers to more quickly connect assessment information to judgments of risk and development of interventions. A note of caution with divergent patterns: The method of assessment should be analyzed as a possible explanation, in which the construct and response process may be questioned. For example, students whose word and passage reading fluency (overall achievement) is low, but their comprehension is high, teachers may need to investigate further, whether the structural nature of measuring fluency or the background knowledge of comprehension is contributing construct irrelevant variance to the constructs of reading. A final issue to consider is that, as noted in the next section, overall achievement in these skill areas may change at different rates, over time. See Table 6.

3.1c Guidance for Interpreting Overall Achievement with Performance Distributions and Percentile Ranks

The basic interpretive guide for reporting overall achievement is students' normative performance, which is reported in detail in Section 3.2. These performance levels are provided for all measures and all grades. When interpreting easyCBM® results, it is important to refer to the percentile rank associated with a given raw score at the time of year the measure was administered. These percentile ranks are based on national norms, which are re-calculated and updated every five years. Because of the way percentile ranks work, performance at or near the 50th percentile rank can be interpreted as average performance for students in that grade level at that time of the year. Scores above the 50th percentile rank indicate performance above average for students in that grade at that time of the year, and scores below the 50th percentile rank indicate performance below average for students in that grade at that time. The easyCBM® system has a variety of reports designed to help teachers identify students at risk, pinpoint what content they may need additional support to master, and track improvements over time.

Districts have options in determining risk using percentile ranks as part of Multi-Tiered Systems of Support (MTSS). The rationale for this option is that districts may differ in their student populations and available resources. Because assignment of special education labels is essentially a social, not medical decision, districts need to consider both components in making a value-driven decision. In a DYK on classification accuracy, we address both sensitivity and specificity in weighing true and false positives for assigning both special education classifications and levels of support. In addition, our research indicates that traditional cut off percentile rank (PR) values of 10th or 15th PR values may need to privilege higher PRs for reaching proficiency on state tests.

Note: Depending upon District policies and practices, specific cut-scores are established for defining risk: See 3.3b and 3.3c Using easyCBM® to Document Individual Differences, Figure 1. Adjusting Benchmark (BM) Test Windows and Risk Values (User's Manual, pages 76-77). In the end, teachers receive classroom level risk for achieving expected levels of performance so they can make decisions about the amount of instructional and specialized supports needed.

Using easyCBM® to Document Individual Differences

With easyCBM®, we approach the validation process from two perspectives: (a) nomothetic and (b) idiographic. Dr. Stan Deno, the key person who founded Curriculum-Based Measurement (CBM) referred to these two perspectives as appreciating individual differences (nomothetic with a view on distributions of students) versus making an individual difference (an idiographic view with a focus on progress over time)[2].

[2] Deno, S. L. (1990). Individual Differences and Individual Difference: The Essential Difference of Special Education: The Essential Difference of Special Education. *The Journal of Special Education*, *24*(2), 160-173. https://doi.org/10.1177/002246699002400205.

With both perspectives, the emphasis on validation is about the decisions, not the measures. “It is the interpretation of test score for proposed uses that are evaluated, not the test itself…each intended interpretation must be evaluated. Statements about validity should refer to interpretations for specified uses. It is incorrect to the unqualified phrase “the validity of the test” (p. 11)[3].

Table 5. Cross-Tabulation of Perspectives and Time

Perspective / View	Point in Time	Over Time
Nomothetic	1. Single Benchmark	2. Benchmark Comparisons
Idiographic	3. Risk Analysis	4. Progress Monitoring

In this section, we focus primarily on cell 1: Single Benchmark performance as an indicator of overall achievement. Note that this can occur at any time of the year. Typically, teachers begin school in the fall and would be grouping students to assist in targeting instruction (either in general skill group in Tier I or in specialized instruction in Tiers 2 and 3). See Figure 2. Over time, teachers may be interested in determining if this grouping structure maintains and may be interested in progress of students over time (relative to the norms) to readjust groups if warranted. See cell 2 and Figure 3. In both decisions (initial grouping and re-grouping, we emphasize a norm-referenced view, which is addressed in Section 3.2 for teachers to predict performance. Note: We address cells 3 and 4 in the last two sections (3.3 and 3.4): Sub scores and progress monitoring, respectively.

Overall Achievement Reports

- Group Reports: Multiple options are provided for viewing data for groups of students, including graphs, summary reports, and item analysis on individual test forms.
- Benchmark Performance Reports: An overview of an individual student's benchmark performance for the current school year.

Table 6. Validation Supporting Decision Making

Figure Number and Display	Information Presented	Decision/Interpretation to Make
Figure 2. Bar chart	Distribution of students in classrooms (with skills)	Grouping students with minimum variance
Figure 3. Time series	Progress of students (in groups) progress over time	Confirm/disconfirm grouping composition

This performance on BMs can also be used to help teachers **group students** and display strengths and weaknesses among students based on an item analysis. **Figure 2** illustrates how benchmark performance combined with item-level skill analysis can be used to group students for instruction. Rather than grouping solely on total scores, this display allows teachers to identify specific skill profiles that can guide targeted small-group instruction. The decision supported is how to form instructional groups that maximize instructional relevance and efficiency. This figure aligns with Testing Standards 3.1 and 4.1 by supporting construct-aligned interpretation and appropriate instructional use. It also reinforces the principle that assessment-driven grouping should be dynamic and skill-based, not static or label-driven.

[3] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC.

Figure 2. Using Benchmark Performance to Group Students with Skill Analysis (User's Manual, page 95)

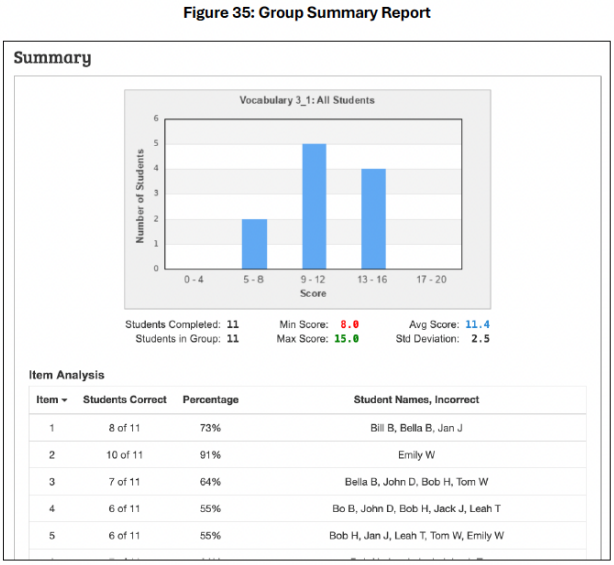


Figure 35: Group Summary Report

Summary

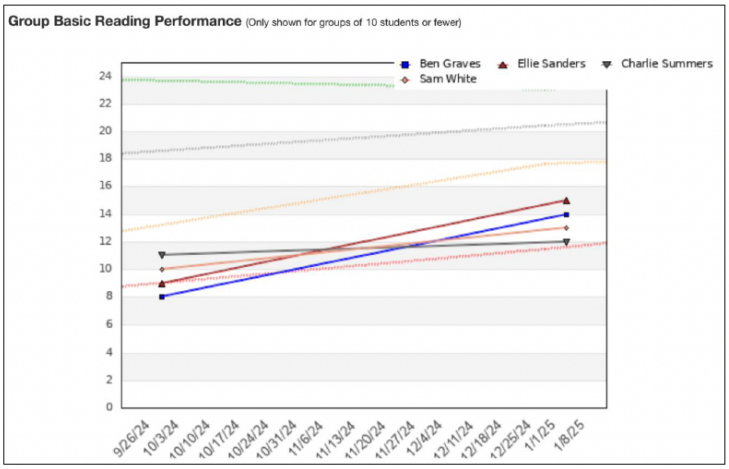
Students Completed: 11 Min Score: 8.0 Avg Score: 11.4
Students in Group: 11 Max Score: 15.0 Std Deviation: 2.5

Item Analysis

Item	Students Correct	Percentage	Student Names, Incorrect
1	8 of 11	73%	Bill B, Bella B, Jan J
2	10 of 11	91%	Emily W
3	7 of 11	64%	Bella B, John D, Bob H, Tom W
4	6 of 11	55%	Bo B, John D, Bob H, Jack J, Leah T
5	6 of 11	55%	Bob H, Jan J, Leah T, Tom W, Emily W

As students are grouped, their progress can also be displayed to allow teachers an opportunity to re-group students. **Figure 3** extends benchmark grouping displays by incorporating progress-monitoring data over time. Teachers can evaluate whether students within an instructional group are responding similarly to instruction. If students show divergent growth patterns, regrouping may be warranted. The figure reinforces that grouping decisions should be revisited regularly based on evidence of student response. This visualization supports ongoing instructional decision-making and aligns with Standards 1.4 and 4.10 by emphasizing monitoring of instructional effectiveness.

Figure 3. Extending Benchmark Performance to View Students Progress (User's Manual, page 96)

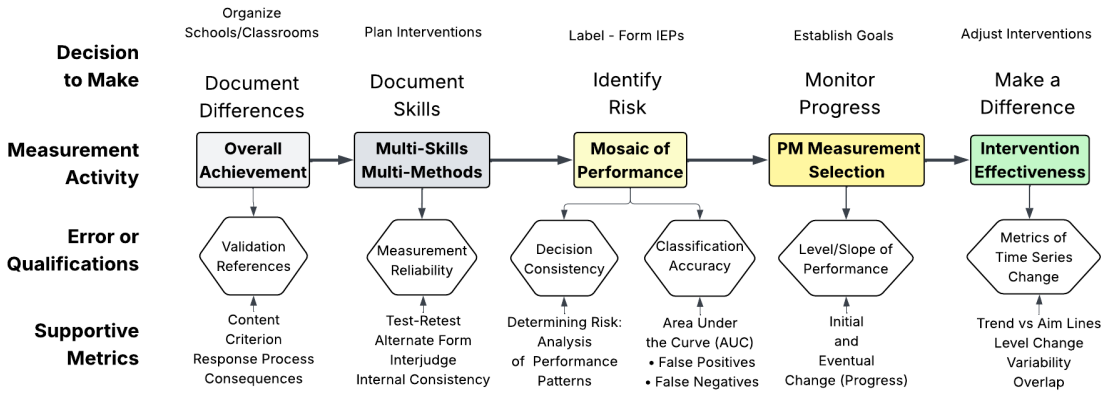


Conclusions with a Validity Argument

easyCBM® provides reports using both an Overall Outcome in each subject area (Reading, Spanish literacy, and Math) as well as a matrix reflecting Multi-Skill Multi Method (MS-MM) approach: multiple skills are assessed using multiple methods (selection and production responses as well as computer-based and paper-pencil with a variety of accommodations). By focusing on skill specific distributions, and the use of normative performance at three benchmark periods, information on student performance quickly bridges overall achievement with sub score analyses (determination of risk) and appropriate progress monitoring where required. Errors in interpretation are also addressed with the convergent and discriminate information that may be contrary to expectations, allowing teachers to follow up with more refined analyses (e.g., using different assessment methods). A User Manual provides teachers extensive information on how to document this process.

A **Complete Validity Argument** is now possible by **beginning with a decision to be made** and then associating it with an appropriate measurement activity, which also carries with it qualifications or potential sources of error. First, easyCBM® documents Overall Achievement in reading, Spanish, and math that is supported with reference to multiple skill performance documented through multiple measurement methods (MS-MM). Overall achievement and multiple skills are interpreted as relative performance, using percentile rank to note individual differences (addressed in **Section 2**). Later, in **Section 3**, further consideration is given to this MS-MM by analyzing this mosaic of separate sub-test performances and classification accuracy to determine risk of failing to succeed and providing students supports (Tiers 2 and 3). The decision-making process then turns from *documenting individual differences* to *making an individual difference* (addressed in **Section 4**). The focus now is on selecting appropriate measures for progress monitoring (PM) and finally in determining intervention effectiveness, through interrupted time series graphic displays for each student. In **Figure 4** below, the top row displays the decision to make; the second row presents the measurement activity (often the most visible event that requires teachers' time and training); in the third row is potential error (or qualification) associated with the measurement activity; finally, in the bottom row is a description of supportive metrics, that lead back to the measurement activity and decision being made. Notice the colors change from neutral (gray) to success with hesitation (yellow) to success (green). This complete validation argument fits within the larger science of education in which conjectures are affirmed only after dubitation is quieted.[4]

Figure 4. Comprehensive Validity Argument in easyCBM®



[4] Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge – 2nd Edition*. New York: Routledge.

Appendix A: Technical Report Table and Figure Titles

Table 1. Usage and Reporting for easyCBM® (User's Manual, page 4)

Table 2. Measures that Map into a Multi-Skill Construct of Reading Achievement (User's Manual, page 9)

Table 3. Measures Defining Various Constructs of Reading (User's Manual, page 13).

Table 4. Measures Defining the Construct of Early Spanish Literacy Proficiency (User's Manual, page 10)

Table 5. Cross-Tabulation of Perspectives and Time

Table 6. Validation Supporting Decision Making

Figure 1: Using easyCBM® to Improve Student Learning (User's Manual, page 6)

Figure 2. Using Benchmark Performance to Group Students with Skill Analysis (User's Manual, page 95)

Figure 3. Extending Benchmark Performance to View Students Progress (User's Manual, page 96)

Figure 4. Complete Validity Argument in easyCBM®

Technical Report References

easyCBM User's Manual (2025). Eugene, OR: University of Oregon – Behavioral Research and Teaching.

Conclusions Supporting Claims for Criterion 3.2: Predicted Student Performance

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of predicted student performance.

This evaluation applies Criterion 3.2 (Indicators 3.2.a–3.2.c) to all sections of the EGCI_3.2 documentation and finds strong evidence that score reports, norms, and supporting resources appropriately facilitate valid interpretations and uses of predicted student performance.

Indicator 3.2.a: The design of score reports and interpretive materials is well aligned with intended users and purposes. Benchmark reports are clearly framed for educators' use in universal screening, progress prediction, and instructional planning. The grain size of reported information—measure-level scores, composites, and percentile ranks by grade and season—is appropriate for classroom, school, and district decision making. Documentation explicitly distinguishes Proficient and Basic measures and includes cautions for use with students performing well below grade level, reducing the risk of misinterpretation. The availability of English and Spanish benchmark configurations further demonstrates attention to diverse user needs.

Indicator 3.2.b: The system provides defensible representations of uncertainty around predicted performance through normative percentile ranks and clear seasonal benchmarks. The 2025 norms are derived from large, stratified national samples and use non-parametric estimation methods well suited to CBM score distributions. Supporting documentation explains distributional properties and seasonal growth patterns, enabling users to understand variability and avoid over-interpretation of small score differences. These features support accurate interpretation of predicted performance across the score range.

Indicator 3.2.c: Sufficient and appropriate guidance is provided to support intended uses. Interpretive guidance is grounded in extensive peer-reviewed research and long-standing CBM practice. Norm tables, composites, and benchmark expectations are explicitly linked to screening and intervention decisions, while warnings highlight conditions under which interpretations may be compromised (e.g., inability to complete grade-level measures). Guidance supports interpretation for students across the full performance continuum, including those at high risk and those exceeding benchmarks.

Overall, EGCI_3.2 demonstrates strong alignment between report design, normative evidence, and interpretive guidance. The documentation provides educators with clear, research-based tools to predict student performance responsibly, while explicitly bounding interpretations to prevent misuse and unintended consequences.

3.2a Intended Interpretations for Benchmark Measures (District User's Manual – Page 9)

The purpose of benchmark testing, or universal screening, is to provide information regarding students' progress toward meeting end-of-year grade-level expectations and to determine which students may require intervention or enrichment. Benchmark assessments (BMs) are given on grade-level material and are administered three times per year: Fall, Winter, and Spring. These BMs provide the most critical datum for predicting performance in two critical ways: Because the norms occur at three times during the year (early in the fall, mid-year in the winter, and end of year in the spring), users can gauge 'normal' development and use this as a guide for setting expectations (e.g. goals). Second, when districts assess their own students as part of the Multi-Tier Support System (MTSS), they can compare their own distributions with those from the national normative sample. Third and finally, these national norms use a well-balanced stratified random sample and thus can provide stability in comparing values for both expectations and local distributions.

Benchmark measures are designed to assess students' performance on the highest-priority instructional targets in reading for each grade level. See the table below for the measures thar are administered at each benchmark assessment period by subject and grade. For the system to provide a reading composite score, all applicable reading benchmark measures for that grade level and assessment period must be administered. Measures listed in italics are not required for the composite score.

Table 1. Measures Deployed as Benchmarks by Grade and Season (User's Manual, page 9)

Grade-Level Benchmark Measures by Assessment Period

	Grade Level	Fall Benchmark	Winter Benchmark	Spring Benchmark
Reading	K	Letter Names Letter Sounds Phoneme Segmenting	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Phoneme Segmenting Word Reading Fluency
	1	Letter Sounds Phoneme Segmenting Word Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency	Letter Sounds Word Reading Fluency Passage Reading Fluency
	2	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading	Passage Reading Fluency Vocabulary Proficient Reading
	3–8	Passage Reading Fluency Vocabulary Proficient Reading *Basic Reading**	Passage Reading Fluency Vocabulary Proficient Reading *Basic Reading**	Passage Reading Fluency Vocabulary Proficient Reading *Basic Reading**

easyCBM District Edition User's Manual © University of Oregon 9

Table 2. Spanish Literacy Measures Deployed as Benchmarks by Grade and Season (User's Manual, page 10)

Spanish Literacy	K	Syllable Segmenting Syllable Sounds	Syllable Segmenting Syllable Sounds	Syllable Segmenting Syllable Sounds
	1	Syllable Segmenting Syllable Sounds Word Reading Sentence Reading	Syllable Segmenting Syllable Sounds Word Reading Sentence Reading	Syllable Segmenting Syllable Sounds Word Reading Sentence Reading
	2	Word Reading Sentence Reading Vocabulary	Word Reading Sentence Reading Vocabulary	Word Reading Sentence Reading Vocabulary
	3–8	Vocabulary	Vocabulary	Vocabulary

3.2b Score Reports Include Information on the Degree of Error

Reading is an optional benchmark assessment. For universal screening, most students should be given Proficient Reading. However, for students performing well below grade level, it may make sense to administer Basic Reading in place of or in addition to the Proficient measures. If a student is unable to complete the Proficient Reading test (e.g., due to low accuracy or fluency), consider giving Basic Reading instead.

Percentile Ranks (PRs) and Successive Raw Scores: Why Differences Behave Differently in Middle vs. Tails
Percentile Ranks (PRs) convert a raw score into the percentage of the norm group that scored at or below that score. Because PRs depend on the *shape* of the score distribution—usually bell-shaped—the relationship between raw-score increments and PR increments is non-linear. This leads to a fundamental rule: A fixed increase in raw score produces large PR changes in the middle of the distribution but small PR changes in the high and low tails.

---

1. Behavior in the Middle of the Distribution

In the middle (around the mean), the distribution is:

- The densest (many students have similar scores)
- The slope of the cumulative distribution function is the steepest

Consequence: A 1-point increase in raw score moves a student ahead of many more peers, because many students cluster near the average.

Example: Suppose scores cluster around 20–25 with Raw scores: 20 → 21 → 22

Approximate PRs:

- 20 → 40th percentile
- 21 → 47th percentile
- 22 → 54th percentile

Successive PR differences:

- 40 → 47 = +7 PR points
- 47 → 54 = +7 PR points

Here, each 1-point jump yields a relatively large, nearly linear PR increase.

---

2. Behavior at the Lower Tail of the Distribution

At the low end, the distribution is:

- Sparse (few students score this low)
- The cumulative curve is flatter

Consequence: A 1-point increase reflects moving past very few additional students.

Example: Raw scores: 3 → 4 → 5

Approximate PRs:

- 3 → 1st percentile
- 4 → 2nd percentile
- 5 → 3rd percentile

Successive PR differences:

- 1 → 2 = 1 PR point
- 2 → 3 = 1 PR point

Small raw-score gains give very small PR changes.

---

3. Behavior at the Upper Tail of the Distribution

Similarly, at the high end:

- The distribution again becomes sparse
- Very few students have those high scores

Consequence: A 1-point increase gains almost no additional PR movement.

Example Raw scores: 38 → 39 → 40

Approximate PRs:

- 38 → 96th percentile
- 39 → 97th percentile
- 40 → 99th percentile

Successive PR differences:

- 96 → 97 = 1 PR point
- 97 → 99 = 2 PR points

Even large raw-score differences translate into small changes in the PR metric.

---

4. Why This Happens: The Mathematical Explanation

Percentile Rank = area under the curve (CDF): A 1-point jump in raw score corresponds to the integral of the density over that interval.

Density is:

- High at the mean → large PR changes for small raw-score changes
- Low at extremes → tiny PR changes even for moderate raw-score changes

Thus, PRs compress at the tails and stretch in the middle. This is a universal property of PRs derived from any unimodal distribution, not just the normal curve.

Table 3. Summary of Key Differences

Location in Distribution	Raw-Score Differences	Effect on PRs	Explanation
Low tail	Successive differences look the same numerically (e.g., 3→4→5)	Very small changes in PRs	Few students score this low; CDF is flat
Middle	Same raw-score differences	**Large** PR jumps	Many students cluster near the mean
High tail	Same raw-score differences	Very small PR changes	Few students score this high; CDF flattens again

Practical Implications

1. PRs exaggerate differences in the middle – Small raw-score gains can look large (e.g., moving from 40th to 55th PR).
2. PRs compress differences at the extremes – Big instructional gains for high-performing students may barely move PRs.
3. Interpretation is non-linear – A 10-point PR gain is not comparable across the scale.
4. Better interval scales (e.g., scale scores, z-scores) are needed for growth modeling or psychometric analyses.

3.2c Interpretations for All: Summary of easyCBM® 2025 Norms Technical Report (Technical Report No. 2508)

This technical report presents the 2025 national norms for the easyCBM® reading assessment system, developed by Behavioral Research and Teaching (BRT) at the University of Oregon. easyCBM® is designed as a universal screening and progress-monitoring system for students in Grades K–8. The 2025 norms update provides educators, researchers, and policymakers with refreshed national benchmarks for interpreting student performance across foundational and advanced reading skills.

The primary purpose of the report is to establish updated percentile-based normative reference points that allow schools to evaluate students' academic standing relative to a nationally representative peer group. These norms support high-stakes educational decisions such as early identification of risk, tiered intervention placement, instructional planning, and monitoring of student growth over time.

Data for the 2025 norms were drawn from student performance during the 2024–2025 academic year using valid easyCBM® district accounts. A rigorous multi-step data screening process was used to ensure the validity and interpretability of scores. Students were excluded if they tested on more than one grade-level measure during the school year, if tests were incomplete, if scores fell outside the allowable operational range for each measure, or if values were missing. Additional seasonal windows were enforced to maintain comparability across administrations: Fall (August 1–October 15), Winter (December 1–February 15), and Spring (March 15–June 15).

For most reading measures, only districts demonstrating broad testing coverage were included to reduce selection bias. Specifically, for Grades K–7, district-measure-season cells were retained only if at least 50% of students were tested. For Grade 8, slightly different proportional thresholds were used due to naturally smaller participation rates. Spanish measures and Rapid Automatized Naming (RAN) relied on all available post-screening data because of smaller participation volumes.

Following district-level filtering, the study employed stratified random sampling to ensure national geographic representation. The United States was divided into four regions (Midwest, Northeast, Southeast, and West). Each region contributed up to 500 randomly selected students per grade and season. This resulted in a consistent target of approximately 2,000 students per grade-season combination for most English reading measures, yielding strong statistical stability for percentile estimation.

Sample Demographics and Score Statistics for every measure and grade are reported using a stratified random sample (of the 2000 students) by region (Midwest, West, Northeast, and Southeast) with representation reported (count and percent) for gender (female, male, X), disability (yes or no), ethnicity (Hispanic, not Hispanic, or unknown), race (American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, Two or more races, White, or Unknown), and English language status (yes or no). For each measure and grade, descriptive statistics are reported, including Sample Size, Minimum score, Maximum Score, Skewness, Mean, and Standard Deviation. Note that in for regions with fewer than 500 students, the Sample Size include all students in that region; otherwise, most typically, the Sample Size was 2,000. Following these demographic and descriptive statistics, the Percentile Ranks (PRs) are reported for six-week periods for Fall, Winter, and Spring; the PRs are reported for every score value of the entire score distribution.

All normative PRs were estimated with an empirical cumulative distribution function (ECDF; ecdf function from the R {stats} package), a non-parametric estimator of the cumulative distribution function of the data samples for a measure, grade, season. This method avoids assumptions about normality and is well-suited for curriculum-based measures, which often exhibit skewness, ceiling effects, and growth-related distributional changes across the school year. All PRs were rounded to the nearest whole number and capped at the 99th percentile. The ECDF shows the proportion of observations in a sample that are less than or equal to a given value. Estimated percentiles were rounded to the nearest integer, and the percentile ceiling was constrained to be equal to 99. All analyses and figures were conducted and created in the R programming environment (R Core Team 2024a) with the following R packages: data.table (Barrett et al. 2024); knitr (Xie 2024); moments (Komsta and Novomestky 2022); readxl (Wickham and Bryan 2023); stats (R Core Team 2024b); and tidyverse (Wickham et al. 2019).

The norming system includes a comprehensive set of early literacy, advanced reading and Spanish reading measures. Foundational literacy measures include Letter Names, Letter Sounds, Phoneme Segmenting, Word Reading Fluency, and Rapid Automatized Naming. More advanced reading measures include Passage Reading Fluency, Vocabulary, Proficient Reading, and the Reading Composite. Parallel norms are also presented for multiple Spanish literacy measures, including Spanish Syllable Sounds, Spanish Syllable Segmenting, Spanish Sentence Reading, Spanish Word Reading, and Spanish Vocabulary. Results illustrate clear seasonal growth patterns across all reading measures.

In kindergarten, for example, *Letter Sounds* increase dramatically from fall to spring, reflecting rapid early phonics development. *Phoneme Segmenting* shows similar accelerated growth across kindergarten and first grade, supporting its role as a sensitive indicator of early phonological development. *Word Reading Fluency* and *Passage Reading Fluency* demonstrate consistent yearly growth across all elementary grades, with effect sizes increasing through Grade 5 and stabilizing through middle school.

Passage Reading Fluency (PRF), which serves as one of the most widely used screening indicators, shows particularly strong developmental progression. Mean winter PRF scores increase from approximately 38 words correct per minute in Grade 1 to well over 130 words correct per minute by Grade 5, with continued growth into upper grades. By Grades 6–8, score distributions reflect increased variability and reduced skew, consistent with advanced reading plateauing patterns.

Vocabulary and Proficient Reading measures extend interpretation beyond fluency into comprehension-related constructs. These measures show increasing means and narrowing distributions across grades, indicating growing consistency in students' academic language development.

The *Reading Composite* represents a major advancement in this edition of the norms. It integrates multiple reading indicators into a single standardized metric, allowing educators to examine overall reading proficiency using a unified scale. Composite norms and associated descriptive statistics are presented across all grades and seasons. *Spanish literacy* norms mirror the structure of the English assessments and provide critical benchmarking tools for bilingual programs and Spanish-speaking populations. Although Spanish samples are smaller than English samples, the norms still demonstrate logical developmental progressions across grades and seasons.

Overall, the easyCBM® 2025 Norms report provides one of the most comprehensive and methodologically rigorous national norming efforts available for curriculum-based measurement systems. By combining large-scale contemporary data, geographic stratification, non-parametric percentile estimation, and wide coverage of reading constructs, the norms offer educators defensible tools for screening, progress monitoring, and instructional decision-making. The 2025 update is particularly valuable for schools transitioning out of pandemic-era disruptions, as it reflects contemporary post-pandemic achievement patterns. As such, these norms are expected to play an important role in guiding early identification, intervention planning, and accountability processes across the United States.

Table 4. Summary of Major easyCBM® 2025 Norms Findings

Domain	Key Findings	Educational Implications
Early Literacy (K–1)	Rapid growth in letter sounds, phoneme segmenting, and word reading across fall–spring.	Supports early universal screening and phonics-based intervention decisions.
Fluency (Grades 1–8)	Consistent increases in PRF mean scores across grades, plateauing in middle school.	Allows accurate benchmarking for RTI placement and growth monitoring.
Vocabulary & Proficient Reading	Steady increases with tightening distributions across grades.	Supports comprehension-focused instructional planning.
Spanish Assessments	Parallel developmental trends with smaller national samples.	Supports dual-language and bilingual instructional programs.
Reading Composite	Unified normed metric across all grades and seasons.	Provides a single defensible indicator of overall reading proficiency.

Appendix A: Technical Report Table Titles

Table 1. Measures Deployed as Benchmarks by Grade and Season (User's Manual, page 9)

Table 2. Spanish Literacy Measures Deployed as Benchmarks by Grade and Season (User's Manual, page 10)

Table 3. Summary of Key Differences

Table 4. Summary of Major easyCBM® 2025 Norms Findings

Technical Report References

Nese, J. F. T., & Wallin, J. (2025). *easyCBM® 2025 norms (Technical Report # 2508)*. Behavioral Research and Teaching, University of Oregon.

Conclusions Supporting Claims for Criterion 3.3: Sub-scores

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of sub-scores.

This evaluation applies Criterion 3.3 (Indicators 3.3.a–3.3.c) to all sections of the EGCI_3.3 documentation and finds strong evidence that easyCBM® sub-score reports and supporting materials appropriately facilitate valid interpretation and use of sub-scores for instructional and system-level decision making.

Indicator 3.3.a: The design of score reports is closely aligned with intended interpretations and users. Reports provide multiple grain sizes—from individual sub-scores to classroom, school, and district summaries—allowing educators to examine domain-specific risk rather than relying on a single overall score. Visual displays (e.g., heat maps, stacked bars, and risk movement charts) are explicitly linked to decisions about tier placement, instructional grouping, and resource allocation. Documentation demonstrates attention to different audiences, including teachers, administrators, and parents, and includes cautions against common misuses such as treating risk status as fixed. Conditions under which interpretations may be compromised are articulated through benchmark windows, cut-score adjustments, and explicit guidance on interpreting risk.

Indicator 3.3.b: The system conveys uncertainty in sub-score interpretation through percentile-based risk bands, color-coded classifications, and longitudinal change displays (e.g., just noticeable difference). While numeric confidence intervals are not always presented, the combination of multiple sub-scores, repeated measurements, and movement across benchmarks provides users with a practical representation of variability and classification error. Supporting guidance emphasizes sensitivity and specificity trade-offs when setting risk cut points, reducing over-interpretation of single sub-scores.

Indicator 3.3.c: Extensive research-based guidance supports appropriate use of sub-scores. Interpretive materials are grounded in CBM theory, MTSS practice, and national testing standards, and were developed with experienced educators. Guidance supports decisions across the full performance continuum, from early identification of risk to monitoring change over time and evaluating instructional impact. The integration of nomothetic (risk classification) and idiographic (change and response) perspectives further strengthens validity.

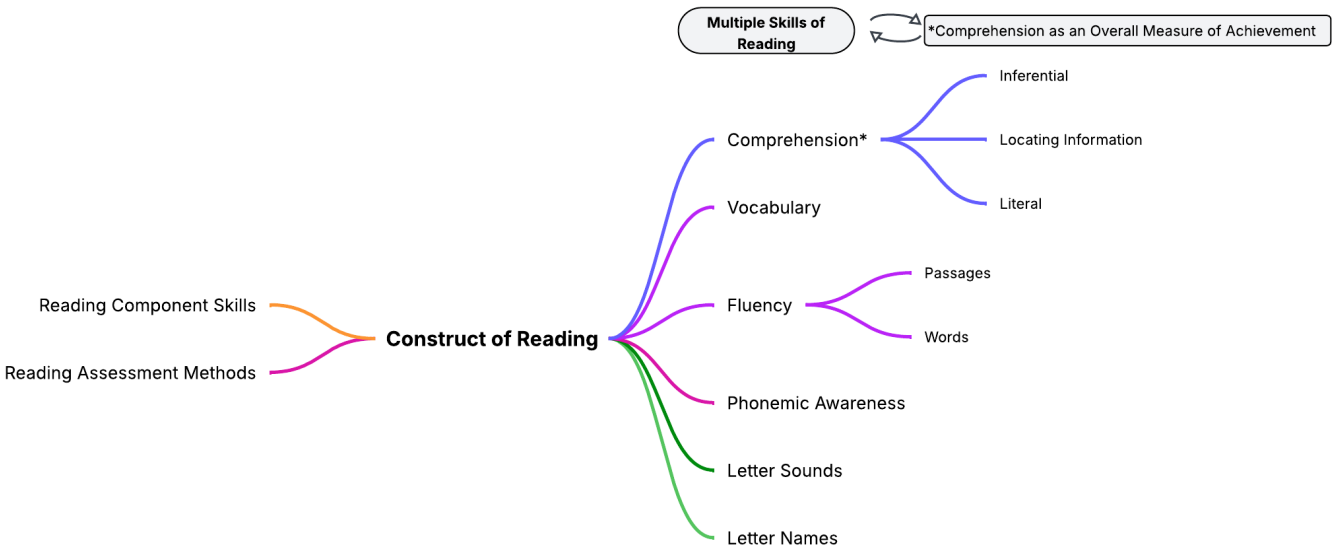
Overall, EGCI_3.3 demonstrates strong coherence among sub-score design, interpretive guidance, and intended instructional uses, supporting defensible, transparent, and equitable decision making.

3.3a Documenting Risk using Sub Scores in Achievement (from User's Manual, pages 91-92)

making. Benchmark reports enable users to identify specific broad constructs (fluency, vocabulary, comprehension, mathematics) in which students are either struggling or meeting expectations, thus facilitating decisions related to programmatic and curricular supports. Group-level reports provide insights into the specific skills students have mastered or with which they are struggling (such as specific letter sounds or skill objectives within math), fostering informed lesson planning based on student needs. Individual reports also enable teachers to monitor the effectiveness of specific interventions for individual students and provide an accessible way to communicate with parents. Sub score achievement (Benchmark [BM] facilitates (a) Identifying students at risk [BM] and (b) monitoring the movement of students across instructional tiers [BM]. Because easyCBM®® is web-based, all reports are available online immediately after administration (for student online testing or live scoring) or after data entry (for paper-and-pencil administration.

In **Figure 1** below, latent **reading** skills and methods are operationalized into a single construct of reading, eventually defined as 'Comprehension' in large scale testing programs. In easyCBM®, overall comprehension scores can be delivered with three measures: Proficient, Basic, and Composite as presented in the earlier Section 3.1. This section, however, presents a model of reading assessment that also includes documentation and analysis of intact sub scores reflecting Multiple Skills (MS) assessed with Multiple Methods (MM). These sub scores can be viewed not only individually but also as comprising a mosaic (pattern) that targets risk of failing to learn skills specific areas. Note that these skills form building blocks upon which subsequent (later) skills are based. For example, letter names and sounds are critical to phonemic awareness and all three skills become key to reading words and passages fluently. With oral reading fluency, comprehension becomes an important correlate. In this developmental sequence, students produce the response, listen and identify responses, and use selection responses. Also, given the use of computer-based and paper-pencil administrations, as well as any number of accommodations (in settings and administration directions) that can be implemented during administration (as described in Section 1.3), these methods of assessment ensure a robust overall measure of achievement concurrent with critical subskills. One final important aspect of this assessment model is that appropriate assessments can be conducted in grades earlier than third. Indeed, for many students, by third grade, overall achievement levels often become intractable.

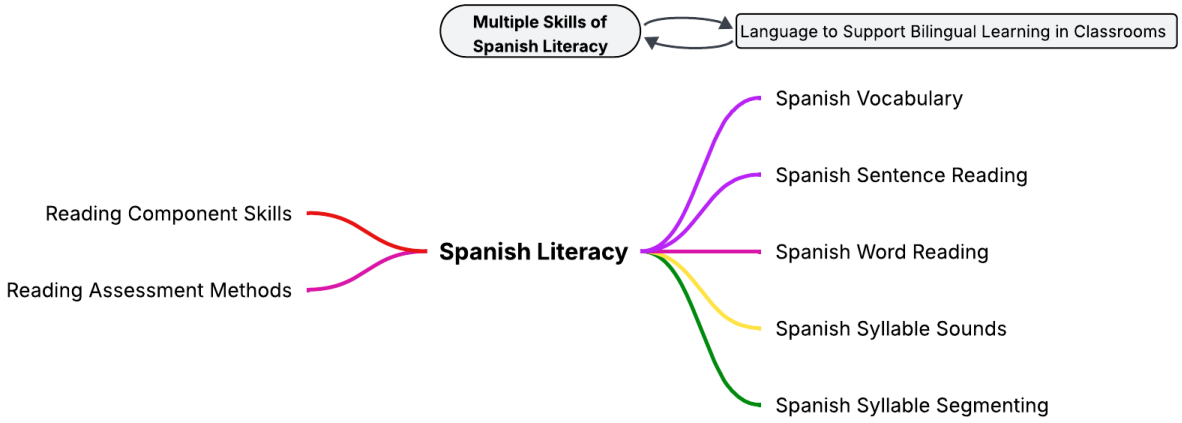
Figure 1. Reading Sub Scores Critical to the Overall Measure of Comprehension



In Figure 2 below, latent **Spanish literacy** skills and methods are operationalized into a single construct of Spanish Literaty, eventually defined as 'Comprehension' in large scale testing programs. In easyCBM®, overall comprehension scores can be delivered with three measures: Proficient, Basic, and Composite as presented in the earlier Section 3.1. As in reading, Multiple Skills (MS) are assessed with Multiple Methods (MM), allowing sub scores to be viewed not only individually but also as comprising a mosaic (pattern) that targets risk of failing to learn skills specific areas. And, like reading…

- The skills form building blocks upon which subsequent (later) skills are based.
- Multiple types of responses are used in the assessment methods.
- Multiple types of accommodations are available in the assessment methodology (primarily in setting and administration directions.
- Computer-based and paper pencil administration is available.
- Important sub skills are provided for monitoring progress.
- Early assessments developmentally allow skills to be documented before Grade 3, when large-scale testing programs are deployed.

Figure 2. Spanish Sub Scores Critical to the Overall Measure of Comprehension



Sub Score Risk Reports

- Composite Report: A tabular report showing scores, percentile ranks, and the new Reading Composite Score **risk** for a group of students on a single benchmark assessment.
- Benchmark Score Reports: A tabular report showing scores, percentile ranks, and **risk** levels for a group of students on a single benchmark assessment.
- Benchmark History Reports: Individual student **risk** history across multiple years.
- Risk Analysis Reports: A breakdown of the change in students' **risk** levels from one benchmark assessment to the next.
- Grade/Measure Comparison Reports: A customizable view of the percent or number of students within each **risk** level, broken down by different grades, measures, and seasons.
- School Comparison Reports: A customizable view of the percent or number of students within each **risk** level by grade and measure, broken down by individual schools.
- RAN Report: A tabular report showing student testing results, including the calculated Correct Digits Per Minute score, for the Rapid Automatic Naming measure which is used for **dyslexia** screening.
- Analytics: A tool to count the number of students tested, broken down by schools or teachers.

These various reports can be classified by the type of visual display provided, the information that is presented, and the decision/interpretation that is warranted: See Table 1.

Table 1. Validation Supporting Decision Making

Figure Number and Display	Information Presented	Decision/Interpretation to Make
Figure 3. Benchmark Windows and Levels	When to activate BMs and set levels of risk	Allocation of Tiers of Support to provide
Figure 4. Heat map	Risk Analysis for students by measures in classrooms	Tier of Support to provide
Figure 5. Heat Map	Risk Analysis Reading Composite by classroom	Tier of Support to provide
Figure 6. Vertical stacked bar chart	Individual standing in risk bands	Continue or change Tier of Support to provide
Figure 7. Heat map	Change in risk by BMs and measures over grades	Confirm/disconfirm risk and Tiers of Support
Figure 8. Heat Map/Bar Chart	Risk Analysis for students across measures for BMs and grades	Tier of Support to provide and confirmation/disconfirmation
Figure 9. Heat map	Change in risk by BMs and measures over grades	Confirm/disconfirm risk and Tiers of support
Figure 10. Table of percentages	Changes in risk level across BMs and grades	School allocation of resources to assign
Figure 11. Stacked bar chart	Percentage of students risk levels by schools	District level and policy formation and allocation of resources across schools

3.3b Cut Scores to Establish Risk Based on Individual Differences

This section continues analyses from the nomothetic view, using reports to display individual differences and in particular highlights students at risk of successful learning? Which students are receiving various levels of support and the effects within and across the school year? This section only addresses Risk with Sub Scores in documenting Individual Differences.

An important component of screening is the establishment of cut-scores for making decisions about who receives what levels of support. As in all Multi-Tier Support Systems (MTSS), three tiers are typically demarcated: Tier 1 in which most students receive the general education curriculum with minor individualization, Tier 2 with considerable adaptations made to provide specialized supports, and Tier 3, in which specialized supports are individualized. This purpose reflects a nomothetic approach where individuals are compared to each other.

Districts have options in determining risk using percentile ranks as part of MTSS. The rationale for this option is that districts may differ in their student populations and available resources. Because assignment of special education labels is essentially a social, not medical decision, districts need to consider both components in making a value-driven decision. In a DYK on classification accuracy, we address both sensitivity and specificity in weighing true and false positives for assigning both special education classifications and levels of support. In addition, our research indicates that traditional cut off percentile rank (PR) values of 10th or 15th PR values may need to privilege higher PRs for reaching proficiency on state tests.

Figure 3 illustrates how adjusting percentile-based risk cut values changes the classification of students within a Multi-Tiered Systems of Support (MTSS) framework. Not only can the Benchmark Measure (BM) window of test administration be adjusted but the display makes explicit that risk designation is not an inherent property of a student score, but a decision driven by policy choices. By manipulating cut points, educators and district leaders can directly observe the resulting balance between false positives and false negatives. This visualization supports decisions about where to set risk thresholds based on district priorities, available instructional resources, and tolerance for classification error. From a validity perspective, the figure emphasizes consequential validity: decisions about risk must be evaluated in terms of their instructional impact rather than statistical convenience alone. Educators can use this display to justify why certain percentile ranks are selected for Tier 2 or Tier 3 placement and

to communicate transparently with stakeholders about the implications of those choices. The figure aligns with Standards 1.1 and 1.4 by clarifying intended interpretations and uses of scores, and with Standards 3.1 and 7.1 by highlighting fairness considerations inherent in classification decisions.

Figure 3. Adjusting Benchmark (BM) Test Windows and Risk Values (User's Manual, pages 76-77)

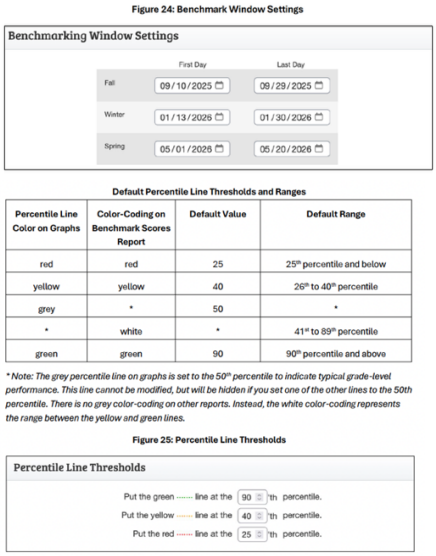


Figure 24: Benchmark Window Settings

Benchmarking Window Settings

	First Day	Last Day
Fall	09/10/2025	09/29/2025
Winter	01/13/2026	01/30/2026
Spring	05/01/2026	05/20/2026

Default Percentile Line Thresholds and Ranges

Percentile Line Color on Graphs	Color-Coding on Benchmark Scores Report	Default Value	Default Range
red	red	25	25th percentile and below
yellow	yellow	40	26th to 40th percentile
grey	*	50	*
*	white	*	41st to 89th percentile
green	green	90	90th percentile and above

*Note: The grey percentile line on graphs is set to the 50th percentile to indicate typical grade-level performance. This line cannot be modified, but will be hidden if you set one of the other lines to the 50th percentile. There is no grey color-coding on other reports. Instead, the white color-coding represents the range between the yellow and green lines.

Figure 25: Percentile Line Thresholds

Percentile Line Thresholds

Put the green line at the 90 th percentile.
Put the yellow line at the 40 th percentile.
Put the red line at the 25 th percentile.

In the end, the outcomes from benchmark measures (BMs) displays the areas in which students are at risk of successfully learning subject specific skills and in what areas are Tier 2 and Tier 3 supports needed. The color codes reflect green with little risk, yellow with some risk, and red with considerable risk. Note that these levels are for each measure, allowing a composite decision to be adjusted according to teachers' professional development.

3.3c Documenting Risk Based on Individual Differences

Figure 4 displays student risk across multiple easyCBM® benchmark measures using a color-coded format. Each measure is shown separately, reinforcing that risk is domain-specific rather than global. This visualization supports instructional decisions by allowing educators to identify whether risk is driven primarily by decoding, fluency, vocabulary, or comprehension. Teachers can use this information to select interventions that directly target the area of need, while administrators can monitor patterns of risk across measures to guide curriculum planning. By highlighting measure-specific risk, the display also supports equitable decision-making, ensuring students receive supports matched to their instructional needs rather than being placed in broad interventions that may not address the underlying skill deficit. The figure supports Testing Standards 1.1 and 3.1 by strengthening construct-relevant interpretation and discouraging overgeneralization from a single score.

Figure 4. Display of Risk for Different Benchmark easyCBM®s (User's Manual, page 93)

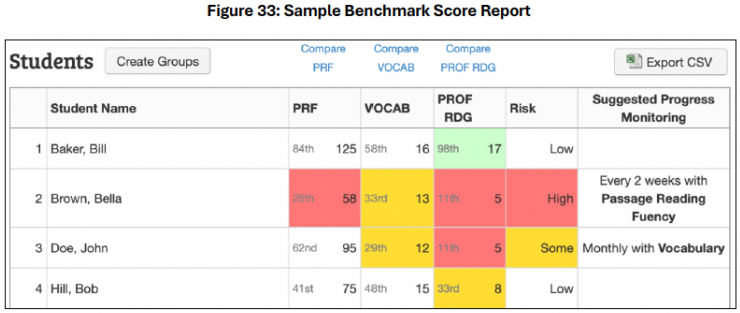


Figure 33: Sample Benchmark Score Report

Students | Create Groups | Compare PRF | Compare VOCAB | Compare PROF RDG | Export CSV

	Student Name	PRF		VOCAB		PROF RDG		Risk	Suggested Progress Monitoring
1	Baker, Bill	84th	125	58th	16	98th	17	Low	
2	Brown, Bella	25th	58	33rd	13	11th	5	High	Every 2 weeks with **Passage Reading Fuency**
3	Doe, John	62nd	95	29th	12	11th	5	Some	Monthly with **Vocabulary**
4	Hill, Bob	41st	75	48th	15	33rd	8	Low	

An important advancement in 2023 was the introduction of a comprehensive (composite) measure of reading in Grades using a composite score of three critical measures: Passage Reading Fluency, Vocabulary, and Proficient Comprehension Reading. **Figure 5** presents risk analysis using a composite reading score derived from Passage Reading Fluency, Vocabulary, and Proficient Comprehension. This display supports holistic decision-making by integrating multiple dimensions of reading into a single interpretive framework. Educators can use this display to justify comprehensive intervention decisions that reflect the multifaceted nature of reading. The primary decision supported is overall reading risk placement. By reducing reliance on a single measure, the composite aligns with Testing Standards 1.1 and 2.1, strengthening validity evidence through construct coverage.

Figure 5. Analyzing Risk in Composite Reading Performance (User's Manual, page 108)

Figure 49: Composite Report

	LN		LS		PS		Composite
Student Name	Score	Percentile	Score	Percentile	Score	Percentile	Percentile
Buckner, Flynn	50	98	35	99	23	91	98
Douglas, Dominique	1	6	0	22	0	34	15
Edwards, Uriah	54	98	25	95	29	95	97
Forbes, Haley	6	20	0	22	0	34	18
Franks, India	17	44	5	51	6	63	39
Glass, Armand	16	43	7	59	3	50	37

This risk analysis can also be used to measure change across successive BMs to determine if a just noticeable difference (JND) is occurring. **Figure 6** displays changes in risk for individuals within specific measures and across benchmark periods simultaneously. This integrated view supports coordinated MTSS decisions by revealing consistent or divergent patterns of risk.

Figure 6. Displaying Risk for Individuals across Measures and Benchmarks (User's Manual, page 101)

Figure 42: Benchmark Performance Report for Individual Students

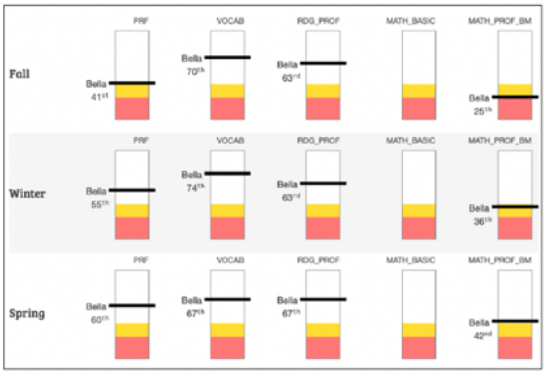
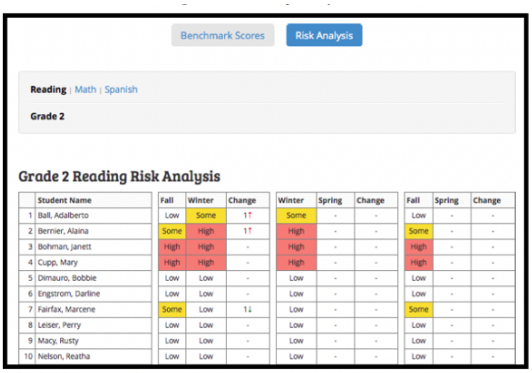


Figure 7 focuses on changes in student risk status across successive benchmark periods. Rather than emphasizing static classification, the display highlights movement over time, allowing educators to evaluate whether instruction is producing a just noticeable difference (JND). This visualization supports decisions about instructional effectiveness and responsiveness. If risk status improves, current supports may be continued or faded; if risk remains stable or increases, instructional intensity may need to be adjusted. The display also supports formative decision-making within MTSS, emphasizing that tier placement is provisional and responsive to student progress. From a validity standpoint, the figure aligns with Testing Standards 1.3 and 1.4 by reinforcing that interpretations should consider change and growth rather than single-point estimates.

Figure 7. Classroom Analysis of Risk and Change in Risk across Benchmarks (User's Manual, page 103)

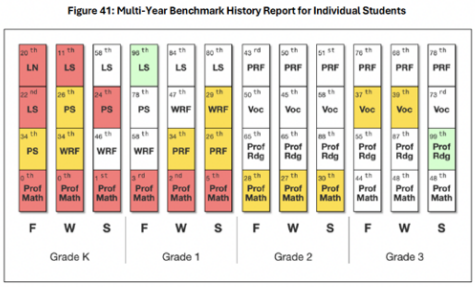


	Student Name	Fall	Winter	Change	Winter	Spring	Change	Fall	Spring	Change
1	Ball, Adalberto	Low	Some	1↑	Some	-	-	Low	-	-
2	Bernier, Alaina	Some	High	1↑	High	-	-	Some	-	-
3	Bohman, Janett	High	High	-	High	-	-	High	-	-
4	Cupp, Mary	High	High	-	High	-	-	High	-	-
5	Dimauro, Bobbie	Low	Low	-	Low	-	-	Low	-	-
6	Engstrom, Darline	Low	Low	-	Low	-	-	Low	-	-
7	Fairfax, Marcene	Some	Low	1↓	Low	-	-	Some	-	-
8	Leiser, Perry	Low	Low	-	Low	-	-	Low	-	-
9	Macy, Rusty	Low	Low	-	Low	-	-	Low	-	-
10	Nelson, Reatha	Low	Low	-	Low	-	-	Low	-	-

Another way to view the Risk Analysis is to summarize risk by Across Measures, Benchmarks, and Grades **(Figure 8)**, for individual measures across Benchmark and Grades numerically documenting change **(Figures 9)**, counts and percentages of students Changing Risk across BMs and Grades **(Figure 10),** and Comparing Schools **(Figure 11)** for making district decisions about resource allocations.

In moving from individual classrooms to individual measures over time, it is possible to adjust decision-making to skill specific resource allocation. Not only within year but across years, teachers can reflect on both the measures and the JNDs in which an individual difference is being made. In **Figures 8, 9, and 10**, the analysis at the grade level can also focus on different instructional programs in the aggregate aimed to ameliorate specific skill deficits.

Figure 8. Displaying Risk across Measures, Benchmarks, and Grades (User's Manual, page 100)



Using a tabular view of change, educators can determine whether individual support needs are persistent or transient across domains as displayed in **Figure 9**. These data can be displayed within a grade level using actual numerical values, which provide a more sensitive view of outcomes.

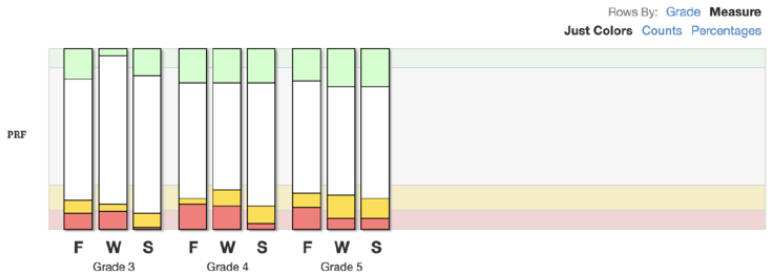
Figure 9. Displaying Risk within Measures and Grade Benchmarks (User's Manual, page 104)

Grade K Reading Risk Analysis

Risk Level	Fall	Winter	Change	Winter	Spring	Change	Fall	Spring	Change
Low	78%	63%	15%↓	63%	39%	24%↓	78%	39%	39%↓
Some	0%	20%	20%↑	20%	46%	26%↑	0%	46%	46%↑
High	22%	17%	5%↓	17%	15%	2%↓	22%	15%	7%↓
Totals	100%	100%	-	100%	100%	-	100%	100%	-

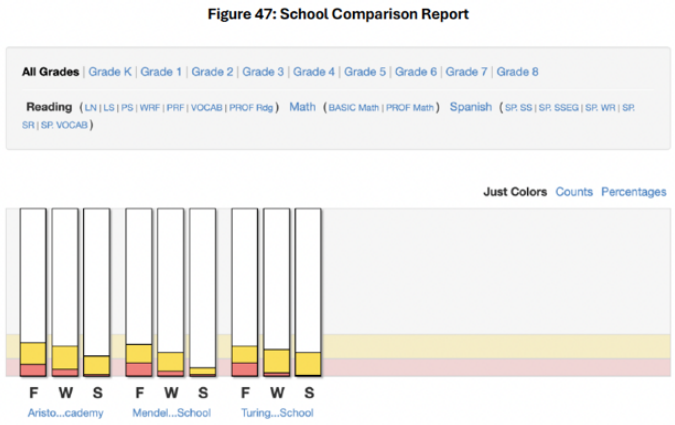
Figure 10 extends risk analysis for individual measures across benchmarks and grades, supporting a broader longitudinal reflection. The decision supported is whether instructional programs are producing sustained improvement. This figure aligns with Testing Standards 1.4 and 7.4 by supporting evaluation of long-term consequences.

Figure 10. Risk within Measures Across Benchmarks and Grade across (User's Manual, page 105)



Finally, **Figure 11** aggregates benchmark risk data at the school level, enabling comparisons across schools within a district. This display supports system-level decision-making related to resource allocation, professional development, and program evaluation. District leaders can identify grades and schools with consistently higher levels of risk and prioritize support accordingly. The figure also enables equity analyses by revealing whether risk patterns are distributed unevenly across schools. It emphasizes that assessment data can inform not only instruction but also systemic improvement efforts. Again, displays are available to document a just noticeable difference (JND) which allows movement across the Tiers 1 to 3 and the levels of support provided across schools. From a Standards perspective, this display aligns with Testing Standards 1.4 and 7.4 by supporting decisions whose consequences extend beyond individual students to organizational practices.

Figure 11. Comparisons of Schools at Risk across Benchmarks (User's Manual, page 106)



Once students' risk are ascertained, addressing them is covered in the last section (Section 3.4 – Progress Monitoring and Interventions), using an idiographic view that focuses on interventions to make an Individual Difference. In that section, we also describe professional development through formal courses that can be used for district-led certifications, blogs on practice posted on brtapps.com and Did You Knows (DYKs) translating research to practice and posted on easyCBM®.com

Appendix A: Technical Report Table and Figure Titles

Table 1. Validation Supporting Decision Making

_____

Figure 1. Reading Sub Scores Critical to the Overall Measure of Comprehension

Figure 2. Spanish Sub Scores Critical to the Overall Measure of Comprehension

Figure 3. Adjusting Benchmark (BM) Test Windows and Risk Values (User's Manual, pages 76-77)

Figure 4. Display of Risk for Different Benchmark easyCBM®s (User's Manual, page 93)

Figure 5. Analyzing Risk in Composite Reading Performance (User's Manual, page 108)

Figure 6. Displaying Risk for Individuals across Measures and Benchmarks (User's Manual, page 101)

Figure 7. Classroom Analysis of Risk and Change in Risk across Benchmarks (User's Manual, page 103)

Figure 8. Displaying Risk across Measures, Benchmarks, and Grades (User's Manual, page 100)

Figure 9. Displaying Risk within Measures and Grade Benchmarks (User's Manual, page 104)

Figure 10. Risk within Measures Across Benchmarks and Grade across (User's Manual, page 105)

Figure 11. Comparisons of Schools at Risk across Benchmarks (User's Manual, page 106)

References

easyCBM User's Manual (2025). Eugene, OR: University of Oregon – Behavioral Research and Teaching.

Conclusions Supporting Claims for Criterion 3.4: Student Progress

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of student growth or progress results.

This evaluation applies Criterion 3.4 (Indicators 3.4.a–3.4.c) to all sections of the EGCI_3.4 documentation and finds strong evidence that easyCBM® progress-monitoring reports and supporting resources appropriately facilitate valid interpretations and uses of student growth data.

Indicator 3.4.a: The design of progress reports is clearly aligned with intended interpretations and users. Reports are differentiated for educators, administrators, parents, and students, with grain sizes ranging from individual time-series displays to classroom, school, and district summaries. Visualizations such as percentile bands, goal and aim lines, benchmark-to-benchmark comparisons, and intervention phase markers directly support decisions about instructional effectiveness, tier movement, and intervention adjustment. The documentation demonstrates explicit attention to audience needs and includes cautions against common misuses (e.g., over-interpreting short or sparse time series, monitoring students on measures outside an appropriate instructional range). Conditions that may compromise interpretation—such as insufficient data points or inappropriate grade-level placement—are clearly articulated in user guidance.

Indicator 3.4.b: The system provides defensible representations of uncertainty in growth interpretation. Although numeric confidence intervals are not always displayed, error and variability are communicated through percentile bands, trend-line stability, overlap, and slope comparisons. Supporting text emphasizes the importance of measurement frequency, sufficient data density, and attention to variability, helping users avoid over-reacting to small or unstable score changes.

Indicator 3.4.c: extensive, research-based guidance supports appropriate use of progress data across the full performance continuum. Guidance is grounded in CBM theory, the *Standards for Educational and Psychological Testing*, and decades of empirical research. Detailed decision rules for selecting measures, adjusting grade levels, setting goals, and evaluating response to intervention are provided for reading, Spanish literacy, and mathematics. The integration of intervention documentation, professional development modules, blogs, and “Did You Know” resources further strengthens interpretive support and promotes coherent, defensible instructional decision making.

Overall, EGCI_3.4 demonstrates strong alignment among report design, treatment of error, and interpretive guidance, supporting valid, transparent, and instructionally meaningful use of student progress data.

3.4a Guidance in Progress Monitoring from User's Manual (pages 11-12)

The purpose of progress monitoring is to measure growth throughout the year for students who are identified as at risk, and to examine their response to intervention on the skill areas that are targeted for instruction. For the **Reading** measures, the progress monitoring forms are designed to be of equivalent difficulty to benchmark forms of the same grade level. For example, a Grade 2 Word Reading Fluency progress monitoring form is equivalent in difficulty to a Grade 2 Word Reading Fluency benchmark form. For **Basic Math and Proficient Math**, longer forms are used for benchmark testing and shorter forms for progress monitoring. You should use percentile ranks, rather than raw scores, to compare performance between benchmark and progress monitoring forms for the math measures.

The progress assessments and reports can be used to:

- Confirm **reading and math proficiency risk** levels at their respective grade (ranging from 'low risk' to 'high risk') and identify grade appropriate progress measures.
- **Identify** intervention or enrichment support for specific students (or groups of students) who may benefit.
- **Monitor the progress** of students during the academic year through both progress and interim benchmark testing (Fall, Winter, Spring).

The first step in determining which students would benefit from progress monitoring is to administer a grade-level benchmark assessment. If benchmark data are not available for the student because they transferred into between benchmark windows, grade-level progress monitoring forms are to be used from the same measures the student would have been tested on at the most recent benchmark assessment. For example, a Kindergarten student who transfers to school after the fall benchmark window but before the winter benchmark window could be given Letter Names, Letter Sounds, and Phoneme Segmenting progress monitoring forms.

Use the student's pattern of performance and professional judgment and knowledge of that student's skills to decide whether they require additional focused instruction and monitoring. Progress monitoring should be conducted using measures that assess the same skill or skills on which the student is receiving additional instruction or intervention. The best measures to use are those in which the student's scores fall within the 10th to 49th percentile range. That is the range in which the measures will be most sensitive to detecting growth as the student makes improvement. As a starting point, we provide preliminary progress monitoring suggestions on the Benchmark Scores Report. In most cases, we recommend focusing instruction and monitoring on a single skill. Note: For the early literacy skills, where students can make rapid progress when given focused instruction, teachers may wish to monitor two skills at a time.

If a student is scoring below the 10th percentile on grade-level materials, those materials may be too difficult for that student to accurately measure their progress. Teachers may need to use materials from earlier grades to identify an appropriate measure and grade level on which to monitor the student. easyCBM® assessments are built on a scale of progressive difficulty, with each grade level becoming more challenging. For the reading and Spanish literacy measures that assess individual skill areas, such as Word Reading Fluency and Passage Reading Fluency, those measures also increase in difficulty as they more closely approximate what could be considered reading (e.g., Passage Reading Fluency is more challenging than Word Reading Fluency).

For students performing below the 10th percentile on Passage Reading Fluency, drop back a grade or two at a time and try a passage from that lower grade. Repeat this process until teachers identify a grade level in which the student is reading between the 10th and 49th percentile in comparison to the norms for students of that grade at that time of year. If the student is reading below the 10th percentile on Grade 1 passages, or if their performance indicates they may have a deficit in earlier skills, you may need to use the early literacy measures (Letter Sounds or Phoneme Segmenting) to identify skill areas on which the student needs targeted instruction and monitoring. If the student is at or above the 50th percentile on the early literacy measures but below the 10th percentile on Passage Reading Fluency, they are likely to need intervention in fluency-building, and the Word Reading Fluency measure would be a good choice to avoid frustration while still monitoring developing fluency.

For a student who is **reading** below grade level on Passage Reading Fluency, the primary instructional focus should be on accurate and fluent reading. Once a student is reading fluently (nearing the 50th percentile on grade-level Passage Reading Fluency), the student probably has sufficient skill in fluency to start focusing more on comprehension and vocabulary.

Students being monitored on out-of-grade measures should be moved up to the next grade level once they are consistently performing at the 50th percentile on the lower-level material. Your goal is to assist the student in moving up in level as quickly as possible so they can catch up to grade-level material. Each student's trajectory is likely to be different and will depend on factors such as the student's initial skill level, the intensity of intervention provided, the ability to benefit from that intervention, motivation to improve, and attendance (the student must be present to benefit from instruction).

Students who are performing well below grade level may not be able to make up all that ground within a single school year. Teachers should make as much progress as possible with those students, with the intention that they will receive support and continue to make progress in subsequent years.

Table 1. Frequency of Progress Monitoring Measurement (User's Manual, page 13)

How often to monitor progress depends on two key questions:
• How quickly is it reasonable to expect to see growth in a particular skill area?
• How much actual intervention has the student received?

The table below provides guidelines on the frequency of progress monitoring for the different measures. To see growth, it's important that students are given focused instruction to address their skill deficits.

Table 1. Frequency of Progress Monitoring Measurement (User's Manual, page 13)

Frequency Guidelines for Administering Progress Monitoring Measures

Measure	Frequency	Comments
Phoneme Segmenting Letter Sounds	Every 1–2 weeks	Students can make rapid progress in these skill areas when they receive in-depth interventions to help accelerate their learning.
Word Reading Fluency Passage Reading Fluency	2 weeks	Students typically take longer to improve in these skill areas.
Vocabulary Basic and Proficient Reading Basic and Proficient Math	3–4 weeks	Students should be tested no more frequently than every three to four weeks with these measures because it takes longer for improvements to take hold and be reflected on test performance.

The following table indicates the grade levels in which progress monitoring forms are available for each of the measures. Some measures have progress monitoring forms for grades in which they are not used for benchmark testing, to support students who are still working on those skills.

Table 2. Content of Reading Progress Monitoring Measurement (User's Manual, page 14)

easyCBM Reading Measures Available for Progress Monitoring by Grade

Grade	Letter Names	Letter Sounds	Phoneme Segmenting	Word Reading Fluency	Passage Reading Fluency	Vocabulary	Basic Reading	Proficient Reading
K	✓	✓	✓	✓				
1	✓	✓	✓	✓	✓			
2				✓	✓	✓		✓
3				✓	✓	✓	✓	✓
4					✓	✓	✓	✓
5					✓	✓	✓	✓
6					✓	✓	✓	✓
7					✓	✓	✓	✓
8					✓	✓	✓	✓

The **Reading and Spanish** progress monitoring forms mirror the benchmark forms for each measure, but for math, the progress monitoring forms are shorter than the benchmark forms. The Basic Math progress monitoring forms take about 8 minutes to complete and Proficient Math progress monitoring forms take about 20 minutes to complete.

3.4b and 3.4c Using easyCBM® to Make an Individual Difference

With easyCBM®, we approach the validation process from two perspectives: (a) nomothetic and (b) idiographic. Dr. Stan Deno, the key person who founded Curriculum-Based Measurement (CBM) referred to these two perspectives as appreciating individual differences (nomothetic with a view on distributions of students) versus making an individual difference (an idiographic view with a focus on progress over time)[1].

With both perspectives, the emphasis on validation is about the decisions, not the measures. "It is the interpretation of test score for proposed uses that are evaluated, not the test itself…each intended interpretation must be evaluated. Statements about validity should refer to particular interpretations for specified uses. It is incorrect to the unqualified phrase "the validity of the test" (p. 11)[2].

This section is organized with these two perspectives with the initial focus on the nomothetic view and reports used to display individual differences. Then, we focus on interventions that can affect both perspectives: Which students are receiving various levels of support and the effects within and across the school year. Finally, to ensure the integrity of the decision-making process, we address professional development through both formal courses that can be used for district-led certifications and through a series of Did You Know (DYKs) offered monthly as a resource with easyCBM®, connecting research to practice. We also offer a series of blogs that are posted on brtapps.com. Note that Individual Differences is in Section 3.3 and Individual Difference is in Section 3.4.

As displayed in See 3.3b (Using easyCBM® to Document Individual Differences), Figure 2. Display of Risk for Different Benchmark easyCBM®s (User's Manual, page 93), teachers can assign students to Tiers of Support (I to III). In the remainder of this document, we address the Progress Monitoring (PM) systems that allows assignment

[1] Deno, S. L. (1990). Individual Differences and Individual Difference: The Essential Difference of Special Education: The Essential Difference of Special Education. *The Journal of Special Education*, *24*(2), 160-173. https://doi.org/10.1177/002246699002400205.

[2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC.

of measures for documenting more specific achievement over time, eventually used to guide instructional adjustments in making an individual difference (see Section 3.4).

In **Figure 1** below, performance proficiency levels for early (Fall) Kindergarten that are low across multiple measures (below the 30th PR as indicated by red cells) indicate PMs for Letter Sounds (LS) and Phoneme Segmenting (PS). Only Phoneme Segmenting PM is indicated if this measure is low but either, or both, of the other measures are medium or high. Letter Sounds PMs are recommended if proficiency levels are low in it. If any of the three measures are high in proficiency values, no follow-up PMs are recommended.

Figure 1. Early Reading Progress Monitoring for Students in Early Kindergarten (User's Manual, page 116)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Reading: Fall of Kindergarten

LN	LS	PS	Monitoring Suggestion
low	low	low	LS, PS
medium	low	low	LS, PS
high	low	low	LS, PS
low	medium	low	PS
medium	medium	low	PS
high	medium	low	PS
low	high	low	-
medium	high	low	-
high	high	low	-
low	low	medium	LS, PS
medium	low	medium	LS, PS
high	low	medium	LS, PS
low	medium	medium	PS
medium	medium	medium	PS
high	medium	medium	PS
low	high	medium	-
medium	high	medium	-
high	high	medium	-
low	low	high	LS
medium	low	high	LS
high	low	high	LS
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

The same logic applies in later Kindergarten (Winter and Spring) and early (Fall) Grade 1 on early reading measures of Letter Sounds (LS) and Phoneme Segmenting (PS). With these benchmark periods, however, the introduction of Word Reading Fluency (WRF) reflects the need for PMs if it is also not at proficient levels (e.g., is below the 30th PR). Again, if performance levels are varied across measures or any of them are at high levels of proficiency, then no PMs are recommended.

Figure 2. Early Reading Progress Monitoring for Students in Kindergarten and Grade 1 (User's Manual, page 117)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Reading: Winter and Spring of Kindergarten, Fall of Grade 1

LS	PS	WRF	Monitoring Suggestion
low	low	low	LS, PS
medium	low	low	PS
high	low	low	PS
low	medium	low	LS
medium	medium	low	WRF
high	medium	low	WRF
low	high	low	LS
medium	high	low	WRF
high	high	low	WRF
low	low	medium	-
medium	low	medium	-
high	low	medium	-
low	medium	medium	-
medium	medium	medium	-
high	medium	medium	-
low	high	medium	-
medium	high	medium	-
high	high	medium	-
low	low	high	-
medium	low	high	-
high	low	high	-
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

As more and different measures are added into the mix for later (Winter and Spring) of Grade 1, the performance/proficiency levels recommended for PMs shift. Of course, if students are low in letter sounds (LS), PMs in that skill are recommended, particularly if this skill is confirmed by Word Reading Fluency (WRF) and Passage Reading Fluency (PRF). Otherwise, PMs are recommended if proficiency levels of WRF are low along with Passage Reading Fluency (PRF). Otherwise, if PRF proficiency levels are low, follow-up PMs are recommended. Again, Again, if performance levels are varied across measures or any of them are at high levels of proficiency, then no PMs are recommended.

Figure 3. Reading Progress Monitoring for Students in Late Grade 1 (User's Manual, page 118)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Reading: Winter and Spring of Grade 1

LS	WRF	PRF	Monitoring Suggestion
low	low	low	LS
medium	low	low	WRF
high	low	low	WRF
low	medium	low	PRF
medium	medium	low	PRF
high	medium	low	PRF
low	high	low	PRF
medium	high	low	PRF
high	high	low	PRF
low	low	medium	-
medium	low	medium	-
high	low	medium	-
low	medium	medium	-
medium	medium	medium	-
high	medium	medium	-
low	high	medium	-
medium	high	medium	-
high	high	medium	-
low	low	high	-
medium	low	high	-
high	low	high	-
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

By Grade 2, the critical measures include Passage Reading Fluency (PRF), Vocabulary (Voc), and Proficient Reading (PR). Determination of PMs follows the privilege of these measures in words before passages (if both are low). PRF progress monitoring is recommended if it is low. Vocabulary PMs are recommended if students' proficiency is low in either or both PRF or Voc. As in earlier grades, if performance levels across measures are at medium or high levels of proficiency, then no PMs are recommended.

Figure 4. Reading Progress Monitoring for Students in Grade 2 (User's Manual, page 119)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Reading: Grade 2

PRF	Vocab	Proficient Reading	Monitoring Suggestion
low	low	low	WRF
medium	low	low	PRF
high	low	low	Vocab
low	medium	low	PRF
medium	medium	low	Proficient Reading
high	medium	low	Proficient Reading
low	high	low	PRF
medium	high	low	Proficient Reading
high	high	low	Proficient Reading
low	low	medium	PRF
medium	low	medium	Vocab
high	low	medium	Vocab
low	medium	medium	PRF
medium	medium	medium	-
high	medium	medium	-
low	high	medium	PRF
medium	high	medium	-
high	high	medium	-
low	low	high	Vocab
medium	low	high	Vocab
high	low	high	Vocab
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

The remaining grades (3 through 8) utilize the following recommendations for proficiency levels with three measures: Passage Reading Fluency (PRF), Vocabulary, and Proficient Reading (PR). If either PRF and/or Vocabulary is low, then PMs are recommended in the measure(s) that are low. If the proficiency level of Basic Reading is low, then teachers are recommended to monitor progress with this measure if either of the other two measures are medium levels. If either of the other two measures is high and Proficient Reading is low, then PMs should be used with it.

Figure 5. Reading Progress Monitoring for Students in Grades 3 – 8 (User's Manual, page 120)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Reading: Grades 3 to 8

PRF	Vocab	Proficient Reading	Monitoring Suggestion
low	low	low	PRF
medium	low	low	Vocab
high	low	low	Vocab
low	medium	low	PRF
medium	medium	low	Basic Reading
high	medium	low	Proficient Reading
low	high	low	PRF
medium	high	low	Basic Reading
high	high	low	Proficient Reading
low	low	medium	PRF
medium	low	medium	Vocab
high	low	medium	Vocab
low	medium	medium	PRF
medium	medium	medium	-
high	medium	medium	-
low	high	medium	PRF
medium	high	medium	-
high	high	medium	-
low	low	high	Vocab
medium	low	high	Vocab
high	low	high	Vocab
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

With the Spanish measures for early (Fall) KIndergarten, the receommendations for PMs simply reflect the skill area (Syllable Segmenting [SSEG] and Syllable Sounds [SS]) in which students show low proficiency levels, particualry if confirmed by the other measure.

Figure 6. Spanish Progress Monitoring for Early Kindergarten Students (User's Manual, page 122)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Spanish: Fall of Kindergarten

SSEG	SS	Monitoring Suggestion
low	low	SSEG
medium	low	SS
high	low	SS
low	medium	SSEG
medium	medium	SS
high	medium	SS
low	high	-
medium	high	-
high	high	-

Spanish: Winter and Spring of Kindergarten

SSEG	SS	Monitoring Suggestion
low	low	SSEG
medium	low	SS
high	low	SS
low	medium	-
medium	medium	-
high	medium	-
low	high	-
medium	high	-
high	high	-

In early Grade 1 (Fall), similar recommendations are made for SSEG and SS. PMs are recommended for Word Reading (WR), however, in Grade 1, when students show low proficiency levels, irrespective of either/both of the other measures.

Figure 7. Spanish Progress Monitoring for Early Grade 1 Students (User's Manual, page 123)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Spanish: Fall of Grade 1

SSEG	SS	WR	Monitoring Suggestion
low	low	low	SSEG
medium	low	low	SS
high	low	low	SS
low	medium	low	SSEG
medium	medium	low	WR
high	medium	low	WR
low	high	low	WR
medium	high	low	WR
high	high	low	WR
low	low	medium	-
medium	low	medium	-
high	low	medium	-
low	medium	medium	-
medium	medium	medium	-
high	medium	medium	-
low	high	medium	-
medium	high	medium	-
high	high	medium	-
low	low	high	-
medium	low	high	-
high	low	high	-
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

In later (Winter and Spring) of Grade 1, Syllable Sounds (SS) are complemented by Word Reading (WR) and Sentence Reading (SR). PM recommendations privilege syllables before words and words before sentences: When SS proficiency is low, PM is recommended. If WR proficiency is low, PMs are required. If SR is low, PMs are recommended, particularly if WR proficiency is low. If performance is high in any of these areas, the recommendation for PMs in each measure are qualified.

Figure 8. Spanish Progress Monitoring for Late Grade 1 Students (User's Manual, page 124)

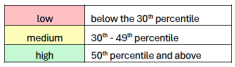
low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Spanish: Winter and Spring of Grade 1

SS	WR	SR	Monitoring Suggestion
low	low	low	SS
medium	low	low	WR
high	low	low	WR
low	medium	low	SR
medium	medium	low	SR
high	medium	low	SR
low	high	low	SR
medium	high	low	SR
high	high	low	SR
low	low	medium	-
medium	low	medium	-
high	low	medium	-
low	medium	medium	-
medium	medium	medium	-
high	medium	medium	-
low	high	medium	-
medium	high	medium	-
high	high	medium	-
low	low	high	-
medium	low	high	-
high	low	high	-
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

In Grade 2, PMs are recommended when performance in that measure is low, particularly if confirmed by low performance in the other two measures: Word Reading (WR), Sentence Reading (SR), and Vocabulary.

Figure 9. Spanish Progress Monitoring for Grade 2 Students (User's Manual, page 125)



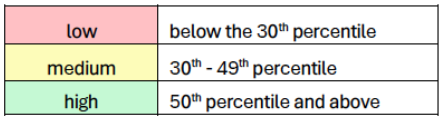
low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Spanish: Grade 2

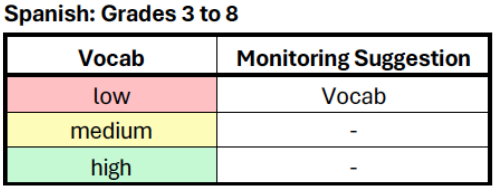
WR	SR	Vocab	Monitoring Suggestion
low	low	low	WR
medium	low	low	SR
high	low	low	SR
low	medium	low	Vocab
medium	medium	low	Vocab
high	medium	low	Vocab
low	high	low	Vocab
medium	high	low	Vocab
high	high	low	Vocab
low	low	medium	WR
medium	low	medium	SR
high	low	medium	SR
low	medium	medium	-
medium	medium	medium	-
high	medium	medium	-
low	high	medium	-
medium	high	medium	-
high	high	medium	-
low	low	high	WR
medium	low	high	SR
high	low	high	SR
low	medium	high	-
medium	medium	high	-
high	medium	high	-
low	high	high	-
medium	high	high	-
high	high	high	-

For Grades 3 through 8, PMs are recommended only using the Vocabulary measures if proficiency levels are low.

Figure 10. Spanish Progress Monitoring for Grade 3 – 8 Students (User's Manual, page 126)



low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above



Spanish: Grades 3 to 8

Vocab	Monitoring Suggestion
low	Vocab
medium	-
high	-

Notice that, in this progression from Figure 1 to Figure 11, with their focus on nomothetic values, the group has iterated between district level (e.g., establishing risk) to successive aggregations from Classrooms and Schools, Measures-Grades-Benchmarks, as well as comparing schools**.**

Yet, the sine qua non of easyCBM® is to evaluate instruction with the first requirement being a display of a time series on progress measures to determine the rate of change. This line graph reflects the progress of each student and with the eventual introduction of interventions, can be used to make evaluations of effectiveness. In these judgments, several different dimensions of progress can be used: (a) relative to the district PRs, (b) relative to the goal and subsequent aim lines, and (c) relative to data characteristics of the time series (changes in level, slope, variation, and overlap). This next section presents graphic displays that increasingly complexify the analysis, reflecting an idiographic approach.

Figure 11 displays individual student progress relative to percentile bands over time. This time-series visualization allows educators to evaluate whether a student's rate of improvement is sufficient compared to peers. The decision supported is whether current instruction is adequate to close achievement gaps. Alignment with Standards 1.3 and 4.10 is evident in the emphasis on growth trajectories rather than static performance.

Figure 11. Monitoring Student Progress Relative to Percentile Bands (User's Manual, page 97)

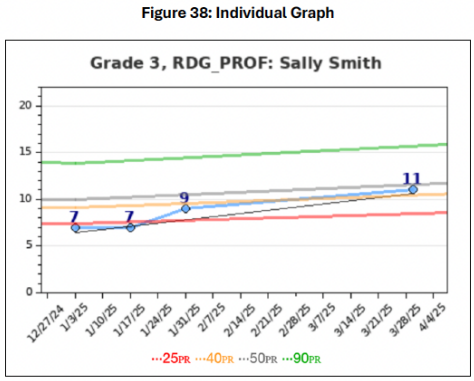
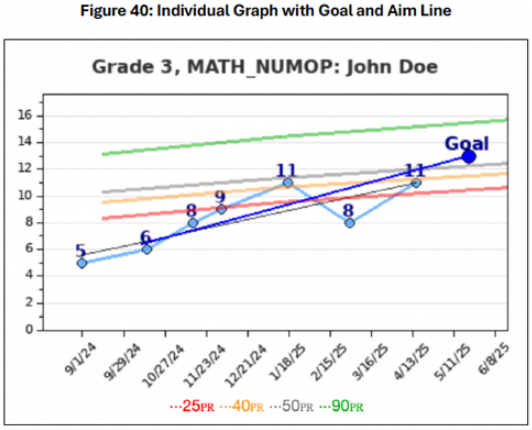


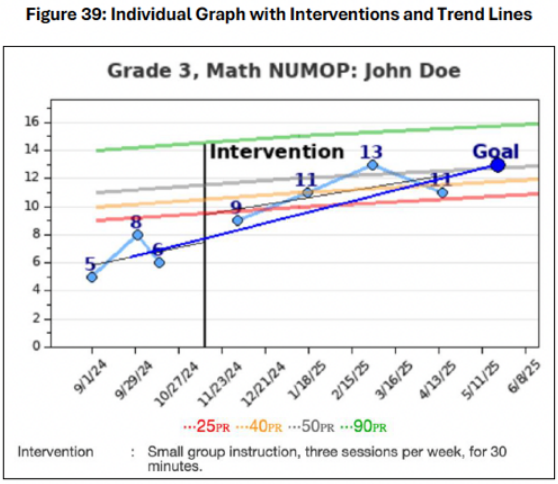
Figure 12 displays change over time relative to instructional goals and the subsequent aim lines displayed to achieve these goals. Notice how decision making can be advanced by ensuring changes are timely (noting the date and level for final progress to be made) and made relative to the variation and slope in attaining this goal. Figure 11 enhances progress monitoring by adding goal lines and aim lines. This visualization integrates two important metrics—trend and goal attainment—to support instructional decision-making. Educators can determine whether observed growth is on track to meet expected outcomes. This figure aligns with Standards 4.1 and 5.1 by supporting consistent interpretation of progress data.

Figure 12. Monitoring Student Progress Relative to Goals and with Aim Lines (User's Manual, page 98)



Finally, the essence of decision-making is impact with the introduction of interventions as displayed in **Figure 13**. And with this interruption in the time series, further tools can be added in the decision-making arsenal: level and overlap. Changes in level refer to the immediate shift in performance following the introduction of the intervention and overlap refers to the data values in common after the intervention with those before it (fewer is better).
The emphasis is on instructional impact, examining changes in level and overlap between baseline and intervention phases. The decision supported is whether instruction is producing meaningful change. This display aligns with Standards 1.4 and 2.1 by focusing on the consequences of instructional decisions rather than student deficit.

Figure 13. Evaluating Instruction Based on Student Progress and Goal Attainment (User's Manual, page 98)



In summary, teachers can make an individual difference by using any of several metrics in adjusting instruction (described below in terms of importance):

- Compare the trend line (which reflects the slope of improvement to the aim line to (which is set according to the end-of-year goal). Obviously, if they diverge with the trend not intersecting with the aim (and not reaching the goal), instruction should be adjusted.
- Compare the effects attained immediately after the introduction of an instructional adjustment to the last data point before the adjustment. This metric is referred to as a change in level.
- Use overlap to compare the range of values (from lowest to highest) before the introduction of the instructional adjustment to the range of values (from lowest to highest) after the introduction of the instructional adjustment. If overlap is great, the instructional adjustment is marginally effective and should be changed.
- Evaluate the slope of progress alone to what is expected. Slope refers to 'rise over run' and when it is steep (and positive), student performance is changing quickly, so teachers should stay the course. In contrast, when the slope is low (almost flat), instructional adjustments are warranted.
- Consider student variation (similar to overlap): Are students performing with minimum variation from data point to data point? If so, their responses to both instruction and assessment appear to be under control. In contrast, if wild swings (up or down) are apparent, then student performance is NOT under control and adjustments may be needed in both the manner the assessment is administered as well as the instructional components that being implemented.

To cap off the these displays of standing for students, a parent report generates a combination of visual displays for their child, providing a rationale for all decisions being made, from placement to treatment. **Figure 14** presents parent-facing reports that integrate multiple visual displays for an individual student. This visualization supports transparent communication and shared decision-making with families. By presenting data clearly, the display aligns with Standards 6.1 and 7.1, ensuring interpretations are accessible and defensible.

Figure 14. Parent Reports Showing Several Different Outcomes (User's Manual, page 102)

Figure 43: Parent Report

easyCBM District Edition

Student Progress Report : As of September 16th, 2025

Student	Aguilar, Harlan
StudentId	862176
Grade	1

Letter Names : LN
Focus on identification skills: letters

Letter Sounds : LS
This is a 60-second test in which the student says variety of letter sounds and digraphs by sight.

Phoneme Segmenting : PS
Focus on blend transfer skills: sounds --> blends

Word Reading Fluency : WRF
This is a 60-second timed test in which the student reads a variety of grade appropriate words.

Passage Reading Fluency : PRF
This is a 60-second timed test in which the student reads a short fictional story

Vocab : VOCAB
Focus on building understanding: increasing lexicon

Proficient Reading
Focus on meaning transfer skills: passage fluency --> deeper comprehension

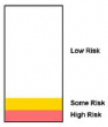
Basic Reading
Focus on meaning transfer skills: passage fluency --> literal comprehension

Basic Math
Focus on building basic understanding and fluency with numbers and math facts/concepts

Proficient Math
Focus on building deeper mathematical understanding

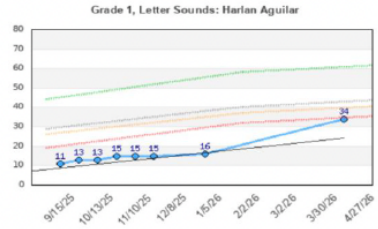
Percentile Rank
The percentage of scores that are equal to or less than a given score. Percentile ranks, like percentages, fall on a continuum from 0 to 100. For example, a test score that is greater than or equal to 75% of the scores of people taking the test is at the 75th percentile rank.

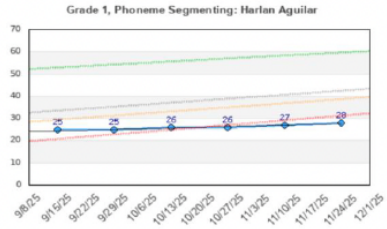
Legend/Key: District Risk Settings

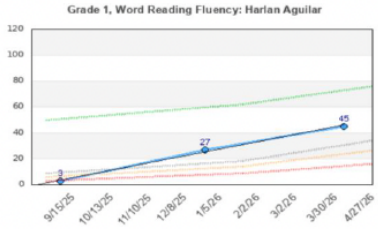


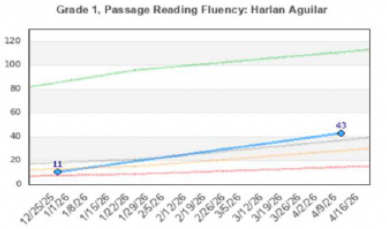
Reading Assessments

	Fall		Winter		Spring	
	score	% ile	score	% ile	score	% ile
LS	11	13	16	7	34	22
PS	25	34				
WRF	3	26	27	66	45	58
PRF			11	31	43	49
Composite	21		22		39	



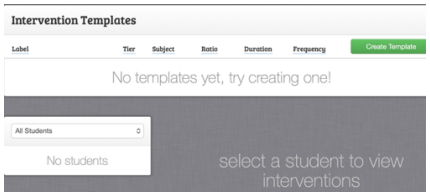
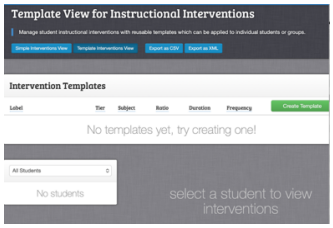
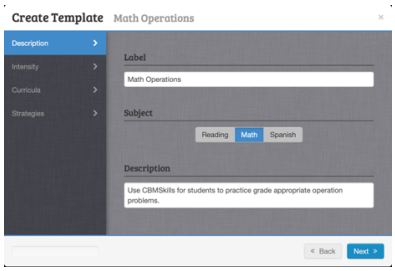
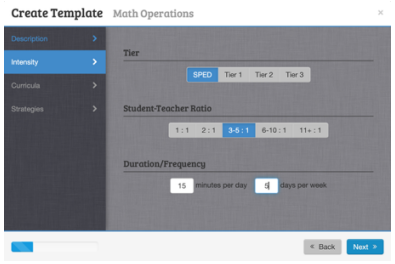
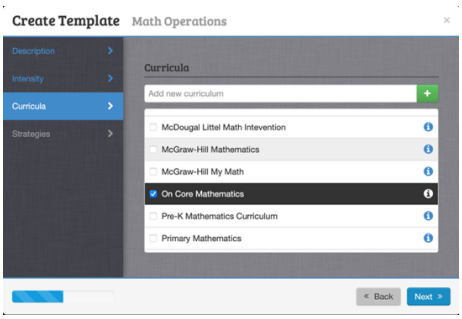


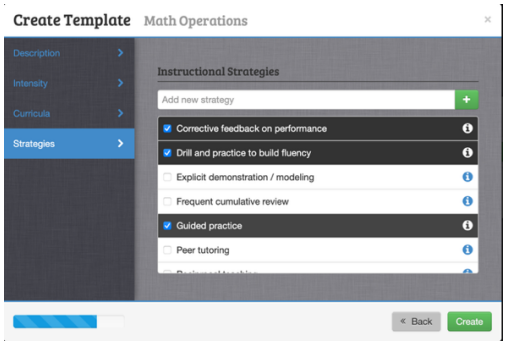
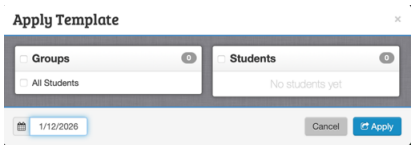




Interventions

The interventions template allows teachers to provide information on the Tier of instruction (1 to 3), the subject area for teaching, the grouping arrangement (which allows tracking the student to teacher ratio), the duration or length of time for proving specialized supports, and the frequency for delivering this package, which also includes a specific curriculum and teacher strategies. In addition, a template can be created to make the process efficient for use with other students.

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>Step 1. Both a label and brief description of the intervention are entered as the initial information to input. Note that the brief description is displayed on the progress (time series) graph.</td>
<td> 
</td>
</tr>
<tr>
<td>Step 2 in creating the template provides information on the intensity of the intervention, which is defined as the Tier (Special Education or Tiers 1 to 3) as well as the ratio of students to teachers and the number of minutes and number of days.</td>
<td> 
</td>
</tr>
<tr>
<td>In Step 3, teachers can either select from many different curricula that have been pre-loaded into the system or they can add their own. Note: An important option allows administrators to pre-define a required curriculum for the school or district.</td>
<td> 
</td>
</tr>
</table>

Step 4 is essentially the last step for defining the intervention. A series of strategies or instructional procedures can be selected or entered if the bank of extant straggles is insufficient. Note that these strategies can be combined in various combinations.	
Given this instructional diet, teachers have the option of applying it to specific groups; to do so, they can either load the students' names from a roster or add them individually (with names separated by a comma).	

These interventions can be considered an operationalization of '*active engaged time*'. For a typical reference, see: Christenson, S. L, Reschly, A. L., & Wylie, C. (2012). *Handbook of Research on Student Engagement*. Springer.

Professional Development Modules with easyCBM®

easyCBM Teacher's Manuals & Documentation

We have a Manual available, and we also provide a Getting Started guide.

In addition, our Help system includes many answers to common issues including these Frequently Asked Questions.

Professional Development Courses

All easyCBM account holders have free access to a variety of professional development courses, available on Thinkific, a learning management system.

To access the courses, simply click on the course links below. You will be taken to the Thinkific enrollment page.

Create an account for yourself in Thinkific by registering with your email and a password you select. You will use this same email/password whenever you visit the easyCBM courses.

Please note: We will not have access to your Thinkific account email/password; it's important that you select sign-in credentials that you will remember because we will be unable to reset them for you.

Reading	Spanish	Math
Using the Early Literacy Measures **Learn about the early literacy measures (Letter Names, Letter Sounds, and Phoneme Segmenting).** This course includes background information on the three measures as well as instruction on how to administer and score each measure. Built-in proficiency quizzes give you the opportunity to document your mastery of the content. • Recommended for anyone who will be administering and scoring the Letter Names, Letter Sounds, or Phoneme Segmenting measures.	Using the Spanish Reading Measures **Learn the key features of the easyCBM Spanish Literacy Measures, which include Syllable Sounds, Syllable Segmenting, Word and Sentence Reading Fluency, and Vocabulary.** The course includes background information on each of the measure types and ends with a quiz where you can assess your learning. • Recommended for anyone who will be administering and scoring any of the easyCBM Spanish Literacy assessments.	Using the Math Measures **This course covers the two different easyCBM mathematics assessments: Basic and Proficient Math.** It includes information about the development of each of the measures, guidance on standardized administration protocols, and a proficiency quiz to demonstrate your understanding of the measures. • Recommended for anyone who will be administering and scoring the easyCBM Basic or Proficient Math measures.

Decision Making with easyCBM

This course focuses on using easyCBM to make decisions. It includes identifying essential features of CBM, using student performance to make identification decisions, and finally, using student progress to make instructional decisions.

- Recommended for educators who need to interpret student performance on CBMs.

Advanced Decision Making: Using Data for RTI/MTSS

Unlock the potential of easyCBM with engaging lessons designed to enhance teaching strategies and improve student literacy outcomes through data-driven decision-making. **This course is ideal for those who enjoy a story-based approach to learning.** Meet our two teacher characters in the first lesson and follow along as they discuss their students' easyCBM literacy scores and different evidence-based instructional strategies they might use to meet their students' identified needs.

- Recommended for teachers, school psychologists, and para-educators tasked with supporting student learning.

Blogs to Support Effective Teaching and Assessment Practices

The focus of blogs is on the practice of teaching and learning assessments.

- easyCBM® Provides Three References for Interpreting Student Performance – Dr. Gerald Tindal, Published on: August 13, 2024
- Post-Pandemic, The Science Of Reading Is Still Clear – Dr. Leilani Sáez, Published on: September 3, 2024
- What to do about Tricky Words – Dr. Leilani Sáez, Published on: August 26, 2024
- The Critical Role of Phonological Sensitivity – Dr. Leilani Sáez, Published on: August 19, 2024
- What is the Grain Size of your Phonics Strategies? –Dr. Leilani Sáez, Published on: August 7, 2024
- The Power in Teachers Thinking Aloud –Dr. Leilani Sáez, Published on: July 29, 2024
- Introduction to CBMSkills: November 2023 Webinar Recording – Dr. Gerald Tindal, Published on: November 15, 2023
- Writing the Write Way – Dr. Gerald Tindal, Published on: October 30, 2023
- Learning Progressions through Diagnostic Assessments – Dr. Gerald Tindal, Published on: October 30, 2023
- Learn about CBMSkills this November – Dr. Gerald Tindal, Published on: October 25, 2023
- University of Oregon Launches Voice-Enabled Assessment to Seamlessly Understand How Early Readers Are Progressing – Dr. Gerald Tindal, Published on: September 13, 2023
- Oral Reading Fluency Research – Dr. Gerald Tindal, Published on: September 5, 2023
- Phonemic Awareness, Phonics, and Fluency Instruction – Dr. Gerald Tindal, Published on: September 5, 2023
- Oral Reading Fluency Practices with Automatic Speech Recognition – Dr. Gerald Tindal, Published on: September 5, 2023
- Oral Reading Fluency (ORF) Norms – Dr. Gerald Tindal, Published on: September 1, 2023
- The Writing Process: Sloppy Copy, Outline, Draft, and Final Essay – Dr. Gerald Tindal, Published on: May 8, 2023
- Finding and Fixing Writing Prompts – Dr. Gerald Tindal, Published on: May 8, 2023

Did You Knows (DYKs) to Support Research-Based Teaching and Assessment Practices

The focus on DYKs is on applications of research to practice so teachers can trust the outcomes.

Accommodations

CBM Uses

CBM Research

CBMSkills Math Diagnosis

Classification Accuracy in Math

Classification Accuracy in Composite Reading

Criterion Validity

Early Reading

Inferences of Curriculum Differences

Normal Distributions

Appendix A: Technical Report Table and Figure Titles

Table 1. Frequency of Progress Monitoring Measurement (User's Manual, page 13)

Table 2. Content of Reading Progress Monitoring Measurement (User's Manual, page 14)

Figure 1. Early Reading Progress Monitoring for Students in Early Kindergarten (User's Manual, page 116)

Figure 2. Early Reading Progress Monitoring for Students in Kindergarten and Grade 1 (User's Manual, page 117)

Figure 3. Reading Progress Monitoring for Students in Late Grade 1 (User's Manual, page 118)

Figure 4. Reading Progress Monitoring for Students in Grade 2 (User's Manual, page 119)

Figure 5. Reading Progress Monitoring for Students in Grades 3 – 8 (User's Manual, page 120)

Figure 6. Spanish Progress Monitoring for Early Kindergarten Students (User's Manual, page 122)

Figure 7. Spanish Progress Monitoring for Early Grade 1 Students (User's Manual, page 123)

Figure 8. Spanish Progress Monitoring for Late Grade 1 Students (User's Manual, page 124)

Figure 9. Spanish Progress Monitoring for Grade 2 Students (User's Manual, page 125)

Figure 10. Spanish Progress Monitoring for Grade 3 – 8 Students (User's Manual, page 126)

Figure 11. Monitoring Student Progress Relative to Percentile Bands (User's Manual, page 97)

Figure 12. Monitoring Student Progress Relative to Goals and with Aim Lines (User's Manual, page 98)

Figure 13. Evaluating Instruction Based on Student Progress and Goal Attainment (User's Manual, page 98)

Figure 14. Parent Reports Showing Several Different Outcomes (User's Manual, page 102)

References

easyCBM User's Manual (2025). Eugene, OR: University of Oregon – Behavioral Research and Teaching.

Note

Page 1 of the Introduction to Technical Report 2604-IAR appears in Tindal, G. (2026) *Overview of 2026 Series Summarizing easyCBM® Research (Technical Report 2603-SUM)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

All the Technical Reports in 2604-IAR are published on the BRT website (https://brtprojects.org). See Swanson, D., & Tindal, G. (2024). *An authoritative bibliography of technical adequacy research conducted on easyCBM – 2024 (Technical Report # 2402.3)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., & McCaslin, S. (2026c). *Test Development for easyCBM® in Grades K-8: Reading (Technical Report # 2603-TDK8R)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026d). *Test Development for easyCBM® in Grades K–8: Mathematics (Technical Report # 2603-TDK8M).* Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026a). *Reliability of easyCBM® in Grades K-8: Mathematics (Technical Report # 2603-RK8M)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026b). *Reliability of easyCBM® in Grades K-8: Reading (Technical Report # 2603-RK8R)*. Eugene, OR.: Behavioral Research and Teaching, University of Oregon.

Tindal, G., & McCaslin, S. (2026e). *Validity Analyses for easyCBM® in Grades K-8: Mathematics (Technical Report # 2603-VK8M)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026f). *Validity Analyses for easyCBM® in Grades K-8: Reading (Technical Report # 2603-VK8R)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G. (2026). *Alignment of easyCBM® with Standards in Grades K-8: Reading and Mathematics (Technical Report # 2603-A38RM)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

These reports reflect the integration of technical adequacy research on easyCBM® conducted over the past 25 years by researchers at *Behavioral Research and Teaching* (BRT – https://brtprojects.org). By integrating this research, reviewers can more easily make informed judgments on adoption of various measures. Each report summarizes the original research, though each study is also available on the BRT website.

The gateways-criteria-indicators are provided for English Language Arts (ELA) by the Center for Assessment and EdReports.org (2023): Center for Assessment, & EdReports.org. (2023). *Review Criteria Interim Assessment English Language Arts Grades 3-8 (v1.0, Final 5/2023).* Dover, NH: EdReports.org and Center for Assessment

Gateway 1: Alignment, Fairness, and Accessibility Criteria and Indicators

Criterion 1.1: Test Development Alignment (8 points).

Focus: design specifications and blueprints that drive high-quality, standards-aligned development.

- Indicator 1.1.a (0/2/4): Design specifications provide clear expectations and detailed guidance. Evidence includes a rationale and research foundation; robust item development documentation; clear scoring rules/rubrics across item types; processes to ensure content accuracy and editorial/technical quality; and specified cognitive-demand ranges sufficient to measure the depth of standards.

- Indicator 1.1.b (0/2/4): Blueprints/specifications emphasize the most important content. For Grades K-8 this means score points concentrate on the "major work" of the standards; for high school courses (if applicable) the distribution reflects the content and skills needed for college and careers.

Criterion 1.2: Item and Form Alignment (16 points)

Focus: alignment of actual items and delivered test events to the standards and intended design.

- Indicator 1.2.a (0/2/4): Delivered forms/events reflect an appropriate distribution of content, score points, and item types—strongly focused on major and supporting clusters (or equivalent priority content).

- Indicator 1.2.b (0/1/2): Items elicit evidence of learning relative to one or more standards without measuring irrelevant knowledge/skills; items align to the design specifications; and items are content-accurate and free of technical/editorial flaws.

- Indicator 1.2.c (0/1/2): Item types and cognitive demand across events are sufficient to assess the full intent and complexity of the targeted standards and align to the blueprint/specifications.

- Indicator 1.2.d (0/1/2): Evidence the assessment addresses standards requiring procedural skill and fluency (or analogous skill components), reflected in item documentation and points distributions.

- Indicator 1.2.e (0/1/2): Evidence the assessment addresses standards requiring conceptual understanding, reflected in items and points distributions.

- Indicator 1.2.f (0/1/2): Evidence the assessment addresses standards requiring application; for HS, modeling expectations (when relevant) are administered to all students.

- Indicator 1.2.g (0/1/2): The assessment includes the discipline-specific practices (e.g., mathematical practices; in ELA, comparable practice/skill expectations) as reflected in specifications and delivered forms, with items requiring the practice to earn full credit when aligned.

Criterion 1.3: Fairness and Accessibility (12 points)

Focus: universal design, accommodations, and technology features that protect score meaning for all students.

- Indicator 1.3.a (0/2/4): Development and review procedures ensure fairness. Evidence includes adherence to universal design principles, item rendering specifications aligned to universal design, review processes that minimize construct-irrelevant variance, bias/sensitivity review, and subgroup/accommodation analyses to evaluate technical quality.

- Indicator 1.3.b (0/2/4): Appropriate accommodations and supports are available for intended populations (including students with disabilities and English Learners). Evidence includes clear documentation of intended test-taking populations, accommodations aligned to intended uses, sufficiency of accommodations, and validity/fairness evidence for interpretations under accommodations, plus clear administration guidance and availability of sample forms/items.

- Indicator 1.3.c (0/2/4): Technology features support validity. Evidence includes platform-access guidance, usable auditory supports (natural voice and adjustable cadence), and an overall visual/digital-tool design (e.g., calculators, highlighters) that is navigable and not distracting.

Gateway 2: Technical Quality Criteria and Indicators

Criterion 2.1: Overall Achievement (8 points)

Focus: defensible achievement scores for the target content domain.

- Indicator 2.1.a (0/1/2): Item and form development procedures yield high-quality test events, with review and piloting aligned to content and statistical quality standards.
- Indicator 2.1.b (0/1/2): Achievement scores are reliable. Evidence includes clear procedures for estimating reliability/precision and obtained indices appropriate for intended use.
- Indicator 2.1.c (0/1/2): Achievement scores support intended interpretations. Evidence may include validity studies, equating/linking methods supporting comparability across events, and documentation that promotes consistent presentation/scaffolding and scoring.
- Indicator 2.1.d (0/1/2): Achievement scores are appropriate for intended uses, supported by sufficient theoretical and empirical justification and consistent articulation of use cases.

Criterion 2.2: Predicted Student Performance (claim-dependent)

Focus: predicted results relative to a state summative or other criterion measure (scored only if claimed).

- Indicator 2.2.a (0/1/2; may be N/C): Design supports prediction by demonstrating construct/content similarity to the criterion and providing evidence for specific prediction claims (e.g., to a named test).
- Indicator 2.2.b (0/1/2; may be N/C): Predicted results are reliable; procedures for estimating reliability of predicted scores/classifications are documented and aligned to the inference (e.g., classification reliability).
- Indicator 2.2.c (0/1/2; may be N/C): Predicted results reflect likely future performance; studies document data, samples, methods, and interpretive logic supporting the predictive relationship.
- Indicator 2.2.d (0/1/2; may be N/C): Predicted results are appropriate for intended uses, supported by adequate theoretical/empirical evidence and clear articulation of use.

Criterion 2.3: Sub-scores (claim-dependent)

Focus: strengths-and-need sub scores at reported strands/objectives (scored only if claimed).

- Indicator 2.3.a (0/1/2; may be N/C): Test events are designed to support reporting at each stated level of granularity and to justify interpreting strengths/needs within the content domain.
- Indicator 2.3.b (0/1/2; may be N/C): Reported sub scores are reliable/precise, with defensible, documented estimation methods and adequacy for intended uses.
- Indicator 2.3.c (0/1/2; may be N/C): Sub scores support intended interpretations and represent distinct sub-domains, supported by empirical evidence.
- Indicator 2.3.d (0/1/2; may be N/C): Sub scores are appropriate for intended uses, with sufficient theoretical/empirical support and clear use statements.

Criterion 2.4: Student Progress (claim-dependent)

Focus: progress/growth interpretations across administrations (scored only if claimed).

- Indicator 2.4.a (0/1/2; may be N/C): The assessment is designed to support growth; content and scale characteristics (within/across grades) match the vendor's growth model.

- Indicator 2.4.b (0/1/2; may be N/C): Growth scores are reliable, including appropriate standard errors and evaluation of precision along the score scale.

- Indicator 2.4.c (0/1/2; may be N/C): Growth scores support intended interpretations; methods are documented, and evidence addresses potential disruptions (e.g., redesign/rescaling) and confirms growth inferences.

- Indicator 2.4.d (0/1/2; may be N/C): Growth scores are appropriate for intended uses, supported by adequate evidence and clearly articulated use cases.

Gateway 3: Score Reports and Interpretive Guides Criteria and Indicators

Criterion 3.1: Overall Achievement (10 points)

Focus: reporting and guidance that support correct interpretation and use of overall achievement results.

- Indicator 3.1.a (0/2/4): Report design and information are consistent with intended interpretations and users (educators, parents, students, administrators). Evidence includes user-centered design attention, studies/focus groups showing users can interpret and use reports, warnings about common misuses, and flags for compromised test integrity with conditions explained.

- Indicator 3.1.b (0/1/2): Reports communicate score uncertainty (e.g., confidence intervals, error bands, probability statements) and provide supports/examples to interpret error and its practical implications.

- Indicator 3.1.c (0/2/4): Guidance and supports (instructional/curricular or interpretive) are sufficient and appropriate, aligned to intended use, grounded in research or educator consultation, and cover the full performance range.

Criterion 3.2: Predicted Student Performance (claim-dependent)

Focus: reporting and guidance for predicted performance results (scored only if claimed).

- Indicator 3.2.a (0/2/4; may be N/C): Reports and materials match intended uses for predicted results and demonstrate that intended users can interpret them; include misuse warnings and integrity flags when relevant.

- Indicator 3.2.b (0/1/2; may be N/C): Reports include uncertainty around predicted results and supports to interpret that uncertainty.

- Indicator 3.2.c (0/2/4; may be N/C): Guidance is provided to support appropriate use across the performance range, aligned to the predictive use claim.

Criterion 3.3: Sub-scores (claim-dependent)

Focus: reporting and guidance for sub scores (scored only if claimed).

- Indicator 3.3.a (0/2/4; may be N/C): Reports and materials are consistent with intended uses for sub scores and show users can interpret them; include misuse warnings and integrity flags when relevant.

- Indicator 3.3.b (0/1/2; may be N/C): Reports include uncertainty around sub scores (error bands or similar) with interpretive supports.

- Indicator 3.3.c (0/2/4; may be N/C): Guidance supports appropriate sub score use for students across performance levels and is aligned to the sub score purpose.

Criterion 3.4: Student Progress (claim-dependent)

Focus: reporting and guidance for growth/progress results (scored only if claimed).

- Indicator 3.4.a (0/2/4; may be N/C): Reports and materials for progress results match intended uses and audiences and demonstrate interpretable, usable displays; include misuse warnings and integrity flags when relevant.
- Indicator 3.4.b (0/1/2; may be N/C): Reports communicate uncertainty around progress estimates with supports to interpret it.
- Indicator 3.4.c (0/2/4; may be N/C): Guidance supports appropriate use of progress information, aligned to the growth model and covering the full performance range.