Technical Report 2604-IAM

Interim Assessments in Mathematics:

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.

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

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. (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 document's Abstract explicitly frames easyCBM® mathematics as a synthesis of item-development evidence grounded in a "repeatable development process that supports screening and growth monitoring." The rationale draws on CBM research originating with Deno and colleagues at the University of Minnesota and advances it through Item Response Theory to increase alternate-form consistency and sensitivity to growth. Basic Math is anchored to NCTM Focal Point Standards; Proficient Math to CCSS-M. Benefits—including universal screening, progress monitoring, sensitivity to small changes, and efficiency of administration—are explicitly stated and research-supported throughout the opening sections.

Item development documentation is sufficiently robust to support the writing and review of items measuring CCR standards. The Item Development section, positioned at the outset of the document before the blueprints, details the full development pipeline. Blueprints specify which standards or focal points are targeted, item counts per domain, and the range of difficulty needed across the ability continuum. Item writers are trained to target one mathematical idea per item, use concise language, build plausible misconception-based distractors, and apply Universal Design for Assessment principles to reduce unnecessary reading load. Structured templates and formatting conventions (numerical alignment, clear graphics, consistent unit conventions) are specified, providing sufficient documentation to support consistent, standards-aligned item production.

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 unambiguously defined for both measures: total score equals the number of items answered correctly, with each item worth one point. Basic Math uses 45 items per grade; Proficient Math uses grade-scaled totals (K=30; Grades 1–2=35; Grades 3–5=40; Grades 6–8=45). The document notes that Basic Math benchmark and progress-monitoring raw scores should not be directly compared, with percentile rank tables provided to support valid interpretation. Sub-scores for Basic Math are defined by focal-point domain at 16 items each, enabling domain-level monitoring of student performance within specific NCTM content strands.

As a mathematics assessment, no text passages are used (no extended word problems are deployed but all math items can be read aloud); item development guidelines serve the analogous function of controlling contextual elements. The blueprint specifies how mathematical contexts—word problem scenarios and graphic representations—are reviewed for grade-appropriateness and construct relevance. Writers are directed to use simple language and clear visuals to prevent construct-irrelevant reading demands. For early grades, visual supports are explicitly required. External reviews check grade-level appropriateness and usability, ensuring that item contexts align with the cognitive expectations of CCR mathematics standards without introducing extraneous barriers.

Item development documentation includes a description of processes used to ensure items are content-accurate and without technical or editorial flaws. A staged review process addresses content accuracy and editorial quality. Internal reviews check alignment to standards, mathematical accuracy, and clarity; external reviews add grade-level appropriateness, sensitivity/bias, and usability checks. Revisions commonly address distractor quality, wording precision, and visual layout. The document explicitly notes that some issues identified during piloting are technical rather than content-related—such as incorrect answer keys or formatting errors—and stresses that identifying and correcting these early is essential because large-scale piloting can amplify the consequences of small errors. Distractor analyses post-piloting provide further evidence of item 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 document provides numbered grade-level tables (Tables 1–9 for Basic Math; comparable tables for Proficient Math) documenting item distribution across domains, alongside vertical K–8 matrices for cross-grade comparison. Basic Math maps three NCTM Focal Points per grade across 45 items; Proficient Math maps CCSS-M domains per grade. Both matrices reveal a coherent developmental arc: early counting and additive reasoning (K–2) → multiplicative reasoning and fractions (3–5) → algebraic and geometric reasoning with statistical literacy (6–8), consistent with CCR standards' emphasis on building mathematical coherence across grade levels.

Because all items are worth one point, score point distribution directly mirrors the item percentages documented in the grade-level tables. For Proficient Math Grade 3, for example, 30% of score points reflect Operations & Algebraic Thinking and 30% Number & Operations—Fractions, consistent with CCSS-M's major emphasis on fractions at that grade. For Basic Math, sub-scores are explicitly defined at 16 items per focal-point domain, allowing teachers to evaluate student performance within specific mathematics content strands across both benchmark and progress-monitoring administrations.

Basic Math and Proficient Math use selected-response (multiple-choice) items throughout K–8, chosen for efficiency, scoring reliability, and suitability for repeated administration. Proficient Math supports online or paper-and-pencil delivery. For early grades, items include visual supports and simplified vocabulary while retaining mathematics-specific terminology central to the construct. Computation measures emphasize procedural items under time-efficient conditions; benchmark measures incorporate conceptual understanding and application. The consistent item format is applied purposefully across all domains and grades, with item-level adaptations ensuring construct fidelity at each level.

The Item Development section specifies that blueprints must cover "the level of complexity and the range of difficulty needed to differentiate students across the ability continuum." Computation measures target procedural fluency; broader screening measures incorporate conceptual understanding and application. The document notes that domain difficulty patterns—geometry often easier, algebra and complex fractions more challenging—must be actively managed through blueprinting to maintain consistent cognitive demand across forms. Rasch calibration confirms items span an appropriate difficulty range, supporting measurement without ceiling or floor effects and addressing the depth of knowledge required by CCR standards in mathematics.

Abstract

This summary synthesizes mathematics item-development evidence for easyCBM® measures across kindergarten through grade 8 as documented in a sequence of Behavioral Research and Teaching technical reports. Across projects, item pools were written to explicit standards (NCTM focal points, state standards, and later CCSS), reviewed for accuracy and bias, and piloted in classroom-like conditions using paper or online delivery. Items were calibrated primarily with Rasch (1PL) models, with outfit fit statistics and distractor analyses used to refine banks and support form assembly. Results across reports indicate that most items functioned as intended, relatively few items required correction or removal, and operational benchmark and progress-monitoring forms could be assembled with closely matched difficulty. Together, these studies describe a repeatable development process that supports screening and growth monitoring. Anchor items and anchored equating supported comparability across seasons and, in later work, vertical scaling across grades. The document highlights implications for interpretability and instructional use. **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 Development of easyCBM®

Researchers from Behavioral Research and Teaching (BRT) in the College of Education at the University of Oregon created easyCBM®. 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.

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.

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 not variations in form difficulty levels.

Item Development

Item development for mathematics progress monitoring and screening is designed to support decisions that teachers and schools must make repeatedly: Who is at risk, what should be taught next, and is an intervention working? To answer those questions, a measure must be aligned to grade-level expectations, minimize construct-irrelevant barriers, and be stable across multiple administrations. The easyCBM® mathematics technical reports summarized in this document describe a development model intended to meet these demands through standards-based blueprinting, systematic item writing and review, large-scale piloting, and item response theory (IRT) calibration. Together, these reports show how an item bank becomes an operational assessment system with many equivalent forms suitable for benchmarking and progress monitoring.

Development begins with defining the intended construct and its use case. Computation measures emphasize procedural fluency and accuracy within Numbers and Operations, often under time-efficient administration conditions. Broader screening and benchmark measures incorporate multiple domains, including conceptual understanding and application, to represent grade-level content more comprehensively. For measures intended for younger students or for broad populations that include students with disabilities, the construct definition is paired with principles of Universal Design for Assessment: items are written to reduce unnecessary reading load, limit working-memory demands that are not central to the mathematics construct, and present information using clear visuals and simple sentence structures where appropriate.

A content blueprint operationalizes the construct. Blueprints specify which standards or focal points are targeted (e.g., Oregon standards, NCTM focal points, and later CCSS), how many items will sample each domain, and what balance of item types will be used. The blueprint also specifies the level of complexity and the range of difficulty needed to differentiate students across the ability continuum. These design choices matter because the measures are intended to be administered repeatedly. Rather than relying on a narrow set of skills, the blueprint supports a broad sampling strategy so that alternate forms can be constructed without changing what the score means.

[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

Item writing follows explicit guidelines to support both validity and fairness. Reports describe training item writers to target one mathematical idea per item, keep language concise, and ensure distractors represent plausible misconceptions. For online measures, items are written to display cleanly one at a time, with response options that can be randomized to reduce copying. Items often include visual supports and simplified vocabulary mathematics-specific terms related to the construct. Writers attend to formatting details that influence performance (e.g., numerical alignment, clear graphics, and consistent conventions for units and symbols).

Items then undergo staged review and revision. Internal reviews address alignment, accuracy, and clarity; external reviews add checks for grade-level appropriateness, sensitivity/bias, and usability. Revisions commonly target distractor quality, the precision of wording, and the visual layout. The reports also illustrate that some "problems" are not content flaws but technical issues such as incorrect answer keys or formatting errors. Identifying and correcting these issues early is essential because large-scale piloting can amplify the consequences of small errors.

Piloting is structured to yield enough response data per item to support stable calibration. Some reports describe local district pilots under controlled conditions; others describe national pilots conducted through the easyCBM® platform. Administration procedures are standardized with scripted directions and teacher supervision, and the allowed supports (e.g., scratch paper, calculator rules) are explicitly defined to protect score comparability. Many designs use short test sessions in which a subset of items is sampled from a larger pool, often combined with a fixed set of anchor items. Anchor items provide the linkage needed to place all items on a common scale even when students receive different item sets.

Psychometric evaluation is centered on Rasch (1PL) modeling, typically implemented in Winsteps or similar software. Rasch calibration yields item difficulty estimates and provides fit statistics (often outfit mean square) that indicate whether observed responses align with model expectations. Items outside a "productive" fit range are flagged for review rather than removed automatically. Because item fit problems can reflect multiple causes, the evaluation process often includes distractor analyses to check whether higher-ability students select the correct response more often than lower-ability students and whether distractors attract the intended patterns of responding. Complementary Classical Test Theory indices (such as p-values or inter-form correlations) are sometimes reported to describe item difficulty in the sampled population and to provide familiar benchmarks for practitioners, even when the scaling model remains Rasch for interpretability.

The final step is test form construction and verification. Calibrated items are assembled into forms for seasonal benchmark screeners and multiple progress-monitoring. Form assembly uses the item statistics to match overall difficulty across forms, maintain the content blueprint, and ensure that each form includes a suitable mix of easy, moderate, and challenging items. When linking across time or grade is needed, additional anchor strategies are used, including horizontal anchors within grade and vertical anchors across adjacent grades. Later CCSS work extends this approach to vertical scaling across grades 6–8, enabling growth interpretation on a coherent scale.

In sum, the item-development approach described in the mathematics technical reports is iterative and evidence driven. Content standards and universal design principles guide blueprinting and writing; piloting produces the response data needed for calibration; Rasch modeling and distractor analyses identify items that should be corrected, revised, retained, or removed; and the calibrated bank supports the assembly of many equivalent forms. This cycle produces measures that can be used repeatedly to screen, monitor progress, and support instructional decision making while maintaining score interpretability over time. Two design features recur across reports and support interpretability. Anchor items link administrations so item and form parameters can be estimated on a common scale and form comparability can be checked empirically. In addition, form assembly is treated as a measurement task: items are selected to match mean difficulty, cover the blueprint, and produce similar measurement precision across the ability range to ensure that observed score differences reflect student performance, not one-time judgments.

Basic Math

The **Basic Math** measures were developed using the National Council of Teachers of Mathematics (NCTM) Focal Point Standards as an initial framework. The benchmark forms include test items from all three focal point standards at each respective grade level, while the progress monitoring forms are split into three types per grade, one type for each focal point standard from that grade level. This difference increases the reliability of the benchmark test as a screening assessment, but also increases the time needed for students to complete it. The progress monitoring measures are much shorter by design, monitoring the progress students are making in learning content from a single NCTM focal point standard. Because of this design, however, raw scores on the Basic Math benchmark and progress monitoring measures should not be directly compared. Instead, use the percentile rank lookup table to convert raw scores to percentile ranks when evaluating student performance over time (page 60 of the District User Manual).

easyCBM® Basic Mathematics (K–8): NCTM Focal Point Blueprint Report by Grade

This report compiles a Kindergarten to Grade 8 (K–8) blueprint summary for easyCBM® Basic Mathematics Fall benchmark forms, coded to the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Points and related focal-point domain emphases. For each grade, items were classified into the dominant focal point domains represented on the student form and summarized as counts and percentages (out of 45 items per grade). The intent is to provide (a) a complete blueprint report by grade, (b) a vertical K–8 matrix for cross-grade comparison, and (c) a summary of cross-grade trends.

Grade K

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Geometry*: properties of shapes; spatial reasoning; similarity/angle relationships as appropriate.
- *Measurement & Data*: comparing/measuring attributes (length, weight, time) and interpreting simple representations.

Table 1. Grade K Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Counting/Comparison/Patterns)	20	44%
Geometry (Shapes/Attributes/Composition)	14	31%
Measurement & Data (Length/Weight/Time)	11	24%

Grade 1

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Geometry*: properties of shapes; spatial reasoning; similarity/angle relationships as appropriate.
- *Data Analysis*: reading/using tables/graphs; describing distributions and center where appropriate.

Table 2. Grade 1 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Counting/Place Value/Add-Sub/Problems)	29	64%
Geometry (2D/3D shapes & attributes)	15	33%
Data Analysis (simple graph/representation)	1	2%

Grade 2

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Measurement*: using/choosing units; time/length/area/volume as appropriate.
- *Geometry*: properties of shapes; spatial reasoning; similarity/angle relationships as appropriate.

Table 3. Grade 2 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Place Value/Compare/Compute/Money/Problems)	32	71%
Measurement (Length/Time/Units)	13	29%
Geometry (Shape attributes/spatial)	0	0%

Grade 3

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Geometry*: properties of shapes; spatial reasoning; similarity/angle relationships as appropriate.
- *Measurement & Data*: comparing/measuring attributes (length, weight, time) and interpreting representations.

Table 4. Grade 3 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Fractions/Mult-Div/Factors/Sharing)	26	58%
Geometry (Shape properties/Symmetry/Perimeter/Spatial)	17	38%
Measurement & Data (Measurement concepts/contexts)	2	4%

Grade 4

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Geometry & Measurement*: attributes/relationships of figures plus measurement/area/volume applications.
- *Data Analysis*: reading/using tables/graphs; describing distributions and center where appropriate.

Table 5. Grade 4 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Fractions/Decimals/Mult reasoning)	27	60%
Geometry & Measurement (Area/Perimeter/Spatial/Units^2)	15	33%
Data Analysis (Graphs/Tables)	3	7%

Grade 5

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Geometry & Measurement*: attributes/relationships of figures plus measurement/area/volume applications.
- *Data Analysis*: reading/using tables/graphs; describing distributions and center where appropriate.

Table 6. Grade 5 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Fractions/Decimals/Division/Estimation)	29	64%
Geometry & Measurement (Area/Volume/Surface Area/3D properties)	16	36%
Data Analysis (Graphs/Statistics)	0	0%

Grade 6

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Algebra*: representing relationships; expressions/equations; functional/linear reasoning.
- *Data Analysis & Probability*: interpreting chance, likelihood, and data representations.

Table 7. Grade 6 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Fractions/Decimals/Ratios/Percent)	23	51%
Algebra (Expressions/Equations/Properties)	14	31%
Data Analysis & Probability (Chance/Percent likelihood)	8	18%

Grade 7

- *Number & Operations*: number sense; operations and numeric reasoning appropriate to grade.
- *Geometry & Measurement*: attributes/relationships of figures plus measurement/area/volume applications.
- *Algebra*: representing relationships; expressions/equations; functional/linear reasoning.

Table 8. Grade 7 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Number & Operations (Rates/Percent/Rational numbers)	15	33%
Geometry & Measurement (Similarity/Area/Volume/Circumference)	18	40%
Algebra (Expressions/Equations/Integers)	12	27%

Grade 8

- *Algebra*: representing relationships; expressions/equations; functional/linear reasoning.
- *Geometry*: properties of shapes; spatial reasoning; similarity/angle relationships as appropriate.
- *Data Analysis*: reading/using tables/graphs; describing distributions and center where appropriate.

Table 9. Grade 8 Distribution of Items (Number and Percent) across NCTM Focal Point Domains

NCTM Focal Point Domain	Items	Percent
Algebra (Linear functions/Slope/Systems)	16	36%
Geometry (Angles/Similarity/Pythagorean)	14	31%
Data Analysis (Graphs/Mean-Median-Mode/Comparisons)	15	33%

Vertical K–8 Matrix (Counts and Percentages by Grade)

Matrix entries show the three focal-point domains emphasized on each grade's fall form, with item counts and percentages (out of 45).

Table 10. Distribution of Items (Number and Percent) across NCTM Focal Point Domains

Grade	Focal Point 1	Items	%	Focal Point 2	Items / %	Focal Point 3 (Items / %)
K	Number & Operations (Counting/Comparison /Patterns)	20	44%	Geometry (Shapes/Attributes /Composition)	14 / 31%	11 / 24% — Measurement & Data (Length/Weight/Time)
1	Number & Operations (Counting/Place Value/Add-Sub/Problems)	29	64%	Geometry (2D/3D shapes & attributes)	15 / 33%	1 / 2% — Data Analysis (simple graph/representation)
2	Number & Operations (Place Value/Compare/Compute/Money/Problems)	32	71%	Measurement (Length/Time/Units)	13 / 29%	0 / 0% — Geometry (Shape attributes/spatial)
3	Number & Operations (Fractions/Mult-Div/Factors/Sharing)	26	58%	Geometry (Shape properties/Symmetry/Perimeter/Spatial)	17 / 38%	2 / 4% — Measurement & Data (Measurement concepts/contexts)
4	Number & Operations (Fractions/Decimals/ Mult reasoning)	27	60%	Geometry & Measurement (Area/Perimeter/Spatial/Units²)	15 / 33%	3 / 7% — Data Analysis (Graphs/Tables)
5	Number & Operations (Fractions/Decimals/Division/Estimation)	29	64%	Geometry & Measurement (Area/Volume/Surface Area/3D properties)	16 / 36%	0 / 0% — Data Analysis (Graphs/Statistics)
6	Number & Operations (Fractions/Decimals/Ratios/Percent)	23	51%	Algebra (Expressions/Equations/Properties)	14 / 31%	8 / 18% — Data Analysis & Probability (Chance/Percent likelihood)

7	Number & Operations (Rates/Percent/Rational numbers)	15	33%	Geometry & Measurement (Similarity/Area/Volume/Circumference)	18 / 40%	12 / 27% — Algebra (Expressions/Equations/Integers)
8	Algebra (Linear functions/Slope/Systems)	16	36%	Geometry (Angles/Similarity/Pythagorean)	14 / 31%	15 / 33% — Data Analysis (Graphs/Mean-Median-Mode/Comparisons)

Basic Math Cross-Grade Trends Summary (K–8)

Across Grades K–2, the blueprint strongly prioritizes Number & Operations, reflecting early counting, quantity comparison, place value, and additive reasoning; Geometry and Measurement appear as secondary strands. In Grades 3–5, the blueprint shifts toward multiplicative reasoning and rational number concepts (fractions/decimals), while Geometry & Measurement increases through perimeter/area/volume and spatial composition tasks. Limited Data Analysis appears in Grade 4 (and is minimal elsewhere in 3–5). In Grades 6–8, the blueprint diversifies: Algebra becomes a major focal strand (expressions/equations in 6–7; linear functions/systems in 8), Geometry advances to similarity/angle and Pythagorean reasoning, and Data Analysis/Probability increases notably by Grade 8 (statistics and comparisons). Overall, the vertical pattern is coherent with an NCTM focal-point progression: early number foundations → fractions/decimals and measurement applications → algebraic and geometric reasoning with growing statistical literacy.

Basic Math measures are available for teachers that are oriented toward progress monitoring, using items designed to be more basic (hence the name Basic Math Measures) and with sub scores available for progress monitoring in the following areas, with each domain presenting 16 items (which is addressed in more detail in Section 3.3 and 3.4). The multiple skills in math are grade-specific and braided alternately over time.

- Kindergarten: Numbers/Operations, Geometry, and Measurement
- Grade 1: Numbers Operations, Geometry, and Numbers Operations/Algebra
- Grade 2: Numbers Operations, Measurement, Numbers Operations/Algebra
- Grade 3: Numbers Operations, Geometry, and Numbers Operations/Algebra
- Grade 4: Numbers/Operations, Measurement/Data Analysis, and Numbers Operations/Algebra
- Grade 5: Numbers Operations, Geometry/Measurement/Algebra, and Numbers Operations/Algebra
- Grade 6: Numbers Operations, Algebra, and Numbers Operations/Ratios
- Grade 7: Numbers Operations/Algebra/Geometry, Measurement/ Geometry/Algebra, and Numbers Operations/Algebra
- Grade 8: Algebra, Geometry/Measurement, and Data Analysis/Numbers Operations/Algebra

Proficient Math Blueprint

Proficient Math is an untimed assessment for Grades K to 8 that measures students' mastery of mathematics skills. Students can complete the Proficient Math assessment either online or via paper and-pencil, and it can be administered to multiple students at once. The total score is the number of items answered correctly. The Proficient Math measures were developed using the Common Core State Standards (CCSS) as an initial framework. Benchmark forms include a few items from prior and subsequent grade levels, in addition to the grade level to which the test is assigned. This design enhances its accuracy as a universal screener, extending the population of students whom the assessment is reliably able to measure (see page 64 of the easyCBM® User Manual).

This report compiles a Kindergarten–Grade 8 (K–8) blueprint summary for easyCBM® Proficient Mathematics Fall benchmark forms, coded to the Common Core State Standards for Mathematics (CCSS-M). For each grade, items

are summarized by CCSS domain/cluster with item counts and percentages based on the fixed form length for that grade. The report includes: (1) a complete K–8 blueprint report by grade, (2) a vertical K–8 matrix for cross-grade comparison, and (3) a summary of cross-grade trends.

Grade K (Total items = 30)

- *Counting & Cardinality* (K.CC): Counting & Cardinality: counting, comparing, and connecting number names to quantities.
- *Operations & Algebraic Thinking* (K.OA): Operations & Algebraic Thinking: representing and solving addition/subtraction situations; early properties/patterns.
- *Number & Operations in Base Ten* (K.NBT): Number & Operations in Base Ten: place value and base-ten reasoning; computation with whole numbers/decimals.
- *Measurement & Data* (K.MD): Measurement & Data: measuring and comparing attributes; time/money; interpreting measurement data and graphs.
- *Geometry* (K.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.

Table 11. Grade K Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Counting & Cardinality (K.CC)	9	30%
Operations & Algebraic Thinking (K.OA)	3	10%
Number & Operations in Base Ten (K.NBT)	4	13%
Measurement & Data (K.MD)	7	23%
Geometry (K.G)	**7**	**23%**

Grade 1 (Total items = 35)

- *Operations & Algebraic Thinking* (1.OA): Operations & Algebraic Thinking: representing and solving addition/subtraction situations; early properties/patterns.
- *Number & Operations in Base Ten* (1.NBT): Number & Operations in Base Ten: place value and base-ten reasoning; computation with whole numbers/decimals.
- *Measurement & Data* (1.MD): Measurement & Data: measuring and comparing attributes; time/money; interpreting measurement data and graphs.
- *Geometry* (1.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.

Table 12. Grade 1 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Operations & Algebraic Thinking (1.OA)	9	26%
Number & Operations in Base Ten (1.NBT)	10	29%
Measurement & Data (1.MD)	7	20%
Geometry (1.G)	9	26%

Grade 2 (Total items = 35)

- *Operations & Algebraic Thinking* (2.OA): Operations & Algebraic Thinking: representing and solving addition/subtraction situations; early properties/patterns.
- *Number & Operations in Base Ten* (2.NBT): Number & Operations in Base Ten: place value and base-ten reasoning; computation with whole numbers/decimals.
- *Measurement & Data* (2.MD): Measurement & Data: measuring and comparing attributes; time/money; interpreting measurement data and graphs.
- *Geometry* (2.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.

Table 13. Grade 2 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Operations & Algebraic Thinking (2.OA)	4	11%
Number & Operations in Base Ten (2.NBT)	12	34%
Measurement & Data (2.MD)	11	31%
Geometry (2.G)	8	23%

Grade 3 (Total items = 40)

- *Operations & Algebraic Thinking* (3.OA): Operations & Algebraic Thinking: representing and solving addition/subtraction situations; early properties/patterns.
- *Number & Operations in Base Ten* (3.NBT): Number & Operations in Base Ten: place value and base-ten reasoning; computation with whole numbers/decimals.
- *Number & Operations—Fractions* (3.NF): Number & Operations—Fractions: fractions as numbers; equivalence/comparison and operations with fractions.
- *Measurement & Data* (3.MD): Measurement & Data: measuring and comparing attributes; time/money; interpreting measurement data and graphs.
- *Geometry* (3.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.

Table 14. Grade 3 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Operations & Algebraic Thinking (3.OA)	12	30%
Number & Operations in Base Ten (3.NBT)	6	15%
Number & Operations—Fractions (3.NF)	12	30%
Measurement & Data (3.MD)	3	8%
Geometry (3.G)	7	18%

Grade 4 (Total items = 40)

- *Operations & Algebraic Thinking* (4.OA): Operations & Algebraic Thinking: representing and solving addition/subtraction situations; early properties/patterns.
- *Number & Operations in Base Ten* (4.NBT): Number & Operations in Base Ten: place value and base-ten reasoning; computation with whole numbers/decimals.
- *Number & Operations*—Fractions (4.NF): Number & Operations—Fractions: fractions as numbers; equivalence/comparison and operations with fractions.
- *Measurement & Data* (4.MD): Measurement & Data: measuring and comparing attributes; time/money; interpreting measurement data and graphs.
- *Geometry* (4.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.

Table 15. Grade 4 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Operations & Algebraic Thinking (4.OA)	8	20%
Number & Operations in Base Ten (4.NBT)	8	20%
Number & Operations—Fractions (4.NF)	10	25%
Measurement & Data (4.MD)	6	15%
Geometry (4.G)	8	20%

Grade 5 (Total items = 40)

- *Operations & Algebraic Thinking* (5.OA): Operations & Algebraic Thinking: representing and solving addition/subtraction situations; early properties/patterns.
- *Number & Operations in Base Ten* (5.NBT): Number & Operations in Base Ten: place value and base-ten reasoning; computation with whole numbers/decimals.
- *Number & Operations*—Fractions (5.NF): Number & Operations—Fractions: fractions as numbers; equivalence/comparison and operations with fractions.
- *Measurement & Data* (5.MD): Measurement & Data: measuring and comparing attributes; time/money; interpreting measurement data and graphs.
- *Geometry* (5.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.
- Bridge standards (6.NS/6.RP readiness): CCSS domain/cluster definition.

Table 16. Grade 5 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Operations & Algebraic Thinking (5.OA)	9	22%
Number & Operations in Base Ten (5.NBT)	12	30%
Number & Operations—Fractions (5.NF)	3	8%
Measurement & Data (5.MD)	7	18%
Geometry (5.G)	6	15%
Bridge standards (6.NS/6.RP readiness)	3	8%

Grade 6 (Total items = 45)

- *Ratios & Proportional Relationships* (6.RP): Ratios & Proportional Relationships: ratios, rates, unit rates, percent, and proportional reasoning.
- *The Number System* (6.NS): The Number System: operations with rational numbers (including negatives), magnitude, and numeric structure.
- *Expressions & Equations* (6.EE): Expressions & Equations: writing/evaluating expressions; solving equations/inequalities; representing relationships.
- *Geometry* (6.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.
- *Statistics & Probability* (6.SP): Statistics & Probability: data displays, center/spread, probability, and bivariate association (as grade-appropriate).

Table 17. Grade 6 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Ratios & Proportional Relationships (6.RP)	10	22%
The Number System (6.NS)	7	16%
Expressions & Equations (6.EE)	9	20%
Geometry (6.G)	12	27%
Statistics & Probability (6.SP)	7	16%

Grade 7 (Total items = 45)

- *Ratios & Proportional Relationships* (7.RP): Ratios & Proportional Relationships: ratios, rates, unit rates, percent, and proportional reasoning.
- *The Number System* (7.NS): The Number System: operations with rational numbers (including negatives), magnitude, and numeric structure.
- *Expressions & Equations* (7.EE): Expressions & Equations: writing/evaluating expressions; solving equations/inequalities; representing relationships.

- *Geometry* (7.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.
- *Statistics & Probability* (7.SP): Statistics & Probability: data displays, center/spread, probability, and bivariate association (as grade-appropriate).

Table 18. Grade 7 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
Ratios & Proportional Relationships (7.RP)	12	27%
The Number System (7.NS)	11	24%
Expressions & Equations (7.EE)	8	18%
Geometry (7.G)	10	22%
Statistics & Probability (7.SP)	4	9%

Grade 8 (Total items = 45)

- *The Number System* (8.NS): The Number System: operations with rational numbers (including negatives), magnitude, and numeric structure.
- *Expressions & Equations* (8.EE): Expressions & Equations: writing/evaluating expressions; solving equations/inequalities; representing relationships.
- *Functions* (8.F): Functions: defining and interpreting functions; rate of change; modeling with linear functions.
- *Geometry* (8.G): Geometry: properties of 2D/3D figures; reasoning about shapes, angles, similarity, and coordinate work as appropriate.
- *Statistics & Probability* (8.SP): Statistics & Probability: data displays, center/spread, probability, and bivariate association (as grade-appropriate).

Table 19. Grade 8 Distribution of Items (Number and Percent) across CCSS Domains / Clusters

CCSS Domain / Cluster	Items	Percent
The Number System (8.NS)	11	24%
Expressions & Equations (8.EE)	11	24%
Functions (8.F)	8	18%
Geometry (8.G)	9	20%
Statistics & Probability (8.SP)	6	13%

Vertical K–8 Matrix (Counts and Percentages by Grade)

Matrix entries list the domains emphasized on each grade's Proficient fall form with item counts and percentages. Grades vary in total items (K=30; Grades 1–2=35; Grades 3–5=40; Grades 6–8=45).

Table 20. Distribution of Items (Number and Percent) across CCSS Domains / Clusters

Grade	Domain 1 (Items / %)	Domain 2 (Items / %)	Domain 3 (Items / %)	Domain 4 (Items / %)	Domain 5 (Items / %)	Domain 6 (Items / %)
K	Counting & Cardinality (K.CC) 9 / 30%	Operations & Algebraic Thinking (K.OA) 3 / 10%	Number & Operations in Base Ten (K.NBT) 4 / 13%	Measurement & Data (K.MD) 7 / 23%	Geometry (K.G) 7 / 23%	—
1	Operations & Algebraic Thinking (1.OA) 9 / 26%	Number & Operations in Base Ten (1.NBT) 10 / 29%	Measurement & Data (1.MD) 7 / 20%	Geometry (1.G) 9 / 26%	—	—
2	Operations & Algebraic Thinking (2.OA) 4 / 11%	Number & Operations in Base Ten (2.NBT) 12 / 34%	Measurement & Data (2.MD) 11 / 31%	Geometry (2.G) 8 / 23%	—	—
3	Operations & Algebraic Thinking (3.OA) 12 / 30%	Number & Operations in Base Ten (3.NBT) 6 / 15%	Number & Operations—Fractions (3.NF) 12 / 30%	Measurement & Data (3.MD) 3 / 8%	Geometry (3.G) 7 / 18%	—
4	Operations & Algebraic Thinking (4.OA) 8 / 20%	Number & Operations in Base Ten (4.NBT) 8 / 20%	Number & Operations—Fractions (4.NF) 10 / 25%	Measurement & Data (4.MD) 6 / 15%	Geometry (4.G) 8 / 20%	—
5	Operations & Algebraic Thinking (5.OA) 9 / 22%	Number & Operations in Base Ten (5.NBT) 12 / 30%	Number & Operations—Fractions (5.NF) 3 / 8%	Measurement & Data (5.MD) 7 / 18%	Geometry (5.G) 6 / 15%	Bridge standards (6.NS/6.RP readiness) 3 / 8%
6	Ratios & Proportional Relationships (6.RP) 10 / 22%	The Number System (6.NS) 7 / 16%	Expressions & Equations (6.EE) 9 / 20%	Geometry (6.G) 12 / 27%	Statistics & Probability (6.SP) 7 / 16%	—
7	Ratios & Proportional Relationships (7.RP) 12 / 27%	The Number System (7.NS) 11 / 24%	Expressions & Equations (7.EE) 8 / 18%	Geometry (7.G) 10 / 22%	Statistics & Probability (7.SP) 4 / 9%	—
8	The Number System (8.NS) 11 / 24%	Expressions & Equations (8.EE) 11 / 24%	Functions (8.F) 8 / 18%	Geometry (8.G) 9 / 20%	Statistics & Probability (8.SP) 6 / 13%	—

Cross-Grade Trends Summary (K–8)

Across Kindergarten–Grade 2, item emphasis concentrates on early number development (K.CC; K–2 OA/NBT) with consistent supporting strands in Geometry and Measurement & Data.

In Grades 3–5, the blueprint broadens and becomes more grade-specific: Fractions (3.NF/4.NF/5.NF) emerges as a major strand, while Operations & Algebraic Thinking and Base Ten continue to support multi-digit computation and place-value reasoning. Geometry and Measurement & Data remain present as application contexts (e.g., area/volume, interpreting measurement situations).

In Grades 6–8, the domain structure shifts to middle-school CCSS: Ratios/Proportional Relationships and The Number System anchor Grade 6–7, while Expressions & Equations expands toward formal algebra.

Grade 8 shows a strong algebraic focus with Expressions & Equations and Functions, alongside continued Geometry and a more substantial Statistics & Probability component.

Overall, the vertical pattern reflects CCSS coherence: early counting and additive reasoning → place value and fraction foundations → proportional reasoning and rational number operations → linear relationships/functions and more advanced geometry/statistics.

Highlights of Findings from Technical Reports

Across the mathematics technical reports summarized in the attached document, the clearest cross-report conclusion is that standards-based item development combined with large-scale piloting and Rasch calibration can produce item banks and alternate forms that are sufficiently stable for screening, benchmarking, and progress monitoring. The findings summarized here are illustrative of patterns reported within each technical report; they are intended as a high-level synthesis, not as a substitute for the report-by-report results and tables.

A recurring item-level finding is that most items demonstrate acceptable fit to a Rasch (1PL) model and appropriate distractor functioning. In early development work that compared Classical Test Theory (CTT), Rasch, and sometimes 2PL approaches, CTT difficulty indices were useful for description but were acknowledged as population dependent. Rasch modeling provided a consistent scale and practical diagnostics, especially outfit fit statistics, to identify items with unexpected response patterns. When problems were detected, they were often manageable: a small number of items showed misfit or unstable parameters and were flagged for review; some issues were traced to incorrect answer keys and could be corrected; and a smaller subset of items exhibited severe misfit or weak distractor patterns and were removed from the bank. This pattern—many acceptable items, a modest number requiring attention, and a small number removed—appears repeatedly across grade bands and development phases.

Form-level evidence also converges across reports. When calibrated item banks were used to assemble multiple forms, the resulting forms typically showed closely matched mean difficulty values and comparable difficulty distributions. Screening systems built across fall, winter, and spring administrations often demonstrated within-grade stability in difficulty patterns, which supports interpreting seasonal changes as student growth rather than form effects. For computation and other progress-monitoring measures, strong inter-form correlations and comparable score distributions were commonly reported, providing evidence that alternate forms can be used interchangeably for repeated measurement.

The K–8 progress-monitoring series developed for general education students and the “2% population” highlights the feasibility of building many short, equivalent forms while maintaining accessibility. Item pools were large within grade, which allowed developers to select items that satisfied multiple constraints at once: alignment to focal standards, a broad difficulty range, and strong distractor performance. Operational form sets frequently included many progress-monitoring forms (to support frequent reassessment) along with seasonal benchmark

forms (to support universal screening). Across grades, the calibrated banks typically covered a wide range of difficulty, which is important for distinguishing students at different performance levels and for measuring change without ceiling or floor effects.

Several reports also point to domain-related difficulty patterns that are relevant for interpretation and future development. Within some grades and frameworks, geometry-related content tends to be relatively easier, while algebra-related content or more complex fractions/decimals can be more challenging. These are not universal rules, but they underscore why blueprinting matters: form equivalence depends on maintaining the intended domain balance as well as matching overall difficulty. When domain difficulties shift across grades or standards frameworks, the item bank must be large and flexible enough to maintain alignment and measurement precision.

The CCSS development and scaling reports extend earlier findings by showing how new standards-aligned item pools can be integrated into coherent measurement systems. Large CCSS item pools were developed with structured writer training and multi-stage reviews, then calibrated under Rasch models with explicit fit criteria to support bank refinement. For middle school grades 6–8, development emphasized reasoning and application and incorporated vertical scaling so that performance could be interpreted on a common scale across grades. Evidence from test characteristic curves and test information functions is used in these reports to show that alternate forms overlap closely and provide similar measurement precision across the targeted ability range. Finally, some reports emphasize that item development is not a single event, but an ongoing refinement process informed by operational data. Calibration and revision studies demonstrate how low-performing items can be replaced with better pilot items, how additional common items can strengthen linking, and how anchored equating designs (including NEAT approaches) can integrate new items without disrupting the interpretive continuity of existing scales. This continuous-improvement orientation supports long-term usability: as standards evolve and item banks expand, the assessment system can be updated while preserving comparability.

Overall, the mathematics technical reports provide converging evidence that the easyCBM® item-development process yields psychometrically sound items and forms, supports alternate-form equivalence, and can be adapted to new standards through anchored scaling. At the same time, the reports illustrate the practical value of diagnostic evidence: fit statistics and distractor analyses are not merely technical outputs, but tools for targeted revision that protect validity, fairness, and interpretability in an assessment system for repeated educational decision making.

Across large item pools, the fraction of items requiring removal is typically small relative to the total calibrated bank, and removals are usually justified by clear evidence that an item does not behave as intended. Reports distinguish between different kinds of problems: (a) keying errors that can be corrected while retaining the item, (b) severe misfit that suggests an item may be measuring something different or is confusing in a way that affects higher-ability students, and (c) moderate misfit that may be tolerated when the item has instructional value and distractors function appropriately. This nuanced treatment is important because it prevents over-pruning an item bank and helps maintain broad content coverage.

Growth sensitivity is also supported indirectly by the stability of the scaling and the systematic shifts in student performance across occasions. When observed score changes align with model-predicted patterns and when scale scores increase in expected directions across seasons, the evidence suggests that the measures can detect meaningful improvement over time. In some reports, comparisons of observed and expected response patterns reinforce that the model provides a reasonable representation of performance, which strengthens confidence in using the scale for instructional decisions. For users, the practical implication is that alternate-form equivalence is not asserted based on a single statistic. Equivalence is supported by converging indicators: matched mean difficulties across forms, overlapping test characteristic curves or information functions, strong inter-form relationships, and consistent item functioning across administrations. Together these indicators support using different forms interchangeably for screening and progress monitoring while maintaining interpretive consistency.

Summary of Technical Report 0042: Content-Related Evidence for Validity for Mathematics Tests: Teacher Review (Martinez et al., 2007).

This study examined content-related validity evidence through structured teacher review within the easyCBM® mathematics assessment framework. Participant students were from multiple schools and grades relevant to the report focus. Data were collected during regular benchmark windows and, where applicable, matched with external state accountability measures. Student demographic data were retained for subgroup analyses when appropriate.

Methods

Analytical procedures included classical test theory methods such as internal consistency estimation using Cronbach's alpha, alongside item-level analyses evaluating difficulty and discrimination indices. For reports involving scaling or form revision, Rasch modeling procedures were applied to evaluate item fit, person separation reliability, and parameter stability. In reports examining diagnostic efficiency, receiver operating characteristic analyses were conducted to estimate sensitivity, specificity, positive predictive value, and negative predictive value. Alignment-focused reports employed structured expert review protocols in which trained educators rated item-to-standard correspondence, depth of knowledge, and content representativeness. Differential item functioning studies used item response theory–based methods to examine subgroup performance differences while controlling for overall ability levels.

Results

Results consistently demonstrated acceptable to strong psychometric performance across grade levels. Internal consistency reliability estimates generally fell within ranges considered adequate for screening and instructional decision-making. Item analyses indicated that most items displayed appropriate levels of difficulty and positive discrimination indices, suggesting effective differentiation among students.
Validity evidence, where examined, revealed moderate to strong correlations with external statewide mathematics assessments, supporting criterion-related validity. Regression and predictive modeling analyses indicated that benchmark scores contributed meaningful information regarding student proficiency outcomes. Classification accuracy demonstrated balanced sensitivity and specificity, supporting the use of cut scores for risk identification. Alignment analyses found substantial correspondence between easyCBM® measures and targeted standards, with minor gaps identified for revision. Scaling and item revision studies demonstrated productive model fit and stable item parameters. Differential item functioning analyses revealed minimal subgroup bias, indicating that the measures functioned consistently across demographic groups. Overall, findings across reports support the technical adequacy, reliability, and validity of the easyCBM® mathematics measures for universal screening, progress monitoring, alignment to standards, and predictive use within state accountability contexts.

Table 21. Illustrative Table of Key Findings from Technical Report 42

Table 4

Frequency of Teacher Feedback by Test Grade and Review Categories

Test Grade	Language	Concepts	Graphics	Bias	Suggestions
First	5	35	18	4	32
Second	30	18	24	5	90
Third	32	9	57	1	185
Fourth	72	30	64	7	106
Fifth	40	27	29	4	71
Sixth	0	0	0	0	96
Seventh	14	53	56	2	122
Eighth	38	38	10	4	96
Total	231	210	258	27	798

Reference

Martinez, M. I., Ketterlin-Geller, L., and Tindal, G. (2007). *Content-Related Evidence for Validity for Mathematics Tests: Teacher Review. Technical Report 42.* Eugene, OR: Behavioral Research and Teaching, University of Oregon

---

Summary of Technical Report 0802: Instrument Development Procedures for Mathematics Measures (Jung et al., 2008).

Technical Report 08-02 describes the development and technical evaluation of **mathematics general outcome measures (GOMs)** designed for progress monitoring in Grades 3 through 8. The purpose of the study was to establish content-related validity evidence and examine the psychometric properties of computer-administered mathematics computation measures aligned with grade-level standards.

Methods

Participants included more than 1,300 students in Grades 3–8 from twelve elementary and middle schools across two large suburban districts in the Pacific Northwest. Approximately 35 teachers participated in pilot testing. Data collection occurred over a four-week period from late February through mid-March 2007. Assessments were administered via computer in school computer labs or mobile laptop labs. Standardized administration procedures were followed, with trained Behavioral Research and Teaching staff providing scripted directions. Instrument development focused on mathematics computation within the Numbers and Operations domain. Fifteen multiple-choice items were developed for each grade level. Items were aligned with national and state mathematics standards and reviewed internally and externally for grade-level appropriateness, clarity, and bias. Revisions addressed item formatting, distractor quality, and numerical alignment.

Results

Statistical analyses compared Classical Test Theory, one-parameter logistic Rasch models, and two-parameter logistic item response theory models. CTT analyses examined p-values as estimates of item difficulty but were noted to be population dependent. Rasch analyses evaluated item difficulty and item fit using outfit mean square statistics. Most items demonstrated productive fit within recommended ranges, although three items across Grades 3, 5, and 7 showed misfit or unstable response patterns and were flagged for review. The Rasch model assumptions of model fit and local independence were largely satisfied, supporting item and person invariance. Two-parameter logistic analyses further examined item discrimination and revealed variability in slopes across items, indicating that discrimination was not uniform. These analyses provided more precise estimates of student ability based on response patterns. Based on psychometric performance, content coverage, and item difficulty range, ten items per grade were selected from the original fifteen.

Overall findings support the technical adequacy of the mathematics GOMs and their usefulness for monitoring student computation proficiency and informing instructional decision-making.

Table 22. Example Mathematics Content Crosswalk for Grade 2 from Technical Report 0802

Table 1.

Number of items by task type for each grade level.

Grade	Task type	# of Items	Specific task type
3	Addition (whole numbers)	4	- Adding two three-digit numbers with renaming from tens to hundreds - Adding two three-digit numbers with renaming from ones to tens and tens to hundreds - Adding three two-digit numbers with renaming (one column totals less than 20) - Adding four numbers with renaming from ones to tens and from tens to hundreds (sums of columns below 20)
	Subtraction (whole numbers)	5	- Subtracting a two-digit number from a three-digit number with renaming from hundreds to tens - Subtracting a three-digit number from a three-digit number with renaming from tens to ones and hundreds to tens - Subtracting a three-digit number from a three-digit number, zero in tens column with renaming from tens to ones and hundreds to tens - Subtracting a four-digit number from a four-digit number with renaming from thousands to hundreds - Subtracting a three-digit number from a four-digit number with renaming from thousands to hundreds
	Multiplication (whole numbers)	3	- One-digit factor times two-digit factor with no carrying - One-digit factor times two-digit factor with carrying - One-digit factor times two-digit factor (problems written horizontally)
	Division (whole numbers)	3	- Two-digit dividend; one-digit divisor; one-digit quotient; no remainder

Table 23. Example Item Difficulty Estimates from Technical Report 0802

Table B1.
Estimates of item difficulty for grade 3.

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Correct	192	181	175	154	140	139	147	149	149	158	141	155	136	142	141
Incorrect	25	36	42	63	77	78	70	68	68	59	76	62	81	75	76
Valid	217	217	217	217	217	217	217	217	217	217	217	217	217	217	217
Missing	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
Valid percent	.88	.83	.81	.71	.65	.64	.68	.69	.69	.73	.65	.71	.63	.65	.65

Table 24. Example Item Statistics from Technical Report 0802

Table C1.
Estimates of item difficulty for grade 3.

Entry	Measure	Count	Score	OUT.MSQ	OUT.ZSTD	PTME	OBSMATCH	EXPMATCH
1	34.77	180	156	2.36	3.14	0.3	87.80	87.10
2	40.23	180	145	1.48	1.81	0.39	80.60	82.10
3	42.73	180	139	1.08	0.45	0.44	82.20	79.60
4	50.17	180	118	1.18	1.28	0.48	69.40	74.60
5	54.53	180	104	1.03	0.33	0.55	71.70	72.80
6	54.84	180	103	0.78	-2.10	0.66	82.20	72.70
7	52.38	180	111	0.82	-1.46	0.62	80.00	73.50
8	51.76	180	113	0.73	-2.28	0.64	81.10	73.60
9	51.76	180	113	0.81	-1.53	0.62	80.00	73.60
10	48.86	180	122	0.89	-0.72	0.56	77.20	75.20
11	54.23	180	105	1.20	1.67	0.49	66.70	72.90
12	49.84	180	119	1.41	2.60	0.41	65.60	74.70
13	55.74	180	100	1.18	1.58	0.53	69.40	72.50
14	53.92	180	106	0.77	-2.14	0.64	77.20	73.00
15	54.23	180	105	0.90	-0.87	0.58	73.30	72.90

Table 25. Key Findings Summary from Technical Report 0802

Category	Summary
Sample	Over 1,300 students in Grades 3–8 from twelve schools
Assessment Forms	Grade-specific mathematics computation GOMs
Analysis Method	CTT, Rasch (1PL), and 2PL IRT models
Items Analyzed	Numbers and Operations computation items
Problematic Items	Three items showed misfit or unstable response patterns
Item Fit	Most items demonstrated productive Rasch model fit
Overall Conclusion	Measures show strong technical adequacy for progress monitoring

Reference

Jung, E., Liu, K., Ketterlin-Geller, L. R., & Tindal, G. (2008). *Instrument development procedures for mathematics measures (Technical Report 0802).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0804: Examining Item Functioning of Math Screening Measures for Grades 1–8 Students (Liu et al., 2008).

This technical report examines the psychometric functioning of **mathematics screening measures** designed for students in Grades 1–8. The purpose of the study was to evaluate item difficulty and item fit across multiple grade levels and testing occasions using Item Response Theory (IRT), ensuring that the measures function as reliable general outcome measures (GOMs) for screening students at risk of not meeting grade-level mathematics standards.

Methods

Participants included approximately 6,500 students in Grades 1–8 from two local school districts in the Pacific Northwest. Students were assessed during the fall, winter, and spring of the 2006–2007 school year as part of regular classroom instruction. Sample sizes varied by grade, ranging from approximately 400 students in Grade 7 to over 1,500 students in Grade 5. No demographic data were collected. Each grade-level assessment consisted of three parallel forms corresponding to the three testing periods, resulting in a total of 24 test forms.

The BRT Math Screening Measures were aligned with the *Oregon Mathematics Curriculum Standards* and covered five domains of mathematics. Each assessment included computation items measuring procedural fluency and application items measuring conceptual understanding and problem-solving. Most items were multiple-choice with four response options, except for Grade 1 computation items, which required constructed responses. Calculators were not allowed for computation items but were permitted for application items. Assessments were administered in a paper-and-pencil format, typically within a 45-minute session. Item analyses were conducted using a one-parameter logistic (1PL) Rasch model implemented in WINSTEPS (version 3.61). Data preparation involved compiling student responses into grade- and season-specific datasets, importing them into SPSS, and then analyzing them in WINSTEPS. Key statistics examined included item difficulty estimates, and outfit mean square (MNSQ) fit statistics. Items with outfit MNSQ values between 0.50 and 1.50 were considered productive. Items outside this range were flagged for further inspection using distractor analysis to determine whether unexpected response patterns were due to incorrect answer keys or item flaws.

Results

Across approximately 1,000 total items, the vast majority demonstrated acceptable fit to the Rasch model. Forty-five items were initially identified as problematic. Of these, nine items contained incorrect answer keys and were retained after correction, six items exhibited severe misfit (outfit MNSQ > 2.0) and were removed from the item bank, and thirty items showed moderate misfit but were retained due to their instructional utility. Item difficulty

distributions within grades were comparable across fall, winter, and spring forms, supporting the use of the measures for progress monitoring. Observed and expected response patterns closely aligned, and student scale scores showed systematic increases over time, indicating sensitivity to growth. Overall, the results support the technical adequacy of the BRT Math Screening Measures for screening and monitoring mathematics performance across Grades 1–8.

Technical Report 804 further describes the development and technical evaluation of **mathematics computation curriculum-based measures** designed for use in progress monitoring with students in Grades 3 through 8. The primary purpose of the study was to examine the reliability, comparability, and sensitivity of computation measures intended for repeated administration within a response-to-intervention framework. The measures focused on grade-appropriate number and operations skills aligned with curricular expectations.

Participants included more than one thousand students recruited from public schools across multiple grade levels. Data collection occurred during scheduled assessment windows, with students completing grade-specific computation forms under standardized testing conditions. Responses were scored using digits-correct procedures, yielding fluency-based scores commonly used in computation CBMs to support instructional decision making.

Item and form development emphasized broad coverage of grade-level computation content while minimizing construct-irrelevant variance. Multiple equivalent forms were constructed for each grade level to allow frequent reassessment without compromising score interpretability. Anchor items were embedded across forms to support equating and evaluation of form comparability.

Statistical analyses incorporated both Classical Test Theory and item response modeling approaches. Rasch (1PL) analyses were conducted to evaluate item difficulty, fit statistics, and measurement precision across grades. Complementary CTT analyses examined score distributions, reliability coefficients, and inter-form correlations. Items demonstrating misfit or unstable parameter estimates were reviewed and removed when appropriate.

Results indicated that most items demonstrated acceptable fit to the Rasch model and contributed meaningfully to measurement precision. Inter-form correlations were strong, supporting the equivalence of alternate forms. Overall findings provide evidence that the mathematics computation measures are technically adequate and suitable for progress monitoring and instructional decision making across elementary and middle school grades.

Table 26. Example Results from Technical Report 0804

Table 2.
Grade 1 Fall 2006 Data.

Item	Measure	Count	Score	Out. Msq.	Out. ZSTD	Obs. Match	Exp. Match
1	-3.67	1262	1143	4.76	8.16	90.7	92.8
2	-2.79	1262	1075	0.67	-1.73	90.4	90.6
3	-3.21	1262	1110	0.8	-0.84	90.8	91.7
4	-0.95	1262	861	1.75	4.65	79.6	84
5	-2.09	1262	1006	2.02	4.5	88.5	88.2
6	-2.51	1262	1049	0.63	-2.07	92.9	89.7
7	2.37	1262	306	8.5	9.91	83.1	84
8	-1.96	1262	992	0.87	-0.73	89.8	87.8
9	3.1	1262	211	1.71	3.25	89.9	87.8
10	2.31	1262	315	2.62	7.09	87.3	83.6
11	5.77	1262	36	1.08	0.38	97.1	97.1
12	-0.45	1262	784	1.07	0.6	82.9	82
13	0.68	1262	591	0.94	-0.43	84.7	78.9
14	-1.79	1262	972	0.65	-2.29	89.8	87.1
15	0.54	1262	616	1.13	1.04	82.1	79.3
16	3.03	1262	220	1.85	3.84	88.1	87.4
17	-1.25	1262	903	0.67	-2.52	88.2	85.2
18	2.6	1262	273	2.03	4.73	88.1	85.2
19	1.23	1262	491	0.97	-0.13	80.6	78.9
20	-0.96	1262	863	0.93	-0.5	86.8	84

Table 27. Key Findings Summary from Technical Report 0804

Category	Summary
Sample	≈6,500 students in Grades 1–8
Assessment Forms	24 total forms (fall, winter, spring for each grade)
Analysis Method	1PL Rasch IRT model (WINSTEPS 3.61)
Items Analyzed	≈1,000 mathematics items
Problematic Items	45 flagged; 6 removed, 9 corrected, 30 retained
Item Fit	Majority within acceptable outfit MNSQ range (0.50–1.50)
Overall Conclusion	Measures demonstrated strong item functioning and growth sensitivity

Reference

Liu, K., Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2008). *Examining item functioning of math screening measures for Grades 1–8 students (Technical Report 0804)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0916: IRT Analysis of General Outcome Measures in Grades 1 – 8 (Alonzo, Anderson, et al., 2009).

This technical report presents an item response theory (IRT) analysis of **mathematics general outcome measures** designed for use in Grades 1 through 8. The primary purpose of the study was to evaluate the scaling properties, item functioning, and technical adequacy of fall screening assessments aligned with the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Point Standards. These measures were intended to support early identification of students at risk for mathematics difficulties and to inform instructional decision-making within progress monitoring and Response to Intervention (RTI) frameworks.

Methods

Participants were drawn from two mid-sized school districts in Oregon during the fall of the 2009 school year. Across grades, sample sizes ranged from approximately 900 to over 2,100 students per grade level, resulting in a large, combined dataset suitable for IRT calibration. Demographic data were approximated using data collected during the prior academic year. Participation was voluntary, and assessments were administered in school computer labs using a standardized, web-based testing platform.

The assessment design consisted of 48-item tests at each grade level. Each test included three 16-item subtests aligned with the major NCTM focal point domains relevant to that grade. Items were developed using a structured item-writing process grounded in principles of universal design, with attention to simplified language, reduced syntactic complexity, and accessibility for diverse learners. Items were reviewed for bias and sensitivity, and mathematics-specific vocabulary was retained to preserve construct validity.

Data collection procedures emphasized standardized administration. Items were presented individually on screen with three response options, which were randomly rotated to reduce copying. Student responses were coded to capture selected option, correctness, and focal point domain. Following data collection, responses were prepared for analysis by organizing item-level data by grade and domain.

Statistical analyses were conducted using a one-parameter logistic (1PL) Rasch model to calibrate items across all grade levels. Analyses focused on item difficulty estimates, standard errors, outfit mean square statistics, point-measure correlations, discrimination indices, and comparisons of observed versus expected performance. Item difficulty estimates were centered around zero within grades, with most items falling between −3 and +3 logits, indicating strong potential to differentiate student ability. Outfit statistics generally clustered near 1.0, suggesting good model fit. Measurement error was low across grades, particularly for most items at each grade level.

Results

Results indicated that items across Grades 1–8 functioned well under the Rasch model. Focal point domains were generally well distributed in difficulty within grades, although relative difficulty varied across domains. Geometry tended to be the easiest domain at most grade levels, while algebra-related domains were typically more challenging. Point-measure correlations were low to moderate but consistent with expectations for broad screening measures and observed scores closely matched model-predicted values.

Overall, findings support the technical adequacy of the mathematics general outcome measures for Grades 1–8. The calibrated item pools provide reliable, standards-aligned assessments capable of distinguishing students across a wide range of abilities and instructional decision-making in progress monitoring systems.

Table 28. Summary of Key Findings from Technical Report 0916

Category	Summary
Grade Levels	Grades 1–8
Participants	Approximately 900–2,100 students per grade
Assessment Structure	48-item tests with three 16-item focal point subtests
Statistical Model	1PL Rasch IRT model
Item Fit	Outfit statistics centered near 1.0
Difficulty Range	Most items between −3 and +3 logits
Primary Outcome	Reliable, scalable mathematics screening measures

Reference

Alonzo, J., Anderson, D., & Tindal, G. (2009). *IRT analysis of general outcome measures in grades 1–8 (Technical Report 09-16).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0921: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Kindergarten (Alonzo & Tindal, 2009b).

This technical report presents the development and validation of Kindergarten **mathematics progress monitoring measures** intended for use with both general education students and the federally defined “2% population” of students with disabilities who are assessed on grade-level content using modified achievement expectations. The primary purpose of the study was to create developmentally appropriate, psychometrically sound measures capable of detecting short-term growth in early mathematics skills while adhering to principles of Universal Design for Assessment (UDA).

Methods

Participants included approximately 2,800 Kindergarten students drawn from schools across the United States. Teachers and schools volunteered to participate through recruitment efforts conducted via the easyCBM™ and DIBELS platforms, professional networks, and district partnerships. To ensure confidentiality, no identifying information about students, teachers, schools, or districts was collected. Piloting took place during November and December of 2008. Assessments were administered online under teacher supervision, with students allowed to use scratch paper if needed. Calculators were not permitted.

Item development was guided by the *National Council of Teachers of Mathematics (NCTM) Focal Point Standards* for Kindergarten mathematics. A team of trained item writers with expertise in mathematics education, early childhood education, special education, assessment, and cognitive development created the Kindergarten item pool. Item writers were instructed to reduce both cognitive and linguistic complexity while maintaining alignment with grade-level standards. Items focused on a single mathematical construct, minimized working memory

demands, and relied heavily on visual representations appropriate for young learners. All items were presented in multiple-choice format with three response options and an "I don't know" option to reduce guessing behavior.

Items were delivered through an online assessment interface designed to support accessibility and consistency across administrations. Each testing session included 25 items. The first 20 items were randomly drawn from the Kindergarten item pool, while the final five items were fixed anchor items. These anchor items spanned focal point domains and difficulty levels and were used to calibrate all items to a common measurement scale in the grade.

Data analysis was conducted using a one-parameter logistic (1PL) Rasch model implemented in Winsteps 3.61. Analyses focused on item difficulty estimates, standard errors, mean square outfit statistics, and distractor functioning. Items with outfit statistics outside the acceptable range of 0.50 to 1.50 were examined individually rather than removed automatically. Distractor analyses evaluated whether students with higher estimated ability selected correct responses more frequently than lower-ability students, providing evidence of item functioning. For Kindergarten, a total of 173 mathematics items were analyzed. Most items demonstrated acceptable fit to the Rasch model and appropriate distractor functioning. Items that did not adequately differentiate between higher- and lower-ability students were removed from the item bank, while others with minor fit issues were retained when distractor patterns supported their validity. The final calibrated item bank covered a wide range of difficulty levels, making it suitable for both general education students and those in the 2% population.

Results

Using the calibrated item bank, researchers developed 30 alternate Kindergarten progress monitoring forms aligned with key NCTM focal point domains. Each form consisted of 16 items with closely matched mean difficulty levels to ensure comparability across administrations. Overall, the findings indicate that the Kindergarten mathematics progress monitoring measures are reliable, valid, developmentally appropriate, and instructionally useful for monitoring early mathematics development.

Table 29. Key Findings Summary from Technical Report 0921

Category	Summary
Sample Size	Approximately 2,800 Kindergarten students nationwide
Items Analyzed	173 Kindergarten mathematics items
Assessment Platform	easyCBM® online system
Statistical Model	1PL Rasch model (Winsteps 3.61)
Item Quality	Most items showed acceptable fit and effective distractor functioning
Progress Monitoring Forms	30 alternate forms, 16 items per form
Design Emphasis	Reduced cognitive load, visual supports, and simple language

Reference

Alonzo, J., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Kindergarten (Technical Report 0921)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0919 : The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 1 (Alonzo & Tindal, 2009a).

This technical report documents the development and validation of Grade 1 **mathematics progress monitoring measures** designed for use with both general education students and the federally defined "2% population" of students with disabilities who are assessed on grade-level content using modified achievement expectations. The primary objective was to develop reliable, sensitive measures capable of detecting short-term growth in early mathematics skills while adhering to principles of Universal Design for Assessment (UDA).

Participants included approximately 2,800 Grade 1 students drawn from schools across the United States. Teachers volunteered to participate through recruitment on the easyCBM® and DIBELS websites, existing district partnerships, and professional networks. No identifying information about students, teachers, schools, or districts was collected. Item piloting occurred between November and December 2008. All assessments were administered online under teacher supervision. Students were allowed to use scratch paper, but calculators were not permitted.

Methods

Item development was guided by the *National Council of Teachers of Mathematics (NCTM) Focal Point Standards*. A team of eight trained item writers with expertise in mathematics education, special education, assessment, and developmental psychology created the Grade 1 item pool. Writers were instructed to reduce cognitive and linguistic complexity while preserving alignment with grade-level content standards. Items emphasized single mathematical constructs, minimized working memory demands, and used simple, developmentally appropriate language. All items were multiple-choice with three response options and "I don't know" to reduce guessing. Items were delivered through an online assessment interface designed to support accessibility and consistency. Each student completed 25 items per testing session. The first 20 items were randomly selected from the Grade 1 item pool, while the final five anchor items were fixed across administrations. These anchor items spanned focal point domains and difficulty levels and were used to place all items on a common measurement scale.

Statistical analyses were conducted using a one-parameter logistic (1PL) Rasch model implemented in Winsteps 3.61. Analyses focused on item difficulty estimates, standard errors, mean square outfit statistics, and distractor functioning. Items with outfit statistics outside the acceptable range of 0.50 to 1.50 were reviewed individually rather than removed automatically. Distractor analyses examined whether students with higher estimated ability consistently selected correct responses while lower-ability students selected incorrect options. For Grade 1, a total of 243 items were analyzed. Overall, most items demonstrated acceptable fit to the Rasch model and appropriate distractor functioning. Items that failed to differentiate adequately between higher- and lower-ability students were removed from the item bank, while others were retained despite minor fit issues when distractor patterns supported their validity. The final calibrated item bank covered a broad range of difficulty levels, supporting use with both general education and 2% populations.

Results

Based on the calibrated items, researchers developed 30 alternate Grade 1 progress monitoring forms aligned with key NCTM focal point domains. Each form consisted of 16 items with closely matched mean difficulty levels to ensure comparability across administrations. Collectively, the findings indicate that the Grade 1 mathematics progress monitoring measures are psychometrically sound, instructionally useful, and well suited for tracking early mathematics development in diverse learner populations.

Table 30. Key Findings Summary from Technical Report 0919

Category	Summary
Sample Size	Approximately 2,800 Grade 1 students nationwide
Items Analyzed	243 Grade 1 mathematics items
Assessment Platform	easyCBM® online system
Statistical Model	1PL Rasch model (Winsteps 3.61)

Category	Summary
Item Quality	Most items demonstrated acceptable fit and effective distractor functioning
Progress Monitoring Forms	30 alternate forms, 16 items per form
Design Focus	Reduced cognitive and linguistic complexity with grade-level alignment

Reference

Alonzo, J., Lai, C. F., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 1 (Technical Report 0919)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0920: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 2 (Alonzo, Lai, et al., 2009c).

This technical report describes the development and validation of Grade 2 **mathematics progress monitoring measures** designed for use with both general education students and the federally defined "2% population" of students with disabilities who are assessed on grade-level content using modified achievement expectations. The overarching objective was to construct psychometrically sound, instructionally sensitive measures that could detect short-term growth in mathematics while adhering to principles of Universal Design for Assessment (UDA).

Methods

Participants consisted of approximately 2,800 Grade 2 students recruited from schools across the United States. Schools and teachers volunteered through the easyCBM® and DIBELS websites, direct district partnerships, and professional networks. To protect confidentiality, no identifying student, teacher, or school information was collected. Item piloting occurred between November 10 and December 5, 2008. All assessments were administered online under teacher supervision. Students were permitted to use scratch paper, but calculators were not allowed.

Item development was grounded in the *National Council of Teachers of Mathematics (NCTM) Focal Point Standards*. Eight trained item writers with backgrounds in mathematics education, special education, assessment, and developmental psychology produced approximately 1,100 Grade 2 items. Item writers were explicitly instructed to reduce cognitive and linguistic complexity while maintaining alignment with grade-level standards. Items were designed to focus on a single mathematical construct, minimize working memory demands, and use vocabulary well below grade level when possible. All items were multiple choice with three response options plus an "I don't know" option to reduce guessing. Items were delivered through an online interface designed to support accessibility. Each item was presented individually on screen with randomized answer order, except for the fixed "I don't know" option. Each testing session included 25 items: 20 randomly selected items from the item pool and five fixed anchor items spanning focal point domains and difficulty levels. These anchor items enabled all items to be on a common measurement scale.

Data analysis employed a one-parameter logistic (1PL) Rasch model using Winsteps 3.61. Key parameters examined included item difficulty (measure), standard error, mean square outfit statistics, and distractor functioning. Items with outfit values outside the acceptable range of 0.50 to 1.50 were reviewed in greater detail. Distractor analyses focused on whether higher-ability students consistently selected correct responses while lower-ability students selected incorrect options.

For Grade 2, a total of 1,167 items were analyzed. Thirty-seven items exhibited overfit statistics and were retained due to appropriate distractor functioning. Ninety-seven items showed underfit; of these, 47 were removed because higher-ability students were more likely to select incorrect answers, while the remaining 50 were

retained. The final calibrated item bank demonstrated a wide range of difficulty suitable for both general education and 2% populations.

Results

Using the refined item bank, researchers constructed 30 alternate progress monitoring forms aligned with three Grade 2 focal areas: Numbers and Operations, Geometry, and Numbers and Operations with Algebra. Each form contained 16 items with closely matched mean difficulty levels. Geometry forms were the easiest on average, followed by Numbers and Operations, while Numbers and Operations with Algebra forms were the most challenging. The results support the reliability, validity, and instructional utility of the Grade 2 progress measures.

Table 31. Key Findings Summary from Technical Report 0920

Category	Summary
Sample Size	Approximately 2,800 Grade 2 students nationwide
Items Analyzed	1,167 Grade 2 mathematics items
Assessment Platform	easyCBM® online system
Statistical Model	1PL Rasch model (Winsteps 3.61)
Item Retention	1,120 retained; 47 removed due to poor distractor functioning
Progress Monitoring Forms	30 forms (10 per focal area), 16 items each
Easiest Domain	Geometry
Most Challenging Domain	Numbers and Operations with Algebra

Reference

Alonzo, J., Lai, C. F., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 2 (Technical Report 0920)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0902: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 3 (Alonzo, Lai, et al., 2009a).

This technical report describes the development and validation of Grade 3 **mathematics progress monitoring measures** designed for use with both the general education population and the federally defined "2% population" of students with disabilities who are assessed on grade-level content using modified achievement standards. The primary goal was to create reliable, sensitive measures aligned to grade-level standards that could detect short-term growth while minimizing construct-irrelevant barriers.

Methods

The Grade 3 pilot involved students recruited nationally through participating teachers using the easyCBM® online assessment platform. Approximately 2,800 students per grade participated across the broader project, with Grade 3 students completing online assessments between November and December 2008. Each student completed a 25-item test: 20 items randomly selected from a large Grade 3 item bank and 5 fixed anchor items. These anchor items, identical across all test administrations and ordered consistently, enabled calibration of all items onto a common measurement scale. Calculators were not permitted, though students could use scratch paper. "I don't know" was included to reduce guessing behavior.

Grade 3 items were aligned to *National Council of Teachers of Mathematics (NCTM) Focal Point Standards*, spanning Number and Operations, Number and Operations with Algebra, and Geometry. Items were intentionally designed using principles of Universal Design for Assessment, emphasizing reduced linguistic and cognitive

complexity while preserving grade-level rigor. A multi-stage review process involving six trained researchers ensured clarity, standard alignment, and technical accuracy before piloting.

Item responses were analyzed using a one-parameter logistic (1PL) Rasch model implemented in Winsteps software. The Rasch approach was selected for parsimony and interpretability. Analyses focused on item difficulty estimates, standard errors, Mean Square Outfit statistics, and distractor functioning. Items with outfit values outside the acceptable range of 0.50 to 1.50 were flagged for closer review. Distractor analyses examined whether higher-ability students consistently selected correct responses while lower-ability students selected distractors.

A total of 1,167 Grade 3 items were analyzed. Of these, 92 items showed overfit and 102 showed underfit statistics. All overfitting items demonstrated appropriate distractor functioning and were retained. Underfitting items were examined by content domain, resulting in the removal of 38 items that failed to function as intended. The final Grade 3 item bank contained 1,111 items.

Results

Using calibrated item difficulty estimates, researchers constructed 30 alternate progress monitoring forms (10 per focal point domain), each consisting of 16 items. Forms within each domain demonstrated highly comparable difficulty levels. Geometry forms were the easiest overall, followed by Number and Operations with Algebra, while Number and Operations forms were the most challenging. These results support the technical adequacy of the Grade 3 measures for monitoring progress across a wide range of student abilities.

Table 32. Key Findings Summary from Technical Report 0902

Category	Summary
Total items analyzed	1,167
Items retained in final bank	1,111
Statistical model	1PL Rasch model
Overfitting items	92 (all retained)
Underfitting items removed	38
Progress monitoring forms created	30 (10 per focal point)
Easiest domain	Geometry
Most difficult domain	Number and Operations

Reference

Alonzo, J., Lai, C. F., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 3 (Technical Report 0902)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0903: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 4 (Alonzo, Lai, et al., 2009b).

This technical report documents the development and validation of Grade 4 **mathematics progress monitoring** measures designed for both the general education population and the federally defined “2% population” of students with disabilities who are assessed on grade-level content with modified achievement expectations. The primary goal was to create reliable, sensitive measures capable of detecting short-term growth in mathematics skills while adhering to principles of Universal Design for Assessment.

Methods

Participants were drawn from a national sample of schools across the United States. Approximately 2,800 Grade 4 students participated during the pilot testing window, which ran from November 10 to December 5, 2008. No identifying information was collected to ensure confidentiality. Students completed the assessments online through the easyCBM® platform under teacher supervision, without calculators. Scratch paper was permitted.

Item development was grounded in the *National Council of Teachers of Mathematics (NCTM) Focal Point Standards* and emphasized reduced cognitive and linguistic complexity. Eight trained item writers with backgrounds in mathematics, special education, and assessment produced approximately 1,100 Grade 4 items. Items were written to focus on a single mathematical concept, minimize language load, and use accessible vocabulary. All items were multiple-choice with three options plus "I don't know" to reduce random guessing. Graphics were professionally developed, and the computer interface was designed to display one item at a time with randomized answer order.

Each student received 25 items per testing session. The first 20 items were randomly drawn from the item pool, while the final five anchor items were constant across administrations to allow all items to be placed on a common measurement scale. Data were analyzed using a one-parameter logistic (1PL) Rasch model implemented in Winsteps 3.61. Item difficulty estimates, standard errors, mean square outfit statistics, and distractor functioning were examined. Acceptable outfit values ranged from 0.50 to 1.50, though some items outside this range were retained if distractor analyses showed appropriate response patterns.

For Grade 4, 1,149 items were analyzed. Eighty-four items exhibited overfit statistics but were retained due to strong distractor functioning. One hundred fourteen items showed underfit; of these, 80 were retained and 34 were removed because higher-ability students did not consistently select the correct answer. Overall, most items demonstrated good psychometric properties and covered a wide range of difficulty levels, supporting use with both general education and 2% populations.

Results

Using the calibrated item bank, researchers developed 30 alternate progress monitoring forms aligned with three Grade 4 focal areas: Measurement and Data Analysis, Numbers and Operations, and Numbers and Operations with Algebra. Each form contained 16 items, and mean difficulty levels were closely matched across forms. Measurement and Data Analysis forms were the easiest on average, followed by Numbers and Operations with Algebra, while Numbers and Operations forms were the most challenging. Grade 4 measures were psychometrically sound, instructionally useful, and suitable for monitoring student progress across diverse learner populations.

Table 33. Key Findings Summary from Technical Report 0903

Category	Summary
Sample Size	Approximately 2,800 Grade 4 students nationwide
Items Analyzed	1,149 Grade 4 mathematics items
Statistical Model	1PL Rasch model (Winsteps 3.61)
Item Retention	Most items retained; 34 removed due to poor distractor functioning
Difficulty Range	Wide range, supporting both general education and 2% populations
Progress Monitoring Forms	30 forms (10 per focal area), 16 items each
Easiest Domain	Measurement and Data Analysis
Most Challenging Domain	Numbers and Operations

Reference

Alonzo, J., Lai, C. F., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 4 (Technical Report 0903)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 0901: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 5 (Lai et al., 2009a).

This technical report documents the development and piloting of Grade 5 mathematics progress monitoring measures designed for use with both the general education population and the federally defined "2% population" of students with disabilities. The primary objective was to create reliable, growth-sensitive measures aligned with grade-level mathematics standards and suitable for use within a Response to Intervention (RTI) framework.

Methods

Participants included approximately 2,800 Grade 5 students drawn from schools across the United States. Teachers were recruited through the easyCBM® and DIBELS websites, direct district outreach, and existing research partnerships. Participation was voluntary, and no identifying information about students, teachers, schools, or districts was collected. Assessments were administered online during November and December of 2008 under teacher supervision. Students completed 25 multiple-choice items per session, were permitted to use scratch paper, and were not allowed calculators.

Item development was guided by the *National Council of Teachers of Mathematics (NCTM) Focal Point Standards* and principles of Universal Design for Assessment. Items were written to minimize linguistic and cognitive complexity while maintaining alignment with grade-level content. Eight trained item writers with backgrounds in mathematics education, special education, and assessment produced approximately 1,150 Grade 5 items. Each item targeted a single sub-domain within a focal point standard with three answer options plus and "I don't know".

Data collection followed a structured piloting design. Of the 25 items administered per session, 20 were randomly drawn from the Grade 5 item pool, while five anchor items appeared consistently across all forms to allow calibration onto a common scale. Response options were randomized to reduce order effects and cheating. All items were delivered through the easyCBM® online interface, designed for accessibility and consistent presentation.

Statistical analyses were conducted using a one-parameter logistic (1PL) Rasch model implemented in Winsteps. Item difficulty estimates, standard errors, mean square outfit statistics, and distractor functioning were evaluated. Items with outfit values outside the recommended 0.50–1.50 range were examined in detail. Overfitting items were generally retained if distractor analyses showed appropriate response patterns; underfitting items were kept or removed. Sixty-five Grade 5 items were removed due to poor distractor functioning.

Results indicated that most Grade 5 items demonstrated acceptable Rasch model fit and effective distractor functioning. The calibrated item bank spanned a wide range of difficulty levels, supporting measurement across diverse student ability levels. Using these calibrated items, researchers constructed ten alternate progress monitoring forms and three benchmark forms for each Grade 5 focal point domain. Mean difficulty values across forms were closely clustered, indicating strong alternate-form equivalence.

Results

Overall, findings support the technical adequacy of the Grade 5 mathematics progress monitoring measures. The assessments demonstrate reliable measurement, alignment with grade-level standards, and sensitivity to short-term growth, making them appropriate tools for instructional monitoring and decision-making in RTI systems.

Table 34. Sample of Key Content Summary from Technical Report 0901

Table 1
Results of Rasch Analysis, Grade 5

Item	Focal Point	Domain	Measure	Count	Score	Error	Mean Square Outfit	Discrim
50001	Number and Operations and Algebra	Apply understanding of models for division (e.g., equal-sized groups, arrays, area models, equal intervals on the number line), place value, properties of operations (commutative, associative, distributive) and the relationship of division to multiplication.	1.35	28	11	0.43	1.13	0.88
50002	Number and Operations and Algebra	Apply understanding of models for division (e.g., equal-sized groups, arrays, area models, equal intervals on the number line), place value, properties of operations (commutative, associative, distributive) and the relationship of division to multiplication.	0.57	37	21	0.36	0.91	1.22
50003	Number and Operations and Algebra	Apply understanding of models for division (e.g., equal-sized groups, arrays, area models, equal intervals on the number line), place value, properties of operations (commutative, associative, distributive) and the relationship of division to multiplication.	1.12	30	14	0.41	1.63	-0.13
50004	Number and Operations and Algebra	Apply understanding of models for division (e.g., equal-sized groups, arrays, area models, equal intervals on the number line), place value, properties of operations (commutative, associative, distributive) and the relationship of division to multiplication.	1.07	37	17	0.37	0.99	0.96
50005	Number and Operations and Algebra	Apply understanding of models for division (e.g., equal-sized groups, arrays, area models, equal intervals on the number line), place value, properties of operations (commutative, associative, distributive) and the relationship of division to multiplication.	0.85	1791	936	0.05	1.25	0.61
50006	Number and Operations and Algebra	Apply understanding of models for division (e.g., equal-sized groups, arrays, area models, equal intervals on the number line), place value, properties of operations (commutative, associative, distributive) and the relationship of division to multiplication.	-0.74	34	26	0.46	1.06	0.70
50007	Number and Operations and Algebra	Apply understanding of models for division (e.g., equal-sized groups, arrays, area models, equal intervals on the number line), place value, properties of operations (commutative, associative, distributive) and the relationship of division to multiplication.	-2.76	35	33	0.74	0.45	1.07

Table 35. Key Findings Summary from Technical Report 0901

Category	Summary
Grade Level	Grade 5
Participants	Approximately 2,800 students nationwide
Assessment Format	Online, multiple-choice with anchor items
Statistical Model	1PL Rasch model
Item Bank Size	Approximately 1,150 Grade 5 items
Forms Developed	10 progress monitoring forms and 3 benchmark forms per domain
Primary Outcome	Reliable, growth-sensitive mathematics measures

Reference

Lai, C. F., Alonzo, J., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 5 (Technical Report 0901)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0907: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 6 (Lai et al., 2009d).

This technical report documents the development, piloting, and psychometric evaluation of Grade 6 **mathematics progress monitoring measures** designed for use with both the general education population and the federally defined “2% population” of students with disabilities. The purpose of the study was to create universally designed, curriculum-aligned assessments capable of detecting short-term academic growth within RTI frameworks.

Methods

Approximately 2,800 Grade 6 students from schools across the United States participated in item piloting during November and December of 2008. Teachers were recruited through the easyCBM® and DIBELS websites, existing district partnerships, and professional networks. Data were collected using an online testing platform. Each student completed a 25-item assessment consisting of 20 randomly selected items from the Grade 6 item pool and five fixed anchor items. Calculators were not permitted, scratch paper was allowed, and an "I don't know" response option to reduce guessing behavior. No identifying student or school data were collected to ensure confidentiality.

Items were aligned to the *National Council of Teachers of Mathematics (NCTM) Focal Point Standards* and written using universal design principles to minimize linguistic and cognitive complexity while preserving alignment to grade-level content. The item pool targeted students across a wide ability range, including those in the 2% population. Extensive expert review ensured clarity, standard alignment, and appropriate distractor construction prior to piloting. Item calibration was conducted using a one-parameter logistic (1PL) Rasch model implemented in Winsteps version 3.61. Analyses focused on item difficulty (measure), standard error, and Mean Square Outfit statistics. Items with outfit values outside the recommended range of 0.50 to 1.50 were examined further through distractor analyses. Items were retained when higher-ability students consistently selected correct responses and lower-ability students selected distractors.

Results

A total of 953 Grade 6 items were analyzed. Of these, 43 items demonstrated overfit and 84 items demonstrated underfit. Distractor analysis supported retention of most items, resulting in the removal of only 16 items from the Grade 6 item bank. The final calibrated item pool supported the development of 30 progress monitoring forms (10 per focal point grouping) and nine benchmark screeners. Mean difficulty values within each focal point grouping were tightly clustered, indicating strong form equivalence. Measures aligned with Number and Operations involving ratios and rates were the least difficult, followed by Algebra measures, while measures focused on fraction and decimal operations were the most challenging. Overall, results support the technical adequacy and instructional utility of the Grade 6 progress monitoring measures.

Table 36. Key Findings Summary from Technical Report 0907

Category	Summary
Sample	≈2,800 Grade 6 students nationwide
Analysis Method	1PL Rasch model (Winsteps 3.61)
Items Analyzed	953 Grade 6 mathematics items
Item Retention	16 items removed after fit and distractor analysis
Forms Developed	30 progress monitoring forms; 9 benchmark screeners
Form Equivalence	Comparable difficulty within focal point groupings
Overall Conclusion	Measures demonstrated strong psychometric performance

Reference

Lai, C. F., Alonzo, J., & Tindal, G. (2008). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 6 (Technical Report 0907)*. Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0908: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 7 (Lai et al., 2009c).

This technical report documents the development and validation of a Grade-level **mathematics progress monitoring measure** intended for use with both general education students and students in the 2% population. The study established technical adequacy of the assessment through systematic item development, field testing, and psychometric evaluation, with particular attention to item functioning, reliability, and validity indicators.

Methods

Participants included a large, diverse sample of elementary students drawn from multiple school districts from both general education students and students eligible for the 2% alternate assessment. Inclusion criteria ensured appropriate grade-level placement, demographic representativeness and subgroup analyses.

Items were developed to align with *National Council of Teachers of Mathematics (NCTM)* focal point standards and administered under standardized testing conditions. Data were collected during scheduled assessment windows using paper-based instruments administered by trained personnel. Student responses were recorded dichotomously and compiled for psychometric analysis.

Results

Analyses were conducted using Item Response Theory (IRT) models to evaluate item difficulty, discrimination, and overall model fit. Classical Test Theory indices, including reliability estimates and item-total correlations, were also computed. Differential item functioning analyses were conducted to assess fairness across student subgroups. Results indicated that most items functioned as intended, with difficulty parameters centered near zero and acceptable fit statistics. Reliability estimates supported the use of the measure for progress monitoring purposes. Items demonstrated strong alignment with grade-level content standards, and score distributions suggested adequate sensitivity to differences in student ability levels.

Table 37. Key Findings Summary from Technical Report 0908

Category	Summary
Sample	≈2,800 Grade 7 students nationwide
Analysis Method	1PL Rasch model (Winsteps 3.61)
Items Analyzed	912 Grade 7 mathematics items
Item Fit	15 overfit, 51 underfit; all retained after distractor analysis
Forms Developed	30 progress monitoring forms; 9 benchmark screeners
Difficulty Structure	Comparable difficulty within focal point groupings
Overall Conclusion	Measures demonstrated strong psychometric performance

Reference

Lai, C. F., Alonzo, J., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 7 (Technical Report 0908)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0904: The Development of K–8 Progress Monitoring Measures in Mathematics for Use With the 2% and General Education Populations: Grade 8 (Lai et al., 2009b).

This technical report describes the development and piloting of Grade 8 **mathematics progress monitoring measures** intended for use with both the general education population and the federally defined “2% population” of students with disabilities. The overarching purpose of the study was to create reliable, growth-sensitive assessments aligned with grade-level mathematics standards and appropriate for use within an RTI framework.

Methods

Participants included approximately 2,800 Grade 8 students from schools across the United States. Teachers were recruited through announcements on the easyCBM® and DIBELS websites, existing district partnerships, and professional networks associated with BRT at the University of Oregon. Participation was voluntary, and no identifying information about students, teachers, schools, or districts was collected. Assessments were administered online during November and December of 2008 under teacher supervision. Students completed 25 multiple-choice items per testing session, were allowed to use scratch paper, and calculators were prohibited.

Item development was grounded in the *National Council of Teachers of Mathematics (NCTM) Focal Point Standards* and principles of Universal Design for Assessment. Eight trained item writers with backgrounds in mathematics education, special education, and assessment created approximately 900 Grade 8 items. Writers were instructed to reduce cognitive and linguistic complexity while preserving alignment with grade-level standards. Each item targeted a single mathematical construct and included three answer choices plus an "I don't know" option to reduce random guessing. Graphics and item presentation were designed to minimize construct-irrelevant barriers.

Data collection followed a structured piloting design. Of the 25 items administered per session, 20 were randomly selected from the Grade 8 item pool, while five anchor items appeared consistently across all test forms. These anchor items enabled calibration of all items onto a common measurement scale. Answer options were randomized for each item to reduce order effects and potential cheating.

Statistical analyses were conducted using a one-parameter logistic (1PL) Rasch model implemented in Winsteps. Item difficulty estimates, standard errors, and mean square outfit statistics were examined to evaluate item fit. Items with outfit values outside the recommended range of 0.50 to 1.50 were reviewed using distractor analyses. Overfitting items were retained when distractor patterns indicated appropriate functioning, while underfitting items were retained or removed based on construct validity with 28 items removed (poor distractor functioning).

Results

Results indicated that most Grade 8 items demonstrated acceptable fit to the Rasch model and effective distractor performance. The calibrated item bank covered a broad range of difficulty levels, supporting accurate measurement across students with varying levels of mathematical proficiency. Using the calibrated items, researchers constructed ten alternate progress monitoring forms and three benchmark forms for each Grade 8 focal point domain. Mean difficulty values across alternate forms were highly consistent with form equivalence. Overall, findings support the technical adequacy of the Grade 8 mathematics progress monitoring measures. The assessments are aligned with grade-level standards, sensitive to short-term growth, and suitable for monitoring student progress and informing instructional decision-making within RTI systems.

Table 38. Key Findings Summary from Technical Report 0904

Category	Summary
Grade Level	Grade 8
Participants	Approximately 2,800 students nationwide
Assessment Format	Online, multiple-choice with anchor items
Statistical Model	1PL Rasch model
Item Pool Size	Approximately 900 Grade 8 items
Forms Developed	10 progress monitoring forms and 3 benchmark forms per domain
Primary Outcome	Reliable, growth-sensitive Grade 8 math measures

Reference

Lai, C. F., Alonzo, J., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 8 (Technical Report 0904)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1314: The Development and Scaling of the easyCBM® CCSS Elementary Mathematics Measures: Grade K (Irvin, Saven, Alonzo, Park, Anderson, et al., 2013).

The purpose of the study was to design progress monitoring and benchmarking assessments that are developmentally appropriate, aligned with CCSS expectations, and capable of reliably measuring growth in early mathematics skills within a Response to Intervention (RTI) framework.

Methods

Participants consisted of a large national sample of Kindergarten students drawn from schools across the United States. Schools and teachers volunteered through existing easyCBM® partnerships and professional outreach efforts. To protect confidentiality, no identifying information about students, teachers, schools, or districts was collected. Assessments were administered online under teacher supervision during scheduled piloting windows, following standardized administration procedures consistent with classroom use. Students were permitted to use basic testing supports appropriate for Kindergarten learners. No instructional assistance was provided during testing.

Item development emphasized alignment with the Kindergarten CCSS mathematics standards and accessibility for diverse learners. Items were written by experienced educators with backgrounds in elementary mathematics instruction and assessment. Writers received training in effective item construction and principles of Universal Design for Assessment, with particular attention to minimizing linguistic complexity, reducing working memory demands, and using visual representations appropriate for young learners. Items targeted a single mathematical concept and were designed to avoid construct-irrelevant barriers that could disadvantage students with disabilities or limited language proficiency.
Data collection involved large-scale piloting of items across participating schools. Responses were analyzed using a one-parameter logistic (1PL) Rasch model. Item calibration placed all Kindergarten items on a common measurement scale, allowing for consistent interpretation of item difficulty and student performance. Key statistics examined included item difficulty estimates, standard errors, mean square outfit statistics, and distractor functioning. Items with poor fit or weak distractor patterns were reviewed and either revised or excluded from operational forms.

Following calibration, the item bank was used to assemble multiple alternate forms of Kindergarten mathematics assessments. Forms were designed for both progress monitoring and benchmarking purposes. Each form consisted of a carefully selected subset of items with comparable overall difficulty to ensure alternate form equivalence. Form equivalence was evaluated using test characteristic curves and test information functions, which demonstrated strong overlap across forms and consistent measurement precision across the ability range.

Results

Results indicated that the majority of Kindergarten items demonstrated acceptable fit to the Rasch model and appropriate distractor functioning. The calibrated item bank covered a wide range of difficulty levels, supporting measurement of students with varying levels of early mathematical understanding. The resulting assessment forms were shown to be psychometrically comparable and sensitive to growth, making them suitable for repeated administration within an RTI framework. Overall, the findings support the reliability, validity, and instructional utility of the easyCBM® CCSS Kindergarten mathematics measures.

The vertically aligned scaling approach contributes to coherent progress monitoring across grade levels, while the Kindergarten measures specifically provide educators with a robust tool for assessing early mathematics development and informing instructional decision-making.

Table 39. Example of Key CCSS Content Alignment Summary from Technical Report 1314

Table 1
Kindergarten Item Writing Plan by CCSS Standard

CCSS Standard	Item Set 1	Item Set 2	Item Set 3	Item Set 4	Existing BM Align	Total
CC1	5	5	6	6	0	22
CC2	1	1	1	1	6	4
CC3	6	6	5	5	1	22
CC4	1	1	1	1	5	4
CC5	5	5	6	6	3	22
CC6	1	1	1	1	5	4
CC7	6	6	5	5	2	22
G1	1	1	1	1	6	4
G2	1	1	1	1	7	4
G3	10	10	11	11	2	42
G4	1	1	1	1	5	4
G5	11	11	10	10	0	42
G6	1	1	1	1	8	4
MD1	4	4	4	4	5	16
MD2	2	2	2	2	9	8
MD3	19	19	19	19	4	76
NBT1	25	25	25	25	1	100
OA1	5	5	5	5	3	20
OA2	5	5	5	5	3	20
OA3	5	5	5	5	0	20
OA4	5	5	5	5	1	20
OA5	5	5	5	5	3	20
Set total	125	125	125	125	79	500

Note. Item Sets 1-4 reflect items written to each CCSS standard based on results from our previous alignment study (Irvin et al., 2012b), reflected in Existing BM Align column. Total represents the number of math items written to a given CCSS standard in the current study.

Table 40. Example of Key Piloting Plan from Technical Report 1314

Table 2
Grades K-2 Piloting Plan

Form	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	Anchor	Unique	Total
1	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	22	32
2	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	2	22	32
3	-	2	3	-	-	-	-	-	-	-	-	-	-	-	-	3	22	32
4	-	-	3	2	-	-	-	-	-	-	-	-	-	-	-	2	22	32
5	-	-	-	2	3	-	-	-	-	-	-	-	-	-	-	3	22	32
6	-	-	-	-	3	2	-	-	-	-	-	-	-	-	-	2	22	32
7	-	-	-	-	-	2	3	-	-	-	-	-	-	-	-	3	22	32
8	-	-	-	-	-	-	3	2	-	-	-	-	-	-	-	2	22	32
9	-	-	-	-	-	-	-	2	3	-	-	-	-	-	-	3	22	32
10	-	-	-	-	-	-	-	-	3	2	-	-	-	-	-	2	22	32
11	-	-	-	-	-	-	-	-	-	2	3	-	-	-	-	3	22	32
12	-	-	-	-	-	-	-	-	-	-	3	2	-	-	-	2	22	32
13	-	-	-	-	-	-	-	-	-	-	-	2	3	-	-	3	22	32
14	-	-	-	-	-	-	-	-	-	-	-	-	3	2	-	2	22	32
15	-	-	-	-	-	-	-	-	-	-	-	-	-	2	3	0	22	32

Form	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	Anchor	Unique	Total
1	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	22	32
2	2	3	-	-	-	-	-	-	-	-	-	-	-	-	-	3	22	32
3	-	3	2	-	-	-	-	-	-	-	-	-	-	-	-	2	22	32
4	-	-	2	3	-	-	-	-	-	-	-	-	-	-	-	3	22	32
5	-	-	-	3	2	-	-	-	-	-	-	-	-	-	-	2	22	32
6	-	-	-	-	2	3	-	-	-	-	-	-	-	-	-	3	22	32
7	-	-	-	-	-	3	2	-	-	-	-	-	-	-	-	2	22	32
8	-	-	-	-	-	-	2	3	-	-	-	-	-	-	-	3	22	32
9	-	-	-	-	-	-	-	3	2	-	-	-	-	-	-	2	22	32
10	-	-	-	-	-	-	-	-	2	3	-	-	-	-	-	3	22	32
11	-	-	-	-	-	-	-	-	-	3	2	-	-	-	-	2	22	32
12	-	-	-	-	-	-	-	-	-	-	2	3	-	-	-	3	22	32
13	-	-	-	-	-	-	-	-	-	-	-	3	2	-	-	2	22	32
14	-	-	-	-	-	-	-	-	-	-	-	-	2	3	-	3	22	32
15	-	-	-	-	-	-	-	-	-	-	-	-	-	3	2	0	22	32

Note. C = CCSS pool anchor item; N = NCTM pool anchor item. Anchor items appearing in a vertical column (both CCSS or NCTM) were shared between the specified forms. For example, form 3 and form 4 shared 3 anchor items from the CCSS pool (set C3) and 2 from the NCTM pool (set N3).

Table 41. Key Findings Summary from Technical Report 1314

Category	Summary
Grade Level	Kindergarten
Assessment Alignment	Common Core State Standards (CCSS)
Assessment Platform	easyCBM® online system
Statistical Model	1PL Rasch model
Item Bank Quality	Most items showed acceptable fit and effective distractor functioning
Forms Developed	Multiple equivalent forms for progress monitoring and benchmarking
Primary Outcome	Reliable, growth-sensitive Kindergarten math measures

Reference

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., Anderson, D., & Tindal, G. (2013). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade K (Technical Report # 1314).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1315: The Development and Scaling of the easyCBM® CCSS Elementary Mathematics Measures: Grade 1 (Saven, Irvin, Park, Tindal, et al., 2013).

This technical report describes the development, piloting, and scaling of the easyCBM® Common Core State Standards (CCSS) **Grade 1 mathematics measures** for use within a *Response to Intervention (RTI)* framework. The primary goal was to create technically adequate benchmark and progress-monitoring assessments aligned with CCSS and appropriate for diverse student populations.

Methods

Participants included 1,124 Grade 1 students taught by 329 teachers across 140 schools in 132 school districts spanning 33 U.S. states. Data were collected during a national online pilot conducted between May 15 and June 15, 2013. To protect confidentiality, no demographic information was collected. Students were automatically assigned one of 15 pilot test forms through the secure easyCBM® online platform, ensuring balanced participation across forms. Each pilot form consisted of 32 multiple-choice items, and student responses were automatically recorded. Calculators were not permitted, and answer options were randomly rotated to minimize cheating.

A total of 500 Grade 1 CCSS-aligned mathematics items were developed as part of a larger K–5 item pool. Item writers and reviewers averaged approximately 14 years of mathematics teaching experience and participated in structured training focused on CCSS alignment, principles of effective item writing, and Universal Design for Assessment. Items underwent three stages of review: contracted expert review, internal university researcher review, and external independent review. Graphics and audio supports were developed where necessary to improve accessibility. Only items meeting criteria for clarity, accuracy, alignment, and lack of bias from piloting.

All items were calibrated using a one-parameter logistic (1PL) Rasch model with concurrent equating, implemented in WINSTEPS version 3.6.8. Horizontal anchor items from both newly developed CCSS items and previously validated NCTM-aligned items were used to link pilot forms to a common scale. Item difficulty estimates (β) and outfit mean square (MNSQ) statistics were examined. Items with MNSQ values outside the acceptable range of 0.50 to 1.50 were removed from the item bank prior to test form construction. Distractor analyses evaluated whether incorrect response options functioned as intended across varying levels of student ability.

Results

Results indicated that most Grade 1 items demonstrated acceptable fit to the Rasch model and effective distractor functioning. Thirteen operational test forms were constructed: three seasonal benchmark forms (fall, winter,

spring) and ten progress-monitoring forms. Benchmark forms included vertical anchor items linking adjacent grade levels to support future vertical scaling. The average difficulty across Grade 1 benchmark and progress-monitoring forms was approximately -0.01 logits, with minimal variation, indicating strong form equivalence. Observed response patterns closely aligned with model expectations, supporting the technical adequacy and growth sensitivity of the Grade 1 easyCBM® CCSS mathematics measures for RTI applications.

Table 42. Key Findings Summary from Technical Report 1315

Item	Summary
Participants	1,124 Grade 1 students across 33 states
Assessment Design	15 pilot forms; online administration
Analysis Method	1PL Rasch model with concurrent equating
Item Pool	500 CCSS-aligned Grade 1 items
Item Fit	Majority within acceptable MNSQ range (0.50–1.50)
Final Output	13 equivalent benchmark and progress-monitoring forms

Reference

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., Anderson, D., & Tindal, G. (2013). *The development and scaling of the easyCBM® Common Core State Standards elementary mathematics measures: Grade 1 (Technical Report 1315).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1316: The Development and Scaling of the easyCBM® CCSS Elementary Mathematics Measures: Grade 2 (Irvin, Saven, et al., 2013a).

This technical report describes the development and scaling of the easyCBM® Common Core State Standards (CCSS*)* **elementary mathematics measures for Grade 2**. The purpose of the study was to create progress monitoring and benchmarking assessments aligned with CCSS expectations that are sensitive to student growth, psychometrically sound, and suitable for use within a Response to Intervention (RTI) framework.

Methods

A large national sample of Grade 2 students were drawn from schools across the United States. Schools and teachers volunteered to participate through established easyCBM® partnerships and outreach efforts from BRT at the University of Oregon. To ensure confidentiality, no identifying information about students, teachers, schools, or districts was collected. All assessments were administered online under teacher supervision during scheduled piloting windows and followed standardized administration procedures reflective of typical classroom use.
Item development emphasized close alignment with the Grade 2 CCSS mathematics standards and accessibility for diverse learners. Items were written by experienced elementary educators and content specialists with backgrounds in mathematics instruction and assessment. Item writers were trained in effective item construction and Universal Design for Assessment principles, with particular emphasis on reducing linguistic complexity, minimizing working memory demands, and eliminating construct-irrelevant barriers. Each item targeted a single mathematical concept, designed to measure conceptual understanding and application not simple fluency.

Data collection involved large-scale piloting of the Grade 2 item pool across participating schools. Student responses were analyzed using a one-parameter logistic (1PL) Rasch model. Item calibration placed all Grade 2 items on a common measurement scale, allowing for consistent interpretation of item difficulty and student performance. Statistical analyses focused on item difficulty estimates, standard errors, mean square outfit statistics, and distractor functioning. Items that exhibited poor model fit or inappropriate distractor patterns were reviewed in detail and either revised or excluded from operational assessment forms. Following calibration, the

refined item bank was used to assemble multiple alternate forms of Grade 2 mathematics assessments. Forms were developed for both progress monitoring and seasonal benchmarking purposes. Each form contained a balanced selection of items with comparable mean difficulty to ensure alternate form equivalence. Equivalence across forms was evaluated using test characteristic curves (TCCs) and test information functions (TIFs), which demonstrated strong overlap and consistent measurement precision across the ability continuum.

Results

Results indicated that most Grade 2 items demonstrated acceptable fit to the Rasch model and effective distractor functioning. The calibrated item bank covered a broad range of difficulty levels, supporting assessment of students with varying levels of mathematical proficiency. The alternate forms were shown to be psychometrically comparable and sensitive to changes in student performance over time.

Overall, the findings support the reliability, validity, and instructional utility of the easyCBM® CCSS Grade 2 mathematics measures. These assessments provide educators with robust tools for monitoring progress, evaluating intervention effectiveness, and informing instructional decision-making within an RTI framework.

Table 43. Example of Key CCSS Content Standard Alignment from Technical Report 1316

Table 1
Second Grade Item Writing Plan by CCSS Standard

CCSS Standard	Item Set 1	Item Set 2	Item Set 3	Item Set 4	Existing BM Align	Total
G1	10	10	11	11	0	42
G2	10	11	10	10	3	41
G3	11	11	10	10	0	42
MD1	0	0	0	0	7	0
MD2	3	3	3	3	3	12
MD3	4	3	4	4	0	15
MD4	3	4	4	4	0	15
MD5	4	4	4	3	0	15
MD6	3	3	3	4	2	13
MD7	3	3	3	3	3	12
MD8	4	3	3	3	2	13
MD9	3	4	4	4	0	15
MD10	4	4	4	3	0	15
NBT1	0	0	0	0	10	0
NBT2	5	5	5	5	0	20
NBT3	5	5	5	5	0	20
NBT4	0	0	0	0	7	0
NBT5	1	1	1	2	5	5
NBT6	5	5	5	5	0	20
NBT7	5	5	5	5	2	20
NBT8	5	5	5	5	2	20
NBT9	5	5	5	5	0	20
OA1	2	1	1	1	7	5
OA2	10	10	10	10	3	40
OA3	10	10	10	10	0	40
OA4	10	10	10	10	1	40
Set total	125	125	125	125	57	500

Note. Item Sets 1-4 reflect items written to each CCSS standard based on results from our previous alignment study (Irvin et al., 2012b), reflected in Existing BM Align column. Total represents the number of math items written to a given CCSS standard in the current study.

Table 44. Key Findings Summary from Technical Report 1316

Category	Summary
Grade Level	Grade 2
Assessment Alignment	Common Core State Standards (CCSS)
Assessment Platform	easyCBM® online system
Statistical Model	1PL Rasch model
Item Bank Quality	Most items demonstrated acceptable fit and effective distractors
Forms Developed	Multiple equivalent forms for progress monitoring and benchmarking
Primary Outcome	Reliable, growth-sensitive Grade 2 mathematics measures

Reference

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., & Tindal, G. (2013). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 2 (Technical Report 1316)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1317: The Development and Scaling of the easyCBM® CCSS Elementary Mathematics Measures: Grade 3 (Saven, Irvin, et al., 2013a).

This technical report describes the development and scaling of the easyCBM® Common Core State Standards (CCSS) **elementary mathematics measures for Grade 3.** The primary goal of the study was to create technically sound, CCSS-aligned benchmark and progress monitoring assessments suitable for use within a Response to Intervention (RTI) framework. Emphasis was placed on alignment to instructional standards, accessibility for diverse learners, and psychometric rigor.

Methods

Subjects included a large national sample of Grade 3 students recruited through existing easyCBM® and Behavioral Research and Teaching (BRT) partnerships. Participation was voluntary and spanned 33 states, involving 1,685 Grade 3 students taught by 329 teachers across 140 schools and 132 districts. No individual student demographic data were collected to preserve confidentiality. Assessments were administered online in classrooms under teacher supervision near the end of the 2012–2013 academic year.

Data collection followed a structured piloting plan designed to support stable item calibration. Fifteen pilot forms were created for Grade 3, each containing unique items and horizontally anchored items linking adjacent forms. Anchor items were drawn from both newly written CCSS-aligned items and previously developed easyCBM® items aligned to NCTM standards but validated as CCSS-consistent. Students were automatically assigned forms through the secure piloting platform to balance form completion rates, and response options were randomized to reduce cheating and order effects.

Statistical analyses were conducted using a one-parameter logistic (1PL) Rasch model implemented in Winsteps. This concurrent equating design allowed all Grade 3 items, including anchor items, to be placed on a common horizontal measurement scale. Item difficulty estimates, standard errors, mean square outfit statistics, and post-hoc discrimination indicators were examined to evaluate item functioning. Items with poor model fit, defined by mean square outfit values outside the recommended range of 0.5 to 1.5, were excluded from operational forms.

Following calibration, the refined item bank was used to construct 13 alternate Grade 3 test forms: three seasonal benchmark forms and ten progress monitoring forms. Test characteristic curves and average item difficulty estimates were used to evaluate form equivalence. Results indicated that the forms were highly comparable in overall difficulty, with mean difficulty values clustered closely around the scale mean. The item bank also demonstrated a broad range of difficulty, supporting accurate measurement across students with varying levels of mathematical proficiency.

Results

Overall findings support the technical adequacy of the Grade 3 easyCBM® CCSS mathematics measures. The assessments demonstrated strong alignment with CCSS domains, acceptable Rasch model fit, effective distractor functioning, and alternate-form equivalence. These results indicate that the Grade 3 measures are reliable, growth-sensitive tools capable of informing instructional decision-making and intervention planning within an RTI framework.

Table 45. Example Results from Technical Report 1317

Table 4

Third Grade Item Difficulties by Test Form (without adjacent grade vertical anchor items)

Item #	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	BMF	BMW	BMS	Mean
1	-0.56	-0.99	-0.67	-0.94	-0.91	-0.89	-0.72	-0.95	-0.58	-1.00	-1.15	-1.10	-1.08	-0.888
2	0.82	0.78	0.75	0.67	0.69	0.73	0.61	0.63	0.63	0.66	0.86	0.86	0.88	0.736
3	1.36	1.53	1.66	1.61	1.31	1.62	1.30	1.40	1.28	1.69	1.44	1.42	1.40	1.463
4	-0.83	-0.68	-0.38	-0.29	-0.15	-0.07	-0.24	-0.28	-0.53	-0.59	-0.95	-0.94	-0.98	-0.532
5	0.60	0.60	0.61	0.62	0.66	0.73	0.76	0.79	0.91	0.88	0.81	0.82	0.83	0.740
6	1.51	1.55	1.58	1.65	1.70	1.70	1.85	1.89	1.89	1.90	1.47	1.47	1.47	1.664
7	-1.74	-1.69	-1.74	-1.68	-1.74	-1.76	-1.77	-1.77	-1.79	-1.79	-1.71	-1.71	-1.71	-1.738
8	-1.28	-1.26	-1.29	-1.24	-1.29	-1.23	-1.21	-1.20	-1.18	-1.17	-1.27	-1.27	-1.27	-1.243
9	-0.70	-0.72	-0.76	-0.78	-0.78	-0.79	-0.79	-0.80	-0.82	-0.82	-0.69	-0.69	-0.69	-0.756
10	-0.35	-0.34	-0.34	-0.33	-0.32	-0.33	-0.31	-0.31	-0.30	-0.29	-0.38	-0.38	-0.36	-0.334
11	0.36	0.43	0.34	0.36	0.38	0.34	0.46	0.46	0.45	0.45	0.32	0.32	0.31	0.383
12	0.85	0.88	0.91	0.89	0.96	0.89	0.97	0.91	0.92	0.99	0.86	0.86	0.86	0.904
13	1.38	1.39	1.40	1.40	1.42	1.44	1.47	1.49	1.50	1.52	1.54	1.53	1.53	1.462
14	-1.54	-1.53	-1.50	-1.50	-1.47	-1.42	-1.41	-1.41	-1.39	-1.47	-1.44	-1.44	-1.45	-1.459
15	-1.19	-1.17	-1.14	-1.12	-1.07	-1.07	-1.15	-1.09	-1.14	-1.10	-1.18	-1.18	-1.18	-1.137
16	-0.81	-0.83	-0.75	-0.83	-0.78	-0.77	-0.87	-0.89	-0.86	-0.72	-0.73	-0.74	-0.74	-0.794
17	-0.22	-0.24	-0.26	-0.30	-0.31	-0.31	-0.26	-0.33	-0.33	-0.34	-0.21	-0.21	-0.21	-0.272
18	0.17	0.17	0.20	0.19	0.27	0.26	0.29	0.26	0.29	0.25	0.22	0.22	0.23	0.232
19	1.22	1.35	1.41	1.20	1.41	1.18	1.19	1.54	1.46	1.54	1.22	1.28	1.28	1.329
20	3.10	2.96	2.73	2.74	2.99	2.97	2.97	3.09	3.16	2.67	3.13	3.13	3.14	2.983
21	-1.21	-1.19	-1.17	-1.21	-1.22	-1.15	-1.13	-1.14	-1.13	-1.17	-1.24	-1.24	-1.23	-1.187
22	-0.89	-0.90	-0.90	-0.87	-0.84	-0.85	-0.84	-0.84	-0.83	-0.84	-0.89	-0.89	-0.89	-0.867
23	-0.55	-0.54	-0.53	-0.54	-0.53	-0.53	-0.52	-0.49	-0.49	-0.48	-0.47	-0.47	-0.46	-0.508
24	0.45	0.44	0.46	0.45	0.45	0.46	0.46	0.48	0.46	0.48	0.48	0.48	0.48	0.464
25	0.72	0.72	0.71	0.71	0.71	0.73	0.76	0.75	0.75	0.75	0.72	0.72	0.72	0.728
26	1.35	1.35	1.38	1.38	1.36	1.40	1.39	1.38	1.39	1.41	1.34	1.34	1.34	1.370
27	2.38	2.18	2.12	2.13	2.21	2.32	2.21	2.34	2.25	2.14	2.46	2.40	2.40	2.272
28	0.10	0.11	0.18	0.20	0.14	0.10	0.20	0.29	0.26	0.30	0.24	0.24	0.25	0.201
29	0.79	0.78	0.73	0.76	0.72	0.72	0.69	0.70	0.68	0.69	0.84	0.82	0.82	0.749
30	1.06	1.05	1.02	1.13	1.14	1.15	1.17	1.17	1.22	1.24	1.06	1.06	1.09	1.120
Mean	0.212	0.206	0.225	0.215	0.237	0.252	0.251	0.269	0.271	0.259	0.223	0.224	0.226	0.236

Note. PM1 to PM10 = Progress Monitoring Form 1 to Form 10; BMF = Benchmark Fall, BMW = Benchmark Winter, BMS = Benchmark Spring; Green = horizontal anchor items.

Table 46. Key Findings Summary from Technical Report 1317

Category	Summary
Grade Level	Grade 3
Participants	1,685 students across 33 states
Assessment Format	Online, multiple-choice
Statistical Model	1PL Rasch model (concurrent equating)
Forms Developed	3 benchmark forms; 10 progress monitoring forms
Item Quality	Most items demonstrated acceptable Rasch fit
Primary Outcome	Reliable, CCSS-aligned, growth-sensitive math measures

Reference

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., Anderson, D., & Tindal, G. (2013). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 3 (Technical Report No. 1317)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1318: The Development and Scaling of the easyCBM® CCSS Elementary Mathematics Measures: Grade 4 (Irvin, Saven, et al., 2013b)**.**

Methods

This technical report documents the development and scaling of the easyCBM® Common Core State Standards (CCSS) **elementary mathematics measures for Grade 4.** The study was designed to create progress monitoring and benchmarking assessments that are aligned with CCSS expectations, sensitive to student growth, and psychometrically sound for use within a Response to Intervention (RTI) framework.

Participants consisted of a large national sample of Grade 4 students drawn from schools across the United States. Schools and teachers volunteered to participate through existing easyCBM® partnerships and outreach efforts conducted by Behavioral Research and Teaching at the University of Oregon. To ensure confidentiality, no identifying student, teacher, school, or district information was collected. Assessments were administered online under teacher supervision during designated piloting windows and followed standardized procedures reflective of typical classroom assessment conditions.

Item development emphasized alignment with Grade 4 CCSS mathematics domains, including Operations and Algebraic Thinking, Number and Operations in Base Ten, Number and Operations—Fractions, Measurement and Data, and Geometry. Items were written by experienced elementary educators and assessment specialists who received formal training in effective item construction and Universal Design for Assessment principles. Item writers focused on minimizing linguistic complexity, reducing construct-irrelevant cognitive demands, and ensuring each item targeted a single mathematical concept. Items emphasized conceptual understanding and problem solving rather than procedural fluency alone.

Data collection involved large-scale piloting of Grade 4 items across participating schools. Student response data were analyzed using a one-parameter logistic (1PL) Rasch model. Item calibration placed all Grade 4 items on a common measurement scale, allowing for consistent interpretation of item difficulty and student performance. Statistical analyses focused on item difficulty estimates, standard errors, mean square outfit statistics, and distractor functioning. Items that failed to meet model fit criteria or demonstrated weak distractor patterns were reviewed and either revised or excluded from operational forms.

Following calibration, the refined item bank was used to assemble multiple alternate Grade 4 assessment forms for both progress monitoring and benchmarking purposes. Each form was constructed to have comparable mean difficulty to support alternate-form equivalence. Form equivalence was evaluated using test characteristic curves (TCCs) and test information functions (TIFs), which showed strong overlap across forms and consistent measurement precision across the ability continuum.

Results

Results indicated that the majority of Grade 4 items demonstrated acceptable fit to the Rasch model and effective distractor functioning. The calibrated item bank covered a broad range of difficulty levels, enabling measurement of students with varying levels of mathematical proficiency. The alternate forms were shown to be psychometrically comparable and sensitive to changes in student performance over time.

Overall, the findings support the reliability, validity, and instructional utility of the easyCBM® CCSS Grade 4 mathematics measures. These assessments provide educators with robust tools for monitoring student progress, evaluating intervention effectiveness, and informing instructional decision-making within an RTI framework.

Table 47. Illustrative Results from Technical Report 1318

Table 4
Fourth Grade Item Difficulties by Test Form (without adjacent grade vertical anchor items)

Item #	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	BMF	BMW	BMS	Mean
1	-0.48	-0.39	-0.36	-0.35	-0.33	-0.32	-0.32	-0.31	-0.22	-0.20	-0.41	-0.41	-0.41	-0.347
2	0.96	0.93	0.91	0.87	0.89	0.90	0.77	0.77	0.82	0.85	0.98	1.01	1.04	0.900
3	1.63	1.69	1.56	1.75	1.69	1.81	1.66	1.88	1.66	1.88	1.55	1.53	1.53	1.678
4	0.39	0.41	0.28	0.33	0.49	0.49	0.47	0.37	0.44	0.43	0.45	0.45	0.45	0.419
5	0.59	0.64	0.64	0.64	0.65	0.66	0.66	0.66	0.69	0.69	0.67	0.67	0.67	0.656
6	1.20	1.29	1.29	1.31	1.00	1.00	1.11	1.34	1.34	1.37	1.32	1.32	1.32	1.247
7	-0.78	-0.74	-0.78	-0.70	-0.78	-0.79	-0.82	-0.81	-0.86	-0.86	-0.76	-0.76	-0.76	-0.785
8	-0.39	-0.38	-0.45	-0.37	-0.40	-0.37	-0.36	-0.35	-0.35	-0.34	-0.43	-0.43	-0.42	-0.388
9	-0.24	-0.22	-0.22	-0.23	-0.25	-0.26	-0.28	-0.27	-0.28	-0.28	-0.24	-0.24	-0.24	-0.250
10	0.13	0.13	0.09	0.11	0.12	0.15	0.15	0.16	0.16	0.16	0.14	0.14	0.14	0.137
11	0.46	0.42	0.44	0.46	0.47	0.45	0.51	0.52	0.42	0.44	0.49	0.49	0.49	0.466
12	0.71	0.73	0.76	0.76	0.79	0.76	0.77	0.79	0.78	0.80	0.72	0.72	0.72	0.755
13	0.98	1.00	1.01	0.90	0.88	0.91	0.91	0.94	0.95	0.98	0.88	0.88	0.90	0.932
14	1.29	1.30	1.30	1.31	1.31	1.32	1.34	1.34	1.36	1.32	1.35	1.35	1.35	1.326
15	1.51	1.52	1.57	1.59	1.61	1.61	1.54	1.60	1.59	1.60	1.55	1.55	1.55	1.568
16	2.05	2.02	2.08	2.03	2.06	2.06	1.97	1.97	2.00	2.13	2.10	2.10	2.10	2.052
17	2.58	2.57	2.56	2.46	2.59	2.68	2.48	2.40	2.59	2.39	2.42	2.42	2.40	2.503
18	-1.78	-1.78	-1.76	-1.77	-1.66	-1.70	-1.65	-1.68	-1.66	-1.75	-1.7	-1.71	-1.71	-1.716
19	-1.47	-1.43	-1.46	-1.48	-1.46	-1.49	-1.49	-1.41	-1.43	-1.41	-1.42	-1.42	-1.42	-1.445
20	-1.29	-1.28	-1.29	-1.29	-1.28	-1.24	-1.25	-1.22	-1.22	-1.26	-1.26	-1.26	-1.26	-1.262
21	-0.91	-0.90	-0.93	-0.91	-0.92	-0.92	-0.87	-0.89	-0.89	-0.93	-0.89	-0.89	-0.89	-0.903
22	-0.72	-0.77	-0.73	-0.71	-0.70	-0.70	-0.70	-0.69	-0.68	-0.69	-0.72	-0.72	-0.72	-0.712
23	-0.22	-0.22	-0.21	-0.20	-0.20	-0.20	-0.19	-0.19	-0.18	-0.17	-0.18	-0.18	-0.18	-0.194
24	0.12	0.10	0.13	0.10	0.12	0.15	0.13	0.16	0.14	0.15	0.16	0.17	0.17	0.138
25	0.53	0.54	0.52	0.52	0.52	0.56	0.59	0.59	0.56	0.56	0.55	0.55	0.55	0.549
26	0.86	0.86	0.85	0.85	0.83	0.89	0.87	0.86	0.89	0.89	0.86	0.86	0.86	0.864
27	1.76	1.78	1.73	1.74	1.78	1.87	1.79	1.89	1.87	1.86	1.80	1.80	1.80	1.805
28	-0.32	-0.29	-0.15	-0.10	-0.21	-0.32	-0.08	-0.05	-0.06	-0.04	-0.04	0.04	0.04	-0.122
29	0.51	0.47	0.53	0.54	0.52	0.40	0.36	0.38	0.35	0.35	0.42	0.41	0.41	0.435
30	1.19	1.18	1.18	1.30	1.32	1.44	1.46	1.21	1.23	1.29	1.51	1.51	1.52	1.334
Mean	0.362	0.373	0.370	0.382	0.382	0.393	0.384	0.399	0.400	0.407	0.396	0.398	0.400	0.388

Note. PM1 to PM10 = Progress Monitoring Form 1 to Form 10; BMF = Benchmark Fall, BMW = Benchmark Winter, BMS = Benchmark Spring; Green = horizontal anchor items.

Table 48. Summary of Key Findings from Technical Report 1318

Category	Summary
Grade Level	Grade 4
Assessment Alignment	Common Core State Standards (CCSS)
Assessment Platform	easyCBM® online system
Statistical Model	1PL Rasch model
Item Bank Quality	Most items showed acceptable fit and strong distractor functioning
Forms Developed	Multiple equivalent forms for progress monitoring and benchmarking
Primary Outcome	Reliable, growth-sensitive Grade 4 mathematics measures

Reference

Irvin, P. S., Alonzo, J., & Tindal, G. (2013). *The development and scaling of the easyCBM CCSS elementary mathematics measures: Grade 4 (Technical Report No. 1318).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1319: The Development and Scaling of the easyCBM® Common Core State Standards Elementary Mathematics Measures: Grade 5 (Saven, Irvin, et al., 2013b).

This technical report documents the development, piloting, and scaling of the easyCBM® Common Core State Standards (CCSS) **Grade 5 mathematics measures** designed for use within a Response to Intervention (RTI) framework. The primary objective was to produce technically adequate benchmark and progress-monitoring assessments aligned with CCSS and suitable for diverse student populations.

Methods

Participants included 1,525 Grade 5 students taught by 329 teachers across 140 schools in 132 districts spanning 33 U.S. states. Students participated during the spring 2013 pilot window using an online assessment platform. Demographic information was not collected to protect confidentiality. Students were randomly assigned one of 15 pilot forms, each consisting of 41 items, and responses were automatically recorded. Answer choices were randomized to minimize cheating, and calculators were not permitted. A total of 500 Grade 5 CCSS-aligned math items were developed as part of a larger K–5 item pool. Item writers and reviewers averaged 14 years of mathematics teaching experience and participated in structured training focused on CCSS alignment, item-writing principles, and Universal Design for Assessment. Items underwent three stages of review: contracted expert review, internal university review, and external review. Graphics and audio supports were developed to enhance accessibility. Only items meeting standards for clarity, alignment, and lack of bias advanced to piloting.

All items were calibrated using a one-parameter logistic (1PL) Rasch model with concurrent equating implemented in WINSTEPS (version 3.6.8). Horizontal anchor items from both newly developed CCSS items and previously validated NCTM-based items linked pilot forms to a common scale. Item difficulty estimates (β) and outfit mean square (MNSQ) statistics were examined. Items with MNSQ values outside the acceptable range of 0.50–1.50 were removed from consideration. Distractor analyses evaluated whether incorrect options functioned as intended across ability levels.

Results

Results indicated that most Grade 5 items demonstrated acceptable Rasch model fit and effective distractor functioning. Thirteen operational test forms were constructed per grade: three benchmark forms (fall, winter, spring) and ten progress-monitoring forms. Average item difficulty across Grade 5 forms was approximately 0.25 logits, with minimal variation, indicating strong form equivalence. Benchmark forms included vertical anchor items linking adjacent grades to support future vertical scaling. Overall, findings support the technical adequacy, CCSS alignment, and growth sensitivity of the easyCBM® Grade 5 mathematics measures for RTI applications.

Table 49. Example Results from Technical Report 1319

Table 6

Grade 5 Pearson Split-test Correlation (PC) and Reliability of Slope (RS) Analyses Results

Measure	*n*	Analytic Approach	Correlation coefficient (*r*)	95% Confidence Interval	
				Lower	Upper
CCSS Math	19	PC	.31	-.17	.67
	19	RS	.42	.00	1.00
Numbers and Operations (NumOp)	69	PC	.23	-.01	.44
	69	RS	.29	.07	.58
Geometry Measurement and Algebra (GeoMeasAlg)	6	PC	.44	-.58	.92
	6	RS	.84	.23	1.00
Numbers Operations and Algebra (NumOpAlg)	6	PC	.28	-.69	.89
	6	RS	.46	.00	1.00

Note. **Lower bound of the confidence interval around the median correlation estimate falls below 0.50 but meets or exceeds 0.40.

Table 50. Key Findings Summary from Technical Report 1319

Category	Summary
Participants	1,525 Grade 5 students from 33 states
Assessment Design	15 pilot forms; online administration
Analysis Method	1PL Rasch model with concurrent equating
Item Pool	500 CCSS-aligned Grade 5 items
Item Fit	Majority within acceptable MNSQ range (0.50–1.50)
Final Output	13 equivalent benchmark and progress-monitoring forms

Reference

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2013b). The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 5 (Technical Report # 1319). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1207: The Development and Scaling of the easyCBM® CCSS Middle School Mathematics Measures (Anderson et al., 2012).

This technical report describes the development and scaling of the easyCBM® Common Core State Standards (CCSS) **middle school mathematics measures designed for use in grades 6–8** within a Response to Intervention (RTI) framework. The primary objective was to create progress monitoring and benchmarking assessments that measure higher-order mathematical reasoning, are sensitive to growth over time, and are psychometrically comparable across grades through vertical scaling.

Methods

Participants included a large national sample of middle school students in grades 6, 7, and 8 who participated in item piloting during the 2011–2012 school year. Schools and teachers volunteered through district partnerships and prior involvement with easyCBM®. To protect confidentiality, no identifying student or school information was reported. Students completed the assessments online under standard testing conditions, consistent with typical classroom administration procedures.

Item development emphasized alignment with the Common Core State Standards for Mathematics and accessibility for diverse student populations. A total of 2,700 items were written, with 900 items developed for each grade. Items were stratified across the five CCSS mathematics domains for grades 6–8 and evenly distributed across standards. Item writing was conducted by experienced middle school mathematics teachers who received formal training in effective item construction and principles of Universal Design for Assessment. Items were designed to minimize construct-irrelevant barriers while targeting conceptual understanding, problem solving, and application rather than fluency alone.

Data collection involved large-scale piloting of items across participating schools. Item responses were analyzed using a one-parameter logistic (1PL) Rasch model. All items were calibrated to a single vertical scale spanning grades 6 through 8, enabling both within-grade and across-grade comparisons of student performance. Item difficulty estimates and fit statistics were examined to evaluate item functioning. Although discrimination parameters were fixed in the Rasch model, post-hoc discrimination indices were reviewed to support interpretation of item quality.

Additional analyses included detailed distractor functioning examinations to ensure that correct response options were most frequently selected by higher-ability students, while distractors were more attractive to lower-ability students. Items that failed to meet model fit or distractor functioning expectations were revised or excluded.

Using the vertically scaled item bank, researchers assembled 13 alternate test forms per grade. Of these, 10 forms were designated for progress monitoring and 3 for seasonal benchmarking. Form equivalence was evaluated using test characteristic curves (TCCs) and test information functions (TIFs), which demonstrated strong overlap across forms within each grade. These results indicate that the alternate forms provide comparable measurement precision across the ability continuum.

Results

Overall findings indicate that the easyCBM® CCSS middle school mathematics measures exhibit strong psychometric properties, adequate item fit, and reliable form equivalence. The vertical scale enables meaningful interpretation of student growth both within and across grades, addressing a significant gap in middle school curriculum-based measurement. The results support the use of these measures for instructional decision-making, progress monitoring, and benchmarking within RTI frameworks.

Table 51. Key Guidelines for Anchor Item Selection from Technical Report 1207

Table 3

Guidelines for Choosing Anchor Items

Guideline #	Guideline
1	Generally requires more than one step to solve, but not too many (i.e. 4+), or only one step if operation is difficult
2	Minimal language requirements outside those needed to solve problem.
3	Free of cultural bias, subjects of questions balanced (e.g., sports questions are not overly represented)
4	Targets standard (and specific substandard) directly, far below skills (i.e. simple addition) that are requisites of the targeted skill in the standard or OK
5	Consistent with universal design test features (simplicity, perceptibility, intuitiveness)

Table 52. Example Item Difficulties by Form Summary from Technical Report 1207

Table 4

Sixth Grade Item Difficulties by Form

Item#	Form													
	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	BMF	BMW	BMS	Mean
1	-1.00	-0.89	-0.89	-0.84	-0.81	-0.81	-0.79	-0.76	-0.74	-0.72	-0.71	-0.71	-0.71	-0.7985
2	-0.44	-0.48	-0.44	-0.56	-0.57	-0.58	-0.59	-0.59	-0.61	-0.61	-0.62	-0.62	-0.63	-0.5646
3	-0.22	-0.31	-0.44	-0.26	-0.27	-0.26	-0.30	-0.32	-0.34	-0.43	-0.38	-0.37	-0.36	-0.3277
4	-0.09	-0.03	0.08	-0.06	-0.09	-0.12	-0.04	-0.05	-0.02	0.07	0.05	0.05	0.01	-0.0185
5	0.49	0.21	0.20	0.39	0.40	0.51	0.35	0.30	0.29	0.19	0.13	0.12	0.16	0.2877
6	-1.11	-0.55	-0.57	-0.99	-0.76	-1.10	-0.74	-0.71	-0.62	-0.53	-0.49	-0.43	-0.41	-0.6931
7	0.15	-0.08	-0.08	0.09	-0.05	0.09	-0.04	0.03	-0.09	-0.13	-0.17	-0.21	-0.24	-0.0562
8	0.34	0.26	0.26	0.34	0.27	0.35	0.27	0.24	0.25	0.28	0.33	0.30	0.31	0.2923
9	0.56	0.43	0.55	0.56	0.52	0.56	0.43	0.40	0.44	0.45	0.47	0.49	0.49	0.4885
10	0.72	1.01	0.73	0.73	0.75	0.77	1.05	1.11	1.01	0.96	0.90	0.89	0.83	0.8815
11	-1.17	-1.35	-1.17	-1.13	-1.12	-1.23	-1.47	-1.49	-1.31	-1.30	-1.28	-1.27	-1.25	-1.2723
12	-0.85	-0.81	-0.80	-0.86	-0.86	-0.72	-0.69	-0.68	-0.86	-0.80	-0.77	-0.86	-0.76	-0.7938
13	-0.45	-0.55	-0.59	-0.55	-0.54	-0.62	-0.62	-0.64	-0.46	-0.55	-0.62	-0.45	-0.62	-0.5585
14	-0.23	-0.18	-0.17	-0.18	-0.20	-0.15	-0.11	-0.15	-0.24	-0.21	-0.11	-0.24	-0.11	-0.1754
15	0.14	0.32	0.41	0.32	0.38	0.26	0.12	0.18	0.14	0.33	0.10	0.17	0.11	0.2292
16	-1.53	-1.88	-1.95	-1.92	-1.92	-1.72	-1.44	-1.71	-1.49	-1.84	-1.45	-1.56	-1.41	-1.6785
17	-1.23	-0.97	-1.07	-0.96	-1.04	-1.18	-1.38	-1.17	-1.29	-1.13	-1.35	-1.20	-1.38	-1.1808
18	-0.88	-0.95	-0.84	-0.94	-0.89	-0.79	-0.60	-0.56	-0.82	-0.67	-0.73	-0.92	-0.76	-0.7962
19	-0.31	-0.26	-0.34	-0.20	-0.20	-0.39	-0.48	-0.49	-0.35	-0.44	-0.41	-0.24	-0.39	-0.3462
20	0.01	-0.08	0.16	-0.08	0.08	0.17	-0.07	-0.03	0.07	-0.01	0.04	-0.01	0.07	0.0246
21	-0.85	-0.79	-1.24	-0.85	-1.34	-1.27	-0.73	-0.85	-1.21	-0.86	-0.96	-1.08	-1.09	-1.0092
22	-0.64	-0.67	-0.45	-0.61	-0.31	-0.43	-0.69	-0.66	-0.32	-0.61	-0.51	-0.48	-0.50	-0.5292
23	-0.24	-0.23	-0.15	-0.29	-0.26	-0.15	-0.30	-0.23	-0.28	-0.30	-0.31	-0.22	-0.16	-0.2400
24	-0.11	-0.09	-0.14	0.02	-0.07	-0.13	0.08	-0.07	0.01	0.05	0.00	-0.02	-0.13	-0.0462
25	0.30	0.24	0.30	0.15	0.23	0.26	0.11	0.22	0.16	0.14	0.21	0.21	0.25	0.2138
Mean	-0.3456	-0.3472	-0.3456	-0.3472	-0.3468	-0.3472	-0.3468	-0.3472	-0.3472	-0.3468	-0.3456	-0.3464	-0.3472	

Note. f1 to f10 = Form 1 to Form 10; BMF = Benchmark Fall, BMW = Benchmark Winter, BMS = Benchmark Spring

Table 5
Sixth Grade Item Residual Analysis

Item#	Form												
	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	BMF	BMW	BMS
1	-0.20	-0.09	-0.09	-0.04	-0.01	-0.01	0.01	0.04	0.06	0.08	0.09	0.09	0.09
2	0.12	0.08	0.12	0.00	-0.01	-0.02	-0.03	-0.03	-0.05	-0.05	-0.06	-0.06	-0.07
3	0.11	0.02	-0.11	0.07	0.06	0.07	0.03	0.01	-0.01	-0.10	-0.05	-0.04	-0.03
4	-0.07	-0.01	0.10	-0.04	-0.07	-0.10	-0.02	-0.03	0.00	0.09	0.07	0.07	0.03
5	0.20	-0.08	-0.09	0.10	0.11	0.22	0.06	0.01	0.00	-0.10	-0.16	-0.17	-0.13
6	-0.42	0.14	0.12	-0.30	-0.07	-0.41	-0.05	-0.02	0.07	0.16	0.20	0.26	0.28
7	0.21	-0.02	-0.02	0.15	0.01	0.15	0.02	0.09	-0.03	-0.07	-0.11	-0.15	-0.18
8	0.05	-0.03	-0.03	0.05	-0.02	0.06	-0.02	-0.05	-0.04	-0.01	0.04	0.01	0.02
9	0.07	-0.06	0.06	0.07	0.03	0.07	-0.06	-0.09	-0.05	-0.04	-0.02	0.00	0.00
10	-0.16	0.13	-0.15	-0.15	-0.13	-0.11	0.17	0.23	0.13	0.08	0.02	0.01	-0.05
11	0.10	-0.08	0.10	0.14	0.15	0.04	-0.20	-0.22	-0.04	-0.03	-0.01	0.00	0.02
12	-0.06	-0.02	-0.01	-0.07	-0.07	0.07	0.10	0.11	-0.07	-0.01	0.02	-0.07	0.03
13	0.11	0.01	-0.03	0.01	0.02	-0.06	-0.06	-0.08	0.10	0.01	-0.06	0.11	-0.06
14	-0.05	0.00	0.01	0.00	-0.02	0.03	0.07	0.03	-0.06	-0.03	0.07	-0.06	0.07
15	-0.09	0.09	0.18	0.09	0.15	0.03	-0.11	-0.05	-0.09	0.10	-0.13	-0.06	-0.12
16	0.15	-0.20	-0.27	-0.24	-0.24	-0.04	0.24	-0.03	0.19	-0.16	0.23	0.12	0.27
17	-0.05	0.21	0.11	0.22	0.14	0.00	-0.20	0.01	-0.11	0.05	-0.17	-0.02	-0.20
18	-0.08	-0.15	-0.04	-0.14	-0.09	0.01	0.20	0.24	-0.02	0.13	0.07	-0.12	0.04
19	0.04	0.09	0.01	0.15	0.15	-0.04	-0.13	-0.14	0.00	-0.09	-0.06	0.11	-0.04
20	-0.01	-0.10	0.14	-0.10	0.06	0.15	-0.09	-0.05	0.05	-0.03	0.02	-0.03	0.05
21	0.16	0.22	-0.23	0.16	-0.33	-0.26	0.28	0.16	-0.20	0.15	0.05	-0.07	-0.08
22	-0.11	-0.14	0.08	-0.08	0.22	0.10	-0.16	-0.13	0.21	-0.08	0.02	0.05	0.03
23	0.00	0.01	0.09	-0.05	-0.02	0.09	-0.06	0.01	-0.04	-0.06	-0.07	0.02	0.08
24	-0.06	-0.04	-0.09	0.07	-0.02	-0.08	0.13	-0.02	0.06	0.10	0.05	0.03	-0.08
25	0.09	0.03	0.09	-0.06	0.02	0.05	-0.10	0.01	-0.05	-0.07	0.00	0.00	0.04
Average	0.03	-0.01	0.03	-0.01	0.00	-0.01	0.00	-0.01	-0.01	0.00	0.03	0.01	-0.01

Figure 1. Example Graph of Relation: Time and Raw Score Performance from Technical Report 1207

Figure 1 - *Line plot of mean total raw score and time taken on test (first hour of test administration)*

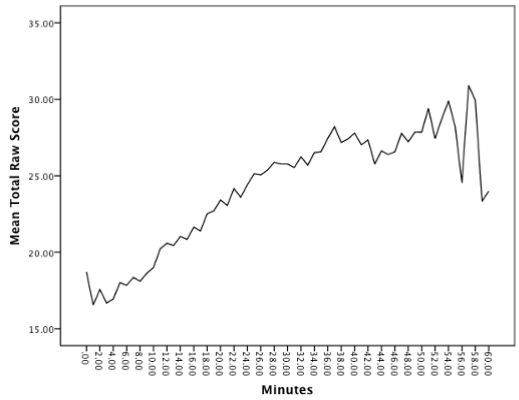


Figure 2. Example Test Information from Technical Report 1207

Figure 2

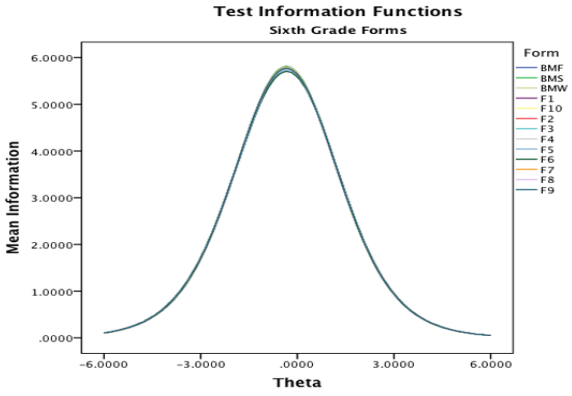


Figure 3. Example Test Characteristic Curves from Technical Report 1207

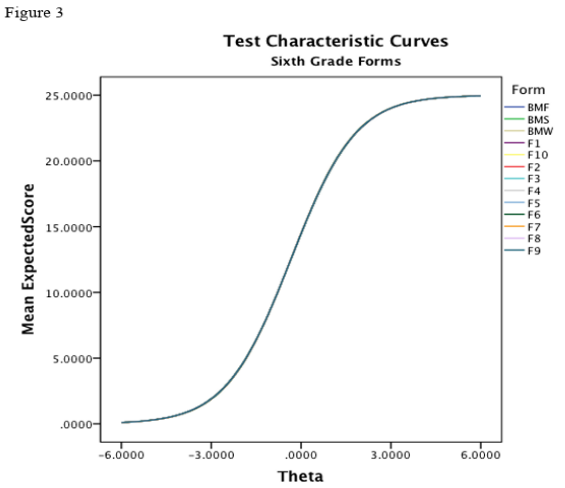


Table 53. Key Findings Summary from Technical Report 1207

Category	Summary
Grades Assessed	Grades 6–8
Items Developed	2,700 total items (900 per grade)
Assessment Alignment	Common Core State Standards (CCSS)
Statistical Model	1PL Rasch model with vertical scaling
Forms Created	13 forms per grade (10 progress monitoring, 3 benchmarking)
Form Equivalence	Supported by TCCs and TIFs
Primary Contribution	Vertically scaled measures enabling cross-grade growth analysis

Reference

Anderson, D., Irvin, P. S., Patarapichayatham, C., Alonzo, J., & Tindal, G. (2012). *The development and scaling of the easyCBM® CCSS middle school mathematics measures (Technical Report 1207)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1313: easyCBM® CCSS Math Item Scaling and Test Form Revision (2012–2013): Grades 6–8 (Anderson et al., 2013).

This technical report documents the piloting, scaling, and revision of easyCBM® Common Core State Standards (CCSS) **mathematics assessments for Grades 6–8.** The primary objectives were to calibrate newly developed CCSS-aligned items onto an existing vertical scale and to revise operational test forms to improve psychometric performance while maintaining alternate-form comparability.

Methods

Participants included students from five schools across five districts located in the Pacific Northwest and Southwest regions of the United States. The total sample comprised 729 Grade 6 students, 1,061 Grade 7 students, and 1,122 Grade 8 students. All participants were users of the district-level easyCBM® online assessment system. Participation was incentivized through district compensation and classroom-level stipends.

Data collection occurred during the Winter 2013 administration of the CCSS mathematics benchmark assessments. Students completed the operational benchmark form for their grade level, followed immediately by 25 pilot items

presented seamlessly as part of a single testing session. Pilot items were administered using a conditional randomization algorithm to ensure a minimum of 200 student responses item while preventing repeated exposure.

Statistical analyses were conducted using a one-parameter Rasch measurement model implemented in Winsteps software. A non-equivalent groups anchor test (NEAT) design was employed, anchoring item difficulties from previously calibrated benchmark items to the established vertical scale. Pilot item difficulties were freely estimated relative to the anchored parameters. Item fit was evaluated using outfit mean square statistics, with acceptable values defined between 0.8 and 1.2. Item discrimination was assessed via point-measure correlations, with a minimum criterion of 0.20 for inclusion.

Results

Results indicated that newly piloted items demonstrated difficulty distributions comparable to anchored items across all grades, supporting successful scale integration. Poorly discriminating items identified in prior reliability analyses were removed and replaced with higher-performing pilot items. Five NCTM-based items aligned with CCSS standards were added to each form to improve accessibility, particularly for lower-performing students. Benchmark forms incorporated additional common items to support both horizontal and future vertical scaling.

Overall, revised test forms exhibited highly comparable mean difficulties, narrow interquartile ranges, and minimal outliers, indicating strong alternate-form equivalence. These findings support the technical adequacy of the revised easyCBM® CCSS mathematics measures for use in progress monitoring and benchmark screening within RTI frameworks.

Figure 4. Sample Item Difficulty Distribution for Anchored Items from Technical Report 1313

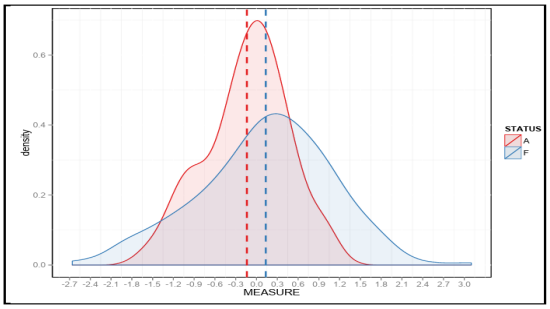


Figure 2. Distribution of Grade 6 item difficulties for anchored and freely estimated items.

Figure 5. Illustrative Box Plots of Item Difficulty from Technical Report 1313

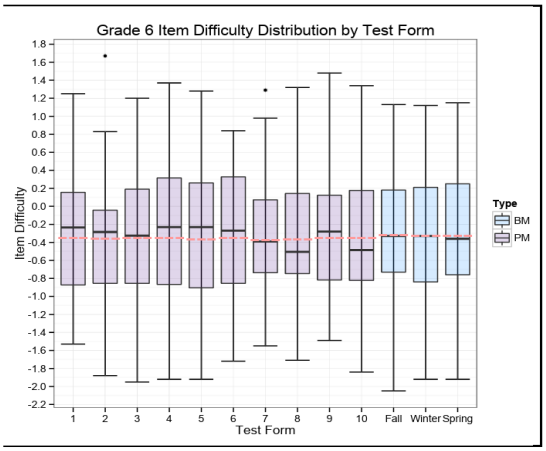


Figure 5. Distribution of item difficulties within Grade 6 test forms. Note that the solid black line within each boxplot represents the median item difficulty for the respective test form, while the hatched red line represents the mean.

Table 54. Key Findings Summary from Technical Report 1313

Category	Summary
Participants	Over 2,900 students across Grades 6–8 from five districts
Item Piloting	25 pilot items administered per student using conditional randomization
Statistical Model	Rasch 1PL model with anchored equating (NEAT design)
Item Fit Criteria	Outfit MNSQ between 0.8–1.2; point-measure ≥ 0.20
Form Revisions	Low-discrimination items replaced; NCTM items added
Overall Outcome	Improved reliability and maintained form comparability

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2012). *easyCBM® CCSS math item scaling and test form revision (2012–2013): Grades 6–8 (Technical Report 1313).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1408: Technical manual: easyCBM® (Anderson et al., 2014).

Methods

Measures. Two sets of mathematics measures were examined: (1) **NCTM Math Measures** (Grades K–8), aligned to the National Council of Teachers of Mathematics Focal Point Standards, consisting of three seasonal benchmarks (originally 48 items, refined to 45) and 30 progress monitoring forms (16 items each); and (2) **CCSS Math Measures** (Grades K–8), aligned to the Common Core State Standards, developed in two phases (middle school 2011–2013; elementary 2012–2013) with 32–50 items per form depending on grade band.

Subjects. The easyCBM® mathematics technical evidence was gathered from large national samples spanning Grades K–8. Criterion validity samples included approximately 2,400–4,400 students per grade level drawn from multiple districts in Oregon and Washington state, as well as a national convenience sample of 76 schools across 26 states for Grades K–2. Norming data were based on a stratified random sample of 500 students per demographic cell, designed to reflect national enrollment proportions by region, race-ethnicity, and gender using the Common Core of Data from the National Center for Education Statistics.

Data Collection Procedures. Item development followed structured multi-stage processes involving trained teacher item-writers, expert review panels, and piloting with approximately 2,800 students per grade (NCTM) or national convenience samples (CCSS). Alignment studies employed teacher raters who independently judged item-standard correspondence using Webb's alignment model or 4-point Likert scales; ratings were analyzed using the many-facets Rasch model (MFRM) to control for rater severity.

Statistical Analyses. Rasch modeling was used to calibrate items, construct equivalent test forms, and evaluate item fit (Mean Square Outfit; acceptable range 0.50–1.50). Internal consistency was assessed via Cronbach's alpha and split-half (Spearman-Brown) reliability. Criterion validity was examined through simple and multiple linear regression, Pearson/Spearman correlations, and diagnostic efficiency statistics (sensitivity, specificity, area under the ROC curve [AUC]). Construct validity was evaluated using confirmatory factor analysis (CFA) comparing one-factor and three-factor models, and bivariate correlations with state assessments (Oregon OAKS, Washington MSP, TerraNova 3, SAT-10).

Results

Both the NCTM and CCSS Math measures were developed through rigorous, iterative processes grounded in Rasch modeling to ensure test form equivalence. For the NCTM measures, Rasch analyses of approximately 1,100 piloted items per grade guided the removal of misfitting items and the construction of forms with equivalent average difficulty and adequate range from easy to difficult items. Items were deemed poorly fitting and removed if they

were overfit (Mean Square Outfit < 0.49) or underfit (> 1.51) and distractor analysis did not support retention. The result was three seasonal benchmarks and 30 progress monitoring forms per grade for Grades K–8.

Alignment of NCTM items to NCTM Focal Point Standards was generally strong. Across grades, benchmark items were rated by 13 trained teacher raters, with the majority of items linked to target standards. Results ranged by grade and focal point but were broadly acceptable, with some grades achieving alignment rates above 95%. A subsequent alignment study comparing NCTM items to the CCSS found reasonable but imperfect alignment: benchmark items tended to align more strongly at the domain level than at the individual standard level, and more strongly to on-grade than prior-grade CCSS. These gaps informed targeted new item development for 2012–2013.

For the CCSS Math measures, the alignment study for middle school grades (6–8) found that 87.73% of the 1,345 reviewed items had adjusted MFRM ratings at or above 2.0 (aligned). Of the remaining 12.27%, fully 97.00% were rated as targeting a requisite skill to the standard. Combined, 99.6% of sampled items were judged aligned with a grade-level CCSS or a requisite skill, representing strong content alignment. Rater consistency was excellent, with mean square outfit statistics for all 15 raters ranging from 0.76 to 1.16.

The NCTM Math measures demonstrated strong internal consistency. Cronbach's alpha ranged from .78 to .91 across all grades (K–8) and benchmark seasons (fall, winter, spring), meeting or exceeding the generally accepted threshold of .80 for most grades and seasons. Split-half reliability coefficients (Spearman-Brown) ranged from .71 to .89 with a median of .82, also indicating acceptable to strong reliability. The CCSS Math measures showed similarly strong or even higher internal consistency, with Cronbach's alpha ≥ .80 across all grades and testing occasions (K–8, fall and winter benchmarks), with alpha values as high as .95 for Grades 6–8. Split-half correlations ranged from .52 to .73 at the lower grades, increasing substantially in the upper grades. An initial reliability concern was identified for CCSS middle school forms prior to revision (alpha < .70); form revisions resolved this issue.

Criterion validity evidence for the NCTM measures was extensive and consistently strong across eight studies for Grades 3–8, with additional studies for K–2. Predictive validity studies compared fall and winter benchmarks to spring administrations of the Oregon OAKS and Washington MSP. For Grades 3–8, fall and winter simple linear regression models accounted for 58–73% of the variance in OAKS scores and 56–72% of the variance in MSP scores, with variance accounted for generally increasing with grade level. For Grades K–2, fall measures predicted 39–54% of variance in the TerraNova 3.

Diagnostic efficiency statistics were robust. AUC statistics for predictive studies ranged from .83 to .94 across state tests and grade levels, indicating excellent discrimination between students who would and would not meet proficiency. Sensitivity of optimal cut scores ranged from .73 to .94 and specificity from .65 to .88. Cross-validation studies confirmed the stability of these cut scores across randomly selected groups of approximately 2,000 students each, with 95% confidence intervals for AUC statistics overlapping across groups, providing strong evidence for cut score generalizability.

Concurrent validity was equally strong. Spring benchmark correlations with state tests ranged from .73 to .82 for OAKS and .68 to .81 for the MSP. Concurrent regression models accounted for 52–67% of the variance in OAKS and 48–67% in the MSP. For CCSS Math measures at Grades 6–8, bivariate correlations with the SAT-10 ranged from .75 to .82, with the winter benchmark accounting for 56–67% of variance in SAT-10 scores.

Construct validity analyses consistently supported a unidimensional mathematics factor for both the NCTM and CCSS math measures. For Grades K–2, Rasch item-fit values for the one-factor model ranged from .50 to 1.30 (Grades K–1) and .60 to 1.79 (Grade 2), indicating adequate model fit. CFA chi-square difference tests comparing one-factor and three-factor models found that three-factor models did not result in significantly better fit at any grade level. This finding held for both the NCTM K–2 and 3–8 analyses. Inter-factor correlations in the three-factor models were moderate to high (.70–.90 for K–2; .60–.80 for 3–8), further supporting the one-factor interpretation. Bivariate correlations between seasonal benchmarks and year-end state tests ranged from approximately .60 to .80, consistent with a single underlying mathematics construct.

Table 55. Summary of Results from Technical Report 1403

Domain	Measure	Metric	Result
Reliability	NCTM (K–8)	Cronbach's α	.78–.91
Reliability	NCTM (K–8)	Split-half	.71–.89 (Mdn = .82)
Reliability	CCSS (K–8)	Cronbach's α	≥.80 (up to .95)
Predictive Validity	NCTM (3–8)	R^2 vs. OAKS/MSP	.56–.73
Predictive Validity	NCTM (K–2)	R^2 vs. TerraNova	.39–.54
Diagnostic Efficiency	NCTM (3–8)	AUC	.82–.94
Diagnostic Efficiency	NCTM (3–8)	Sensitivity	.73–.94
Diagnostic Efficiency	NCTM (3–8)	Specificity	.65–.88
Concurrent Validity	NCTM (3–8)	r vs. OAKS/MSP	.68–.82
Concurrent Validity	CCSS (6–8)	r vs. SAT-10	.75–.82
Alignment	NCTM to NCTM Standards	% Items Linked	65–99% by grade/focal point
Alignment	CCSS Math (6–8)	% Items Aligned	99.6% (aligned or requisite skill)
Construct Validity	NCTM & CCSS (K–8)	CFA Model Fit	One-factor model best fit at all grades

Reference

Anderson, D., Alonzo, J., & Tindal, G. (Eds.). (2014). Technical manual: easyCBM® (Technical Report No. 1408). Behavioral Research and Teaching, University of Oregon.

Appendix A: Technical Report Table and Figure Titles

Technical Report References

Alonzo, J., Anderson, D., & Tindal, G. (2009). *IRT analysis of general outcome measures in Grades 1-8 (Technical Report # 0916)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Lai, C. F., & Tindal, G. (2009a). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 3 (Technical Report # 0902)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Lai, C. F., & Tindal, G. (2009b). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 4 (Technical Report # 0903)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Lai, C. F., & Tindal, G. (2009c). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and the general education populations: Grade 2 (Technical Report # 0920)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009a). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 1 (Technical Report # 0919)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009b). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Kindergarten (Technical Report # 0921)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2013). *easyCBM® CCSS math item scaling and test form revision (2012-2013): Grades 6-8 (Technical Report # 1313)*. 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., Irvin, P. S., Patarapichayatham, C., Alonzo, J., & Tindal, G. (2012). *The development and scaling of the easyCBM® CCSS middle school mathematics measures (Technical Report # 1207)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., Anderson, D., & Tindal, G. (2013). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade K (Technical Report # 1314)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., & Tindal, G. (2013a). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 2 (Technical Report # 1316)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., & Tindal, G. (2013b). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 4 (Technical Report # 1318)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Jung, E., Liu, K., Ketterlin-Geller, L. R., & Tindal, G. (2008). *Instrument development procedures for mathematics measures (Technical Report # 0802)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009a). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 5 (Technical Report # 0901)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009b). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 8 (Technical Report # 0904)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009c). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general populations: Grade 7 (Technical Report # 0908)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009d). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and the general education populations: Grade 6 (Technical Report # 0907)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Liu, K., Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2008). *Examining item functioning of math screening measures for grades 1-8 students (Technical Report # 0804)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Martinez, M., Ketterlin-Geller, L. R., & Tindal, G. (2007). *Content-related evidence for validity for mathematics tests: Teacher review (Technical Report # 42)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2013a). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 3 (Technical Report # 1317)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2013b). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 5 (Technical Report # 1319)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saven, J. L., Irvin, P. S., Park, B. J., Tindal, G., & Alonzo, J. (2013). *The development and scaling of the easyCBM® CCSS elementary mathematics measures: Grade 1 (Technical Report # 1315)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Conclusions Supporting Claims for Criterion 1.2: Item and Form Alignment

Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.

1.2.a Test forms/events delivered to students reflect an appropriate distribution of content and related score points and item types within forms/events.

Test forms focus strongly on major and supporting CCSS clusters; forms focus on content and skills for college and career readiness. Both the Basic Math (NCTM-aligned) and Proficient Math (CCSS-aligned) forms distribute items across the major mathematical domains appropriate to each grade. TR 1002 documents that 75–100% of benchmark and progress monitoring items across Grades K–8 link to NCTM Focal Points, with particularly strong coverage in Numbers and Operations, Geometry, Measurement, and Algebra. TR 2101 documents item distribution across all 11 CCSS domains in the Proficient Math measures (Table 9), with Geometry, Measurement and Data, Number and Operations in Base 10, and Operations and Algebraic Thinking most heavily represented across benchmark forms. Each seasonal benchmark form contains 45 items. The Proficient Math measures were developed explicitly with CCSS—the primary CCR framework for K–8 mathematics—as the foundational blueprint, ensuring forms reflect content and skills identified as necessary for college and career readiness.

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 identified as measuring one or more CCR standards; items align to design specifications; items are content-accurate. Items can be clearly identified as measuring one or more CCSS standards through structured expert review. TR 1208 used 15 raters and a four-point alignment scale to map each item to its intended CCSS; after Rasch adjustment, 87% of items were rated as directly aligned and 99.6% aligned when requisite prerequisite skills were included. TR 1228 (K–2), TR 1229 (Grades 3–5), and TR 1230 (Grades 6–8) used the Distributed Item Review (DIR) system, with four to five trained expert reviewers per grade independently assigning CCSS codes and alignment strength ratings. Items were developed by experienced mathematics teachers targeting specific CCSS domains, and content accuracy was evaluated through multi-rater review with consensus discussion resolving disagreements. No systematic content inaccuracies were reported across reports.

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

Items reach depth and complexity of CCR standards; there is an appropriate range of cognitive demand; item types and demand align to blueprints. TR 1002 explicitly evaluated cognitive demand using Webb's Depth of Knowledge (DOK) framework. Results indicated that most items function at DOK Level 1 (Recognition and Reproduction) and Level 2 (Skill and Concept), with few reaching DOK Level 3 (Strategic Thinking). This distribution is consistent with the CBM design rationale: benchmark and progress monitoring measures emphasize efficient, repeatable assessment of core mathematics skills rather than extended-response application. All items are multiple-choice, which constrains the range of item types but supports standardized administration. Alignment across CCSS clusters is documented in domain-level tables in TR 2101 (Table 10), covering clusters from foundational counting to ratios, expressions, functions, and statistical reasoning. Rater reliability for alignment was strong (ICCs .80–1.0), though DOK consensus was more moderate, reflecting the subjective nature of cognitive demand distinctions.

1.2.d The assessment is aligned to the procedural skill and fluency expectations of CCR standards.

Item development documentation and distributions of points directly address standards requiring procedural skill and fluency. Procedural skill and fluency are addressed primarily through items functioning at DOK Level 1 and lower Level 2, which characterize the majority of easyCBM® math items across grades (TR 1002). The multiple-choice, time-limited CBM format is well-suited to measuring procedural fluency in computation, number operations, and fact recall. TR 1228 and TR 1229 document strong alignment to CCSS clusters in Number and Operations in Base 10 and Operations and Algebraic Thinking—domains that include explicit procedural skill and fluency expectations at each grade. TR 1229 notes some overrepresentation of Number and Operations in Base Ten

in Grade 5. While the document does not formally disaggregate items by procedural versus conceptual categories, the DOK distribution and domain coverage reflect appropriate emphasis on procedural standards expected for K–8 mathematics screening.

1.2.e The assessment is aligned to the conceptual understanding expectations of CCR standards.

Item development documentation and distributions of points directly address standards requiring conceptual understanding. Conceptual understanding is addressed through DOK Level 2 (Skill and Concept) items, which constitute a significant portion of the item pool (TR 1002). Alignment to conceptually demanding CCSS clusters is documented across reports: TR 1228 shows coverage of fractions as numbers, algebraic thinking, and geometric relationships in K–2; TR 1229 confirms alignment to conceptual fraction, geometry, and proportional reasoning standards in Grades 3–5; TR 1230 documents coverage of Expressions and Equations, The Number System, and Ratios and Proportional Relationships in Grades 6–8. The DIR alignment process required raters to identify both direct and prerequisite skill alignments, supporting identification of items targeting conceptual foundations. The document does not formally separate items by CCSS category (procedural vs. conceptual), but the domain and DOK distributions suggest adequate representation of conceptually oriented standards across all grade bands.

1.2.f The assessment is aligned to the application expectations of CCR standards.

Item development documentation and distributions of points directly address standards requiring application; for high school assessments, items attend to the full intent of the modeling process. Application is addressed through higher-DOK items, though TR 1002 indicates few items reach DOK Level 3 (Strategic Thinking), which would most directly correspond to applied mathematical reasoning. The multiple-choice format limits the range of application tasks possible within the CBM design. Items in domains such as Ratios and Proportional Relationships, Geometry, and Statistics and Probability—documented in TR 2101 (Table 9) and TR 1230—address applied mathematical contexts appropriate to middle school grades. As a K–8 screening and progress monitoring system, easyCBM® is not designed as a high school assessment, and modeling standards for high school do not apply. Item development by experienced classroom teachers incorporated mathematical contexts intended to engage students in applied problem-solving within the constraints of the multiple-choice, benchmark administration format.

1.2.g The assessment includes mathematical practices as described in CCR standards.

Assessment design specifications addressing mathematical practices are reflected in items; items are connected to CCR mathematics; forms reflect distribution of mathematical practices. Mathematical practices (MPs) are not explicitly labeled or coded at the item level in the technical reports reviewed. The alignment studies (TR 1002, TR 1208, TR 1228, TR 1229, TR 1230) focus primarily on content standard alignment and DOK rather than practice standard alignment. However, the DOK framework used in TR 1002 captures aspects of mathematical reasoning (DOK Level 2–3) that overlap with practices such as reasoning abstractly, constructing arguments, and making use of structure. Items targeting CCSS content at Level 2–3 DOK inherently require students to engage in practice-adjacent behaviors. The document does not provide a distribution table of mathematical practices or explicit item-level practice alignments. This represents an area where additional documentation would strengthen the alignment evidence, particularly as CCSS-M places mathematical practices at the center of the standards framework.

1.2a – 1.2g Alignment for Math: Evaluation Indicators Alignment in Math

This section of the technical report provides information on two versions of the easyCBM® mathematics measures:
1. Basic math is aligned to National Council of Teachers of Mathematics in (NCTM).
2. Proficient math is aligned to the Common Core State Standards (CCSS).

Note that the Proficient Math is used in screening students with disabilities and Basic Math used in progress monitoring. The reason for this distinction is that, when CCSS standards were developed, they were designed to increase the rigor over those earlier promulgated with the original NCTM standards. This section presents data from six 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. Again, more specific information can be obtained from these technical reports with the findings presented in this summary considered exemplary.

Technical Report 1002 (2012) – The Alignment of easyCBM® Math Measures to Curriculum Standards.
Technical Report 2101 (2021) – The Alignment between easyCBM® Mathematics and Literacy Assessments and State and National Standards.
Technical Report 1228 (2012) – Alignment of easyCBM® Grades K–2 Math Measures to the Common Core Standards.
Technical Report 1229 (2012) – The Alignment of the easyCBM® Grades 3-5 Math Measures to the Common Core Standards.
Technical Report 1230 (2012) – The Alignment of the easyCBM® Grades 6-8 – Math Measures to the Common Core Standards.
Technical Report 1208 (2012) – The Alignment of the easyCBM® Middle School Mathematics CCSS Measures to the Common Core State Standards.

Each summary addresses the methods (subjects, settings, data collection, and analytic procedures) and then the results from these technical reports. Where present, actual tables are included; if the tables were too extensive (e.g., item response theory [IRT] presentations of difficulties and misfits), a representative example is presented. Finally, each section ends with an APA reference to the technical report.

Basic Math

Summary of Technical Report 1002: The Alignment of easyCBM® Math Measures to Curriculum Standards (Nese et al., 2010).

This technical report examined the alignment of easyCBM® mathematics benchmark and progress monitoring measures with the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Points for grades **Kindergarten through 8**. Using Webb's alignment framework, the study focused on evaluating content alignment and depth of knowledge (DOK) through expert judgment.

Methods emphasized structured expert review. Thirteen certified teachers with experience using easyCBM® served as raters. They were trained to evaluate item-to-standard alignment and DOK using a four-point alignment scale and a three-level DOK taxonomy. All benchmark forms for Grades K, 1, and 3–8 were reviewed, along with a substantial subset of progress monitoring forms. Each form contained 16 items, and each item was rated independently by at least two raters. Alignment ratings were later dichotomized to identify items clearly linked to standards. Reliability of ratings was evaluated using intraclass correlations (ICC) derived from cross-classified hierarchical linear models.

Results showed generally strong alignment across grades and focal points. For most grades, between 75% and 100% of benchmark and progress monitoring items were rated as linked to the intended NCTM focal points. Kindergarten and Grade 1 showed particularly strong alignment, with minor weaknesses in Kindergarten measurement.

Grades 3 through 7 maintained high alignment across Numbers and Operations, Geometry, Measurement, Algebra, and related focal points. Grade eight exhibited weaker alignment, especially in Geometry/Measurement, though Data Analysis showed more consistent alignment.

Depth of Knowledge analyses indicated that most items reflected lower to moderate cognitive demand, with the majority rated at DOK levels 1 (Recognition and Reproduction) or 2 (Skill and Concept). Few items reached DOK level 3 (Strategic Thinking), and consensus on DOK ratings was lower than for alignment judgments, highlighting the subjective nature of cognitive complexity ratings. Reliability estimates were strong for alignment ratings (ICCs generally above .80) and moderate for DOK ratings.

Across the report, tables are systematically organized by grade, focal point, and assessment type. Early tables summarize alignment frequencies and percentages for benchmark versus progress monitoring forms. Subsequent tables disaggregate results by rater, DOK level, and reliability indices, providing transparency and replication value. Overall, findings support strong content validity for easyCBM® math measures relative to NCTM Curriculum Focal Points. The main finding was strong alignment across grades K–7; weaker in Grade 8 Geometry with 75–100% items linked; ICCs .80–1.0

Table 1. Example of Analyses from Technical Report 1002

Table 63

Grade 3 Benchmark Measures: Individual Rater's Ratings on Strength of Link Between Items and Standards

			Raters		
Focal point	Term	Ratings	E	G	T
Number and operations	Fall	Not Linked (0)	0	0	0
		Vaguely Linked (1)	18.8	0	0
		Somewhat linked (2)	31.2	25.0	6.2
		Direct Linked (3)	50.0	75.0	93.8
	Winter	Not Linked (0)	0	0	0
		Vaguely Linked (1)	18.8	6.1	0
		Somewhat linked (2)	37.5	31.2	0
		Direct Linked (3)	43.8	62.5	100
	Spring	Not Linked (0)	0	0	0
		Vaguely Linked (1)	0	0	6.2
		Somewhat linked (2)	25.0	6.2	0
		Direct Linked (3)	75.0	93.8	93.8
Geometry	Fall	Not Linked (0)	0	0	0
		Vaguely Linked (1)	25.0	12.5	6.3
		Somewhat linked (2)	18.8	6.3	18.8
		Direct Linked (3)	56.3	81.3	75.0
	Winter	Not Linked (0)	0	0	0
		Vaguely Linked (1)	6.3	0	6.3
		Somewhat linked (2)	37.5	31.3	0
		Direct Linked (3)	56.3	68.8	93.8
	Spring	Not Linked (0)	6.3	0	6.3
		Vaguely Linked (1)	12.5	6.3	0
		Somewhat linked (2)	31.3	6.3	0
		Direct Linked (3)	50.0	87.5	93.8

Reference

Nese, J. F. T., Lai, C.-F., Anderson, D., Park, B. J., Tindal, G., & Alonzo, J. (2010). *The alignment of easyCBM® math measures to curriculum standards (Technical Report 1002)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Proficient Math

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

The mathematics alignment analysis follows the same coding framework as reading but reflects the fact that easyCBM® mathematics measures were developed explicitly with the CCSS in mind. Alignment evidence is drawn from structured coding in the Math 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® Math 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® math measures reflect those standards.

For CCSS **Adopted** states, nearly all mathematics standards were coded as CCSS Exact Matches, yielding consistent Yes alignment decisions across domains such as Operations and Algebraic Thinking, Number and Operations, Fractions, Geometry, and Measurement and Data.

CCSS **Revised** states showed a balance of Exact and Partial Matches. Partial matches typically reflected language modifications or reorganization of clusters rather than changes in mathematical intent. easyCBM® continued to align positively with most of these standards.

State **Unique** mathematics standards introduced a higher frequency of CCSS Deviations and Non-CCSS outcomes, including extensions beyond CCSS grade-level boundaries. Alignment with easyCBM® remained present for core concepts but was less comprehensive for unique or extended content.

Table 2. easyCBM® Mathematics Alignment Summary by State Group

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

Table 3. easyCBM® Mathematics Alignment Summary by Domain for Benchmark (BM) and Progress Measures (PM)

Domains	BM1	BM2	BM3	PMs
Counting & Cardinality	8	9	10	5
Expressions and Equations	18	17	15	35
Functions	7	6	8	15
Geometry	56	51	50	60
Measurement & Data	32	28	29	30
Number & Operations in Base 10	30	34	33	46
Numbers & Operations - Fractions	24	27	24	3
Operations & Algebraic Thinking	28	29	34	37
Ratios and Proportional Relations	17	9	20	6
Statistics and Probability	14	23	18	21
The Number System	23	24	22	25
Grand Total	**257**	**257**	**263**	**283**

Table 4. easyCBM® Mathematics Alignment Summary by Cluster for Benchmark (BM) & Progress Measures (PM)

Clusters	BM1	BM2	BM3	PMs
Add and subtract within 20.	2	3	3	0
Operations & Algebraic Thinking	2	3	3	0
Analyze and solve linear equations and pairs of simultaneous linear equations.	2	0	2	4
Expressions and Equations	2	0	2	4
Analyze patterns and relationships.	2	1	3	0
Operations & Algebraic Thinking	2	1	3	0
Analyze proportional relationships and use them to solve real-world and mathematical problems.	14	7	17	6
Ratios and Proportional Relations	14	7	17	6
Analyze, compare, create, and compose shapes.	1	2	2	6
Geometry	1	2	2	6
Apply and extend previous understandings of arithmetic to algebraic expressions.	2	0	1	3
Expressions and Equations	2	0	1	3
Apply and extend previous understandings of multiplication and division to divide fractions by fractions.	0	1	2	6
The Number System	0	1	2	6
Apply and extend previous understandings of multiplication and division.	1	1	1	0
Numbers & Operations - Fractions	1	1	1	0
Apply and extend previous understandings of numbers to the system of rational numbers.	3	3	1	0
The Number System	3	3	1	0
Apply and extend previous understandings of operations with fractions.	8	9	12	11
The Number System	8	9	12	11
Build fractions from unit fractions.	2	4	3	0

Numbers & Operations - Fractions	2	4	3	0
Classify objects and count the number of objects in each category.	1	1	2	0
Measurement & Data	1	1	2	0
Classify two-dimensional figures into categories based on their properties.	4	2	3	7
Geometry	4	2	3	7
Compare numbers.	0	1	1	5
Counting & Cardinality	0	1	1	5
Compute fluently with multi-digit numbers and find common factors and multiples.	2	4	0	8
The Number System	2	4	0	8
Count to tell the number of objects.	6	7	7	0
Counting & Cardinality	6	7	7	0
Define, evaluate, and compare functions.	6	3	4	6
Functions	6	3	4	6
Describe and compare measurable attributes.	2	2	3	0
Measurement & Data	2	2	3	0
Develop understanding of fractions as numbers.	10	10	10	0
Numbers & Operations - Fractions	10	10	10	0
Develop understanding of statistical variability.	2	4	4	0
Statistics and Probability	2	4	4	0
Draw and identify lines and angles and classify shapes by properties of their lines and angles.	1	1	1	0
Geometry	1	1	1	0
Draw construct and describe geometrical figures and describe the relationships between them.	10	9	10	5
Geometry	10	9	10	
Grand Total	**81**	**75**	**92**	**67**

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.

Summary of Technical Report 1228: Alignment of easyCBM® Grades K–2 Math Measures to the Common Core Standards (Irvin et al., 2012b).

Technical Report 1228 examined the alignment of easyCBM® Grades K–2 mathematics benchmark assessments with the Common Core State Standards (CCSS), with an emphasis on supporting valid instructional decisions in formative assessment and response-to-intervention frameworks. The study employed a two-phase expert review design. Participants included experienced general education teachers, special educators, and district-level mathematics specialists from multiple U.S. states, with average mathematics teaching experience ranging from approximately 10 to 12 years. In Phase 1, one expert per grade reviewed all 135 benchmark items (45 items each from fall, winter, and spring) and identified links to on-grade and prior-grade CCSS. Phase 2 expanded the review to four additional experts per grade who completed structured training and conducted independent online reviews.

Data collection was conducted using the Distributed Item Review (DIR) system, which presented items individually and required reviewers to assign CCSS codes and rate strength of alignment (0 = no alignment, 1 = somewhat aligned, 2 = directly aligned). Reviewers could also indicate alignment to prerequisite skills necessary for on-grade mastery. Data preparation involved collapsing sub-standards, correcting rating inconsistencies, and aggregating ratings across reviewers. Analyses focused on identifying primary and secondary standards for each item based on frequency counts and computing average alignment strength ratings. Results indicated strong overall alignment across grades, with approximately 94% of kindergarten, 99% of Grade 1, and 96% of Grade 2 items aligned to on-grade or prior-grade CCSS. Alignment was consistently stronger at the domain level than at the individual standard level. However, notable gaps were identified, including underrepresentation of certain Geometry, Measurement and Data, and Operations and Algebraic Thinking standards, informing targeted assessment redevelopment.

Table 5. easyCBM® Mathematics Alignments and Key Findings (Grades K–2)

Grade	Overall Alignment	Key Findings
Kindergarten	≈94% aligned	Strong domain coverage: gaps in NBT and select CC standards
Grade 1	≈99% aligned	On-grade alignment strong; Measurement & Data underrepresented
Grade 2	≈96% aligned	Geometry and select OA standards underrepresented

Table 2
CCSS domain and grade level alignment results for the easyCBM® kindergarten fall benchmark in mathematics.

Domain	# ps items	# sec items
K.CC	8	14
K.G	16	12
K.MD	5	13
K.NBT	0	1
K.OA	3	7
Grade K (overall)	32	47

Note. Domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 4
CCSS domain and grade level alignment results for the easyCBM® kindergarten winter benchmark in mathematics.

Domain	# ps items	# sec items
K.CC	8	13
K.G	16	7
K.MD	7	6
K.NBT	0	0
K.OA	4	8
Grade K (overall)	35	34

Note. Domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 6
CCSS domain and grade level alignment results for the easyCBM® kindergarten spring benchmark in mathematics.

Domain	# ps items	# sec items
K.CC	7	12
K.G	15	7
K.MD	7	13
K.NBT	0	0
K.OA	6	11
Grade K (overall)	35	43

Note. Domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 8
CCSS domain and grade level alignment results for the easyCBM® first grade fall benchmark in mathematics.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
K.CC	1	2	1.G	3	8
K.G	11	5	1.MD	1	3
K.MD	0	0	1.NBT	11	9
K.NBT	0	0	1.OA	10	15
K.OA	0	0			
Grade K (overall)	12	7	Grade 1 (overall)	25	35

Note. On- and prior-grade domains are labeled using the unique CCSS identification code. # ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 10
CCSS domain and grade level alignment results for the easyCBM® first grade winter benchmark in mathematics.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
K.CC	2	11	1.G	4	9
K.G	6	12	1.MD	1	0
K.MD	0	0	1.NBT	13	6
K.NBT	0	1	1.OA	10	18
K.OA	2	2			
Grade K (overall)	10	26	Grade 1 (overall)	28	33

Note. On- and prior-grade domains are labeled using the unique CCSS identification code. # ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 12
CCSS domain and grade level alignment results for the easyCBM® first grade spring benchmark in mathematics.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
K.CC	5	10	1.G	7	10
K.G	6	10	1.MD	2	2
K.MD	0	0	1.NBT	14	5
K.NBT	0	0	1.OA	6	26
K.OA	2	5			
Grade K (overall)	13	25	Grade 1 (overall)	29	43

Note. On- and prior-grade domains are labeled using the unique CCSS identification code. # ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 14
CCSS domain and grade level alignment results for the easyCBM® second grade fall benchmark in mathematics.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
1.G	0	0	2.G	0	3
1.MD	4	3	2.MD	13	4
1.NBT	1	6	2.NBT	18	8
1.OA	1	5	2.OA	5	6
Grade 1	6	14	Grade 2	36	21

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 16
CCSS domain and grade level alignment results for the easyCBM® second grade winter benchmark in mathematics.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
1.G	0	0	2.G	0	2
1.MD	2	0	2.MD	11	3
1.NBT	5	5	2.NBT	19	8
1.OA	1	2	2.OA	4	4
Grade 1	8	7	Grade 2	34	17

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 18
CCSS domain and grade level alignment results for the easyCBM® second grade spring benchmark in mathematics.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
1.G	0	0	2.G	0	2
1.MD	4	0	2.MD	15	9
1.NBT	4	3	2.NBT	15	19
1.OA	0	4	2.OA	6	4
Grade 1	8	7	Grade 2	36	34

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Reference

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012*). The alignment of the easyCBM® Grades K–2 math measures to the Common Core State Standards (Technical Report 1228).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1229: The Alignment of the easyCBM® Grades 3-5 Math Measures to the Common Core Standards (Park et al., 2012).

Technical Report 1229 examined the alignment of easyCBM® Grades 3–5 mathematics benchmark assessments with the Common Core State Standards (CCSS) to support valid instructional decision-making within response-to-intervention frameworks. The study used a structured two-phase expert review design. Participants included experienced general education teachers, special education teachers, and district math specialists from multiple U.S. states, averaging approximately 10–12 years of mathematics teaching experience. All reviewers demonstrated familiarity with CCSS and formative assessment systems.

Data collection involved reviewing all 135 benchmark items per grade (45 items each for fall, winter, and spring). In Phase 1, one reviewer per grade grouped items by mathematical skill and identified alignment to on-grade and prior-grade CCSS. Phase 2 expanded reviews to four additional trained reviewers per grade using a secure Distributed Item Review (DIR) system. Reviewers identified aligned standards and rated alignment strength (0 = none, 1 = somewhat aligned, 2 = directly aligned), including identification of prerequisite skills for on-grade mastery.

Analyses included data cleaning, frequency counts of selected standards, identification of primary and secondary alignments, and calculation of mean alignment strength ratings. Results showed strong overall alignment: approximately 98% of Grade 3 items, 100% of Grade 4 items, and 97% of Grade 5 items aligned to on-grade or prior-grade CCSS. Alignment was strongest in Operations and Algebraic Thinking and Number and Operations–Fractions. Geometry and some Measurement and Data standards were underrepresented, particularly in Grade 5, while several Number and Operations in Base Ten standards were overrepresented. Findings guided targeted item development to improve balance and coverage across CCSS domains.

Table 6. easyCBM® Mathematics Alignments and Key Findings (Grades 3–5)

Grade	Items Aligned	Strongly Represented Domains	Underrepresented Domains
Grade 3	≈98%	OA, NF	NBT, MD
Grade 4	100%	MD, NF	Geometry
Grade 5	≈97%	NBT, NF	Geometry, OA

Table 2
CCSS domain and grade level alignment results for the easyCBM® third grade fall benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
2.G	5	9	3.OA	17	19
2.MD	0	1	3.NBT	1	0
2.NBT	0	0	3.NF	14	3
2.OA	0	3	3.MD	1	1
			3.G	3	16
Grade 2	5	13	Grade 3	36	39

Note. On- and prior-grade domains are labeled using the unique CCSS identification code. # ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 4
CCSS domain and grade level alignment results for the easyCBM® third grade winter benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
2.G	5	11	3.OA	14	22
2.MD	0	1	3.NBT	0	0
2.NBT	0	0	3.NF	14	8
2.OA	0	0	3.MD	2	0
			3.G	0	9
Grade 2	5	12	Grade 3	30	39

Note. On- and prior-grade domains are labeled using the unique CCSS identification code. # ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 6
CCSS domain and grade level alignment results for the easyCBM® third grade spring benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
2.G	4	8	3.OA	14	19
2.MD	0	0	3.NBT	0	0
2.NBT	0	0	3.NF	8	6
2.OA	0	0	3.MD	1	0
			3.G	1	11
Grade 2	4	8	Grade 3	24	36

Note. On- and prior-grade domains are labeled using the unique CCSS identification code. # ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 8
CCSS domain and grade level alignment results for the easyCBM® fourth grade fall benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
3.OA	4	9	4.OA	2	10
3.NBT	1	1	4.NBT	3	3
3.NF	2	3	4.NF	5	8
3.MD	11	9	4.MD	11	12
3.G	0	5	4.G	0	0
Grade 3	18	27	Grade 4	21	33

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 10
CCSS domain and grade level alignment results for the easyCBM® fourth grade winter benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
3.OA	1	1	4.OA	1	7
3.NBT	0	2	4.NBT	8	3
3.NF	1	3	4.NF	7	5
3.MD	11	16	4.MD	9	6
3.G	0	0	4.G	0	0
Grade 3	13	22	Grade 4	25	21

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 12
CCSS domain and grade level alignment results for the easyCBM® fourth grade spring benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
3.OA	4	2	4.OA	4	7
3.NBT	1	2	4.NBT	3	6
3.NF	1	2	4.NF	6	8
3.MD	11	5	4.MD	8	15
3.G	0	0	4.G	0	1
Grade 3	17	11	Grade 4	21	37

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 14
CCSS domain and grade level alignment results for the easyCBM® fifth grade fall benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
4.OA	0	0	5.OA	0	1
4.NBT	4	10	5.NBT	11	11
4.NF	0	11	5.NF	8	7
4.MD	2	2	5.MD	3	13
4.G	0	0	5.G	0	0
Grade 4	6	23	Grade 5	22	32

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 16
CCSS domain and grade level alignment results for the easyCBM® fifth grade winter benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
4.OA	0	0	5.OA	0	1
4.NBT	5	7	5.NBT	7	18
4.NF	6	6	5.NF	5	2
4.MD	1	1	5.MD	10	10
4.G	0	0	5.G	0	0
Grade 4	12	14	Grade 5	22	31

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 18
CCSS domain and grade level alignment results for the easyCBM® fifth grade spring benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
4.OA	0	0	5.OA	0	1
4.NBT	4	10	5.NBT	11	11
4.NF	0	11	5.NF	8	7
4.MD	2	2	5.MD	3	13
4.G	0	0	5.G	0	0
Grade 4	6	23	Grade 5	22	32

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Reference

Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM® Grades 3–5 math measures to the Common Core State Standards (Technical Report 1229)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1230: The Alignment of the easyCBM® Grades 6-8 Math Measures to the Common Core Standards (Irvin et al., 2012a).

Technical Report 1230 examined the alignment between easyCBM® mathematics benchmark assessments for grades 6–8 and the Common Core State Standards (CCSS). The study used a two-phase expert review design. In Phase 1, one experienced educator per grade reviewed 135 benchmark items (45 each for fall, winter, and spring) to identify potential on-grade and prior-grade CCSS alignments. Phase 2 expanded the review to four additional educators per grade, all of whom completed training and conducted item-level alignment ratings using a secure Distributed Item Review (DIR) system. Reviewers identified applicable CCSS standards, rated alignment strength, and indicated whether items measured prerequisite skills for on-grade mastery.

Data analyses involved cleaning reviewer responses, collapsing CCSS sub-standards, and calculating frequencies of standards selected per item. Primary standards were defined as those most frequently selected by reviewers, with secondary standards also recorded. Average alignment strength ratings were calculated using Phase 2 data. Results showed strong overall alignment across grades, with approximately 99% of grade 6 items, 93% of grade 7 items, and 96% of grade 8 items aligned to on- or prior-grade CCSS. Alignment was generally stronger to on-grade standards than prior-grade standards.

Despite strong overall alignment, findings revealed systematic over- and underrepresentation of specific domains and standards. Ratios and Proportional Relationships and Expressions and Equations were often overrepresented, while Statistics and Probability, Geometry, and Number System standards were underrepresented in several grades. These results informed targeted assessment development plans to improve CCSS coverage and strengthen the validity of instructional decision-making based on easyCBM® math assessments.

Table 7. easyCBM® Mathematics Alignments and Key Findings (Grades 6–8)

Grade	Overall Alignment	Key Domain-Level Findings
6	≈99% aligned	Overrepresentation of Ratios & Expressions; underrepresentation of Statistics, Geometry
7	≈93% aligned	Strong on-grade alignment; Geometry and Number System overrepresented
8	≈96% aligned	Functions overrepresented; Number System and Statistics underrepresented

Table 2
CCSS domain and grade level alignment results for the easyCBM® sixth grade fall benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
5.OA	0	0	6.RP	11	11
5.NBT	2	6	6.NS	0	2
5.NF	8	7	6.EE	15	26
5.MD	1	1	6.G	0	3
5.G	0	0	6.SP	0	4
Grade 5	11	14	Grade 6	26	46

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 4
CCSS domain and grade level alignment results for the easyCBM® sixth grade winter benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
5.OA	2	2	6.RP	10	9
5.NBT	3	2	6.NS	5	4
5.NF	4	6	6.EE	11	20
5.MD	0	0	6.G	0	0
5.G	0	0	6.SP	0	6
Grade 5	9	10	Grade 6	26	39

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 6
CCSS domain and grade level alignment results for the easyCBM® sixth grade spring benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
5.OA	1	4	6.RP	12	5
5.NBT	2	6	6.NS	2	3
5.NF	5	8	6.EE	16	25
5.MD	0	0	6.G	1	2
5.G	0	0	6.SP	0	5
Grade 5	8	18	Grade 6	31	40

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 8
CCSS domain and grade level alignment results for the easyCBM® seventh grade fall benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
6.RP	11	3	7.RP	0	2
6.NS	1	9	7.NS	3	4
6.EE	1	2	7.EE	1	1
6.G	1	1	7.G	11	7
6.SP	0	1	7.SP	1	1
Grade 6	14	16	Grade 7	16	15

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 10
CCSS domain and grade level alignment results for the easyCBM® seventh grade winter benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
6.RP	6	4	7.RP	0	6
6.NS	2	1	7.NS	7	8
6.EE	1	6	7.EE	4	4
6.G	0	0	7.G	16	2
6.SP	0	0	7.SP	2	2
Grade 6	9	11	Grade 7	29	22

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 12
CCSS domain and grade level alignment results for the easyCBM® seventh grade spring benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
6.RP	6	9	7.RP	3	2
6.NS	1	4	7.NS	8	11
6.EE	0	3	7.EE	4	4
6.G	0	2	7.G	14	2
6.SP	0	0	7.SP	1	2
Grade 6	7	18	Grade 7	30	21

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 14
CCSS domain and grade level alignment results for the easyCBM® eighth grade fall benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
7.RP	0	2	8.NS	0	0
7.NS	0	0	8.EE	3	6
7.EE	4	3	8.F	7	13
7.G	5	7	8.G	5	9
7.SP	4	15	8.SP	1	0
Grade 7	13	27	Grade 8	16	28

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 16
CCSS domain and grade level alignment results for the easyCBM® eighth grade winter benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
7.RP	1	2	8.NS	0	0
7.NS	0	1	8.EE	1	6
7.EE	1	3	8.F	11	7
7.G	2	7	8.G	9	11
7.SP	0	14	8.SP	0	1
Grade 7	4	27	Grade 8	21	25

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Table 18
CCSS domain and grade level alignment results for the easyCBM® eighth grade spring benchmark in math.

Domain	# ps items	# sec items	Domain	# ps items	# sec items
7.RP	0	0	8.NS	0	1
7.NS	0	0	8.EE	2	6
7.EE	1	1	8.F	10	14
7.G	5	7	8.G	7	10
7.SP	2	17	8.SP	2	3
Grade 7	8	25	Grade 8	21	34

Note. On- and prior-grade domains are labeled using the unique CCSS identification code.
ps items = the number of items identified as aligned as primary standard to a given CCSS domain; # sec items = the number of items identified as aligned as secondary standard to that CCSS domain.

Reference

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM® grades 6–8 math measures to the Common Core State Standards (Technical Report 1230)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1208: The Alignment of the easyCBM® Middle School Mathematics CCSS Measures to the Common Core State Standards (Anderson et al., 2012).

Technical Report 1208 documents a formal alignment study examining the extent to which the easyCBM® Middle School Mathematics measures correspond to the Common Core State Standards (CCSS). The study provides critical validity evidence supporting the instructional use of easyCBM® within a formative and Response to Intervention (RTI) framework. The report focuses on alignment quality, rater agreement, and consistency across grades and mathematical domains.

Methods

The alignment study employed a structured, standards-based review process. Approximately 50% of the available easyCBM® middle school mathematics item pool across grades 6–8 was randomly selected for analysis. Items were originally developed by a team of experienced middle school mathematics teachers to reflect CCSS domains including Ratios and Proportional Relationships, The Number System, Expressions and Equations, Geometry, Functions, and Statistics and Probability.

Fifteen practicing middle school mathematics teachers from across the United States served as alignment raters. All raters completed standardized training delivered via webinar, which introduced the CCSS, alignment definitions, rating criteria, and scoring procedures. Raters conducted their reviews using the Distributed Item Review (DIR) online platform, which ensured consistency in item presentation and data collection.

Each item was evaluated against its intended CCSS using a four-point alignment scale ranging from 0 (no alignment) to 3 (direct alignment). For analytic purposes, ratings were collapsed into aligned (direct or requisite) versus not aligned categories. Rasch measurement modeling was applied to examine rater severity, item 'endorsability', and overall model fit, allowing for control of rater effects in estimating alignment outcomes.

Results

Results indicated strong overall alignment between easyCBM® middle school mathematics measures and the CCSS. After adjusting for rater severity, 87% of items were classified as directly aligned to their intended standards. When items measuring requisite or prerequisite skills were included, alignment rose to 99.6%, indicating near-complete coverage of CCSS-relevant content.
Rater effects were present but modest. Rasch analyses demonstrated acceptable fit statistics, suggesting that differences in rater severity did not meaningfully distort alignment conclusions. Overall rater agreement was sufficient to support the reliability of the alignment judgments.

Analyses across grades and mathematical domains revealed only minor variation in alignment rates. No grade-level or domain-specific weaknesses were identified that would undermine the use of the measures for formative or progress-monitoring purposes. Alignment consistency across grades 6–8 supports the vertical coherence of the easyCBM® mathematics system.

Appendix Table Structure

The appendices provide detailed transparency into the alignment process. Appendix A documents rater training materials and alignment criteria. Appendix B presents Rasch model outputs, including rater severity estimates and item 'endorsability' statistics. Appendix C contains item-level alignment tables organized by grade, domain, CCSS code, and rater ratings, enabling replication and secondary analyses. **Overall**, the study provides strong evidence that easyCBM® middle school mathematics measures are well aligned to the CCSS and suitable for instructional decision-making.

Table 8. easyCBM® Mathematics Alignments and Key Results with Evidence Leading to Interpretations

Finding Area	Key Result	Methodological Evidence	Interpretation
Overall Alignment	87% direct alignment; 99.6% including requisite skills	Rasch-adjusted alignment ratings	Strong correspondence to CCSS
Rater Effects	Minimal rater severity bias	Rasch rater severity estimates	Alignment judgments are reliable
Grade-Level Patterns	Consistent across grades 6–8	Grade-disaggregated Rasch analyses	Vertical coherence supported
Domain Coverage	High alignment across all CCSS domains	Domain-level alignment summaries	Broad CCSS content coverage

Reference

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM® middle school mathematics CCSS measures to the Common Core State Standards (Technical Report 1208).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Conclusion

This document summarizes the alignment of easyCBM® mathematics measures with state and national standards, primarily focusing on the Common Core State Standards (CCSS) and NCTM Focal Points.

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

Expert Review: Experienced teachers and specialists reviewed items to assess alignment with specific CCSS domains/clusters and NCTM Focal Points. This involved rating the strength of alignment on a scale (e.g., 0-3 or specific categories like "Very High" to "Minimal").

Content Blueprinting: A content blueprint operationalized the construct by specifying targeted standards, the number of items per domain, and the balance of item types, complexity, and difficulty.

Data Analysis: Alignment was quantified by calculating the frequency of items aligned to primary and secondary standards and computing mean alignment strength ratings. Rasch measurement modeling was used to control for rater severity. For CCSS-aligned states, alignment was coded as Exact Match, Partial Match, Deviation, or Non-CCSS.

Results: The easyCBM® mathematics measures demonstrated generally strong alignment with both NCTM Focal Points and CCSS standards across grades K-8.

NCTM Alignment: For the Basic Math measures, alignment with NCTM Focal Points was generally strong across grades K-7, with between 75% and 100% of items rated as linked. Grade 8 showed weaker alignment in Geometry/Measurement, but Data Analysis was more consistent.

CCSS Alignment

- K-2 Math: Overall alignment was high, with approximately 94% of kindergarten, 99% of Grade 1, and 96% of Grade 2 items aligned to on-grade or prior-grade CCSS. Alignment was strongest at the domain level, though some underrepresentation of specific standards (e.g., Geometry, Measurement & Data, Operations & Algebraic Thinking) was noted, informing future item development.
- 3-5 Math: Strong overall alignment was found, with approximately 98% of Grade 3, 100% of Grade 4, and 97% of Grade 5 items aligned. Alignment was strongest in Operations & Algebraic Thinking and Number & Operations-Fractions. Geometry and Measurement & Data were underrepresented in Grade 5.
- 6-8 Math: Strong overall alignment was observed, with 99% of Grade 6, 93% of Grade 7, and 96% of Grade 8 items aligned. Alignment was generally stronger to on-grade standards. However, Ratios & Proportional Relationships and Expressions & Equations were often overrepresented, while Statistics & Probability and Geometry were underrepresented in some grades.

Overall, the findings indicate that the easyCBM® mathematics measures are well-aligned to relevant standards, providing a valid basis for instructional decision-making.

Appendix A: Technical Report Table Titles

Technical Report References

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM middle school mathematics CCSS measures to the Common Core State Standards (Technical Report # 1208)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012a). *The alignment of the easyCBM Grades 6-8 math measures to the Common Core Standards (Technical Report # 1230)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012b). *The alignment of the easyCBM Grades K-2 math measures to the Common Core Standards (Technical Report # 1228)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nese, J. F. T., Lai, C. F., Anderson, D., Park, B. J., Tindal, G., & Alonzo, J. (2010). *The alignment of easyCBM math measures to curriculum standards (Technical Report #1002)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM grades 3-5 math measures to the Common Core Standards (Technical Report # 1229)*. 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 MGCI_1.3 documentation and finds strong evidence that easyCBM® mathematics assessments are fair, accessible, and appropriate for the full intended test-taking population.

For **Indicator 1.3.a**, item and test-event development procedures demonstrate consistent adherence to principles of Universal Design for Assessment (UDA). Student-facing math measures are presented in uncluttered formats with clear typography, generous spacing, high contrast, and minimal visual distractions, reducing construct-irrelevant variance. Consistency across paper-and-pencil and online formats supports comparability of scores. Items undergo systematic content and bias review by experienced educators and researchers, followed by national field testing and psychometric evaluation, ensuring technical quality and subgroup appropriateness consistent with the *Standards for Educational and Psychological Testing*.

Regarding **Indicator 1.3.b**, the documentation provides a comprehensive, well-differentiated framework for accessibility supports, accommodations, and non-allowable modifications. The intended test-taking population is clearly defined, and accommodations are explicitly aligned with intended uses of the assessment. A wide range of setting, administration, and response accommodations is supported, drawing from NCEO, Smarter Balanced, and NAEP guidance. Crucially, the documentation clearly distinguishes accommodations that preserve validity from modifications that alter the construct (e.g., calculators for operational math items or extended time on fluency measures). This clarity supports equitable access while protecting the validity of score interpretations for students with disabilities and English Learners.

For **Indicator 1.3.c**, the range of technology-enhanced administration options further supports accessibility without compromising validity. Online administration includes scalable text, compatibility with assistive technologies, built-in timers, live scoring, and flexible navigation, while maintaining a clean, non-distracting interface. Paper-and-pencil options remain fully supported, ensuring access in varied school contexts. Guidance is provided for appropriate platform use, administration conditions, and testing times.

Overall, MGCI_1.3 demonstrates a rigorous, standards-aligned approach to fairness and accessibility, with strong procedural, technical, and documentation evidence supporting equitable access and defensible interpretation of mathematics assessment results.

1.3a Math Measurement Presentation and Fair Access

This section provides formal measure descriptions for the easyCBM® Basic Math and Proficient Math measures. Descriptions are written in accordance with the Standards for Educational and Psychological Testing (AERA, APA, NCME), with emphasis on how the measures appear to students and teachers, administration modes, accessibility features, and consistency across paper-and-pencil and online formats.

The **Basic Math** and **Proficient Math** assessments are untimed, group-administered multiple-choice measures of mathematics skills. For benchmark testing, both assessments measure a range of skills closely aligned with a variety of state content standards in mathematics. The Basic Math measures were developed using the National Council of Teachers of Mathematics (NCTM) Focal Point Standards as an initial framework, with benchmark forms including test items from all three focal point standards at each respective grade level. The Proficient Math measures were developed using the Common Core State Standards (CCSS) as an initial framework. In addition to items aligned with the respective grade level, the Proficient Math benchmark measures also include a small number of items from prior and subsequent

grade levels to enhance the test's accuracy as a universal screener, thereby extending the population of students whom it reliably measures. The Basic Math measures were designed to be more easily accessible (fewer cognitive demands for processing what is being asked) and to assess a more foundational understanding of math, making them most appropriate for students who are performing substantially below their grade-level peers. The Proficient Math measures were designed to be more challenging, in line with high expectations of grade-level performance.

Student-Facing Materials

Overall appearance

From the student's perspective, the measures appear as a single, uncluttered pages of text and graphics containing only the items. No scoring marks, numbers, or cues are visible to the student. The design minimizes visual distractions so that performance reflects math skill rather than navigation or formatting demands.

Layout (Online and Paper-Pencil)

- Two-column layout with consistent margins.
- A prompt presented in a box with three options below.
- Minimum reading demands.

Font and typography (Online and Paper-Pencil)

- Sans-serif font (e.g., Arial or comparable): Chosen for clarity and readability across grades.
- Large, grade-appropriate font size: Font size increases for lower grades and gradually decreases across upper grades while remaining comfortably legible.
- Standard capitalization and punctuation: Reflects authentic grade-level text rather than simplified or artificially segmented print.

Spacing and readability (Online and Paper-Pencil)

- Generous line spacing: Reduces visual crowding and supports smooth eye movement.
- Consistent word spacing: Avoids compression that could artificially inflate error rates.
- Clear line breaks with prompts and four items per page.

Accessibility considerations (student)

- High contrast (black text on white background).
- No background shading or watermarks.
- Compatible with magnification and print enlargement.
- Passages can be printed or displayed digitally without altering layout.

Teacher/Assessor-Facing Materials

Overall appearance

- The assessor copy is designed for later error marking.

Layout (Online and Paper-Pencil)

- Separate scoring form for paper-pencil.
- Includes fields for: Skill Area, Student name, and Date.

Font and spacing (assessor)

- Same font and font size as student copy to preserve alignment.
- Adequate spacing to support rapid eye tracking during oral reading.
- Clear visual separation between prompt and options.

Administration experience

- The student experiences PRF as “problem-solving”, with no visible scoring or performance pressure cues.
- The assessor experiences measure as a highly structured, standardized scoring task, supported by consistent layout and marking space.

Accessibility and universal design features

- Minimal visual clutter reduces cognitive load.
- Consistent formatting across grades supports comparability.
- Print and digital compatibility allows administration in varied school contexts.
- Oral directions ensure reading performance is not confounded by comprehension of written instructions.

Grade K–2: Basic Math Measures

Basic Math measures for grades Kindergarten through Grade 2 are designed to assess foundational number sense, operations, and mathematical reasoning appropriate for early learners. Student-facing forms consist of single-page or short multi-page booklets with uncluttered layouts, large sans-serif fonts, generous spacing between items, and minimal visual distractions. Items are presented one per line or with ample white space to support visual tracking. Teachers receive standardized administration directions, scoring guides, and recording forms.

Paper-and-pencil administration uses high-contrast print optimized for photocopying, while the online format presents items one at a time or in vertically spaced lists. Navigation controls are simple and consistent. Accessibility features include screen-reader compatibility, keyboard navigation, and accommodation support consistent with individualized education plans.

Grade 3–5: Basic and Transition to Proficient Math

In Grades 3 through 5, easyCBM® math measures emphasize computation, problem solving, and application aligned to grade-level standards. Student forms use readable fonts, structured item groupings, and consistent response areas. Visual elements such as number lines or simple graphics are used sparingly and only when essential to the construct being measured.

Teacher materials include administration scripts, scoring rubrics, and guidance for interpreting results. Both paper and online versions maintain equivalent layouts to minimize mode effects. Online forms support zoom, text-to-speech where appropriate, and clear progress indicators.

Grade 6–8: Proficient Math Measures

Proficient Math measures for Grades 6 through 8 assess grade-level mathematical proficiency, including conceptual understanding, procedural fluency, and application. Student-facing assessments present problems in a clean, well-structured format with clear separation between items, consistent mathematical notation, and sufficient space for computation.

Teacher-facing materials emphasize standardized administration and scoring consistency. Online delivery mirrors paper formats while allowing for dynamic resizing, assistive technology integration, and secure test delivery. Accessibility features are designed to reduce construct-irrelevant variance while preserving the integrity of the mathematical tasks.

Administration Modes and Accessibility

Across all grades, easyCBM® Math measures are designed for equivalence across administration modes. Paper-and-pencil formats support traditional classroom administration, while online delivery supports remote or computer-based testing. Accessibility considerations include font clarity, spacing, contrast,

compatibility with assistive technologies, and accommodation flexibility, consistent with universal design principles and testing standards.

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® afterward. 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.

Paper-and-Pencil

The easyCBM Basic Math measures are optimized for online administration, which includes read-aloud options built into the system. If you choose to administer these measures in paper-and-pencil format, you will need to read out loud any items that include text. To do so, you will need to direct the students to complete each item one by one, then raise their hands when they're ready to move on to the next item. Keep the whole class together so you know you're reading the text for the item each student is answering.

1. Hand out student copies of the test form to each student.
2. Say, **"Today you'll be taking a math test. It doesn't count toward your grade, but it will help us decide what we should cover in class. Please try your best. Start by writing your name and the date at the top of the page. Circle your answers. Ready? Begin."**
3. During the test, walk around the room to monitor what students are doing and help them stay on task.

Alignment with Testing Standards

Measure design, layout, administration procedures, and accessibility features align with the Standards for Educational and Psychological Testing[1] by supporting fairness, validity, reliability, and appropriate score interpretation. Consistency of presentation across modes and grades supports comparability and defensible use of results for screening, progress monitoring, and instructional decision-making.

1.3b Math 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: 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).

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)

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
[2] Rose, D. H. (2006). *A Practical Reader in Universal Design for Learning*. Harvard Education Press.
[3] Tindal, G. (2025). *Rethinking "Standardization" for NAEP to Increase Equity and Access*. University of Oregon: Behavioral Research and Teaching. Technical Report 2510.

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.

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 1. 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 2. 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 3. 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® afterward. 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 4. 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	

Testing Times

Basic and Proficient Math measures each take approximately 20 minutes for students in Grades K-2 and 30 minutes for students in Grades 3-8. Please note that these times do not include reading directions or transitioning between tests.

Appendix A: Technical Report Table Titles

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 and are reliable. The document describes a multi-stage development process applied across multiple technical reports. TR 0908 details the development of Grade 7 Basic Math items: 912 items were generated through a structured pipeline, piloted with approximately 2,800 students, and analyzed via Rasch calibration and distractor analysis before forms were assembled. TR 1405 documents large-scale operational piloting across Grades K–8 with more than 135,000 students, establishing the breadth of the item pool supporting Proficient Math and CCSS-aligned forms. Items are reviewed internally for standards alignment, mathematical accuracy, and clarity, and externally for grade-level appropriateness, bias/sensitivity, and usability. Distractor analysis is applied post-piloting to evaluate item quality and flag items for revision or removal. For online administration, items are displayed one per screen with randomizable response options, standardizing the test event across administrations. Administration is standardized across online and paper-and-pencil delivery contexts, reducing examiner-level error. Taken together, these development, review, and piloting procedures constitute a systematic, evidence-based approach to producing math test events that meet technical quality standards prior to operational deployment.

Assessment design specifications and test development and review procedures ensure test events meet content and statistical quality criteria. Statistical quality criteria are documented and applied through both Item Response Theory and Classical Test Theory frameworks across all technical reports. Rasch calibration (1PL, implemented in Winsteps) provides item difficulty estimates on a common logit scale and fit statistics to identify items with unexpected response patterns. Anchor item designs in TR 1405 support form comparability by linking item parameters across administrations and ensuring score-scale consistency. CTT-based analyses complement Rasch results: Cronbach's alpha evaluates internal consistency, and inter-form correlations assess alternate-form equivalence. TR 2602 uses Rasch marginal reliability as the primary reliability index for Proficient Math, providing an IRT-based analogue to classical alpha. TR 1312 reports generalizability coefficients alongside alpha for CCSS Grades 6–8 research forms, partitioning variance to evaluate score consistency across forms. Content quality criteria—alignment to NCTM Focal Points or CCSS-M by grade, blueprint-controlled item distributions, and visual and linguistic accessibility requirements—are reviewed before and during piloting. These intersecting statistical and content quality criteria provide a rigorous framework for ensuring that math test events are both psychometrically sound and standards-aligned before operational use.

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 across Basic Math, Proficient Math, and CCSS-aligned forms. Blueprint-based form construction specifies item counts by domain and difficulty range per grade, ensuring that each form samples the construct consistently and reduces variance attributable to form-construction differences. For computation measures, item difficulty is calibrated through Rasch modeling prior to form assembly, ensuring forms are parallel in difficulty and produce interchangeable scores. Online administration under standardized conditions removes examiner variability as a source of error. The document notes that form length is a recognized reliability constraint and recommends that longer forms—or optimized item targeting via IRT—are appropriate remedies when reliability is insufficient. Blueprint constraints, review procedures, and IRT-guided form assembly are therefore reliability-aware by design, establishing at the development stage the conditions necessary for dependable student scores across administrations.

Procedures for calculating and evaluating reliability are well-documented and appropriate. Multiple reliability methods are documented and applied across technical reports, each matched to the measure type and intended use. Cronbach's alpha and split-half reliability are the primary internal consistency indices for Basic Math and Proficient Math total scores (TR 0915, TR 1006, TR 1405). Rasch marginal reliability is reported for Proficient Math forms (TR 2602), providing an IRT-based precision index that accounts for the distribution of student ability relative to item difficulty. Generalizability coefficients and Phi coefficients are reported for CCSS research forms in Grades 6–8 (TR 1312), partitioning variance to evaluate relative and absolute decision consistency. Standard errors of measurement are computed from reliability estimates and score standard deviations. Slope reliability for growth monitoring is estimated via two-level hierarchical linear models across fall–winter–spring (TR 1804), reported by grade and quartile. This multi-method approach is appropriate to the range of measures, grades, and uses documented throughout the technical reports.

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 total-score measures and variable for growth indices. Cronbach's alpha for Basic Math total scores ranges from .80–.87 across Grades 1–8 (TR 0915) and .78–.89 across Grades K–2 (TR 1006), reaching .81–.95 with a median of .90 in the large-scale TR 1405 sample of more than 135,000 students. Rasch marginal reliability for Proficient Math ranges from .81–.87 across all grades and seasons (TR 2602), supporting use for benchmark screening. Research-version CCSS forms in Grades 6–8 (TR 1312) yielded lower alpha (.55–.79) and generalizability coefficients (.62–.79), with the document recommending revision before operational deployment. Slope reliability from HLM analyses (TR 1804) is variable across grades and quartiles, with some grades showing moderate values but many showing lower estimates, indicating that growth scores require additional administrations or longer measurement windows to support high-stakes individual decisions. Together, these values support total-score math measures for screening and benchmark purposes, with supplementary administrations recommended where slope precision is needed.

Abstract

This technical document summarizes primary studies conducted at Behavioral Research and Teaching (BRT) in the development and validation, specifically reliability, of easyCBM® Math measures for Grades K – 8. All studies present summaries of results and then illustrative findings with screen shots of exemplary tables. Note that all primary studies can be obtained at https://brtprojects.org. **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.

Definition and Types of 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).

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.

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.

Summary of Technical Report Findings

This section summarizes reliability-related findings reported in the attached synthesis of Behavioral Research and Teaching (BRT) technical reports focused on easyCBM® mathematics measures. These findings are illustrative: this document is a review of primary studies conducted by BRT researchers and cataloged on the BRT technical report site (https://brtprojects.org). The studies vary in grade span, sample composition, item types, administration seasons, and analytic methods; therefore, the patterns below should be interpreted as evidence within those specific study contexts rather than as universal constants. Across the reviewed reports, internal consistency evidence is consistently strong for total mathematics scores used for screening and benchmarking.

Technical Report 0915 examined internal consistency of **mathematics general outcome measures** in Grades 1–8 using large grade-level samples from a mid-sized Oregon school district. Cronbach's alpha for total scores was reported in the strong range across grades (approximately .80 to .87). The report also documented item-level descriptives and inter-item correlations consistent with broad-domain outcome measures: modest but positive correlations and a spread of difficulties that supported differentiation among students. Subtest reliabilities were lower, which the report attributed largely to the reduced number of items per focal point domain.

Technical Report 1006 extended reliability evidence to the **primary-level mathematics** and situated reliability within a broader technical adequacy argument for Grades K–2. Using large regional samples (thousands of students per grade), the report evaluated benchmark measures across seasons using Cronbach's alpha and split-half estimates. Across K–2 and seasons, internal consistency was strong (alphas roughly .78 to .89). The split-half coefficients were more moderate, as expected for shorter halves, but provided converging evidence of score consistency. Importantly, the report also examined growth reliability using two-level hierarchical linear models. Slope reliability was adequate for students in the lower three achievement quartiles but lower for the top quartile, suggesting that growth inferences may be less stable at the upper end of performance, potentially because of ceiling effects or reduced score variability.

Technical Report 1405 provided large-scale evidence for internal consistency of the easyCBM® CCSS mathematics benchmark measures for Grades K–8 using extant national data from fall and winter administrations. The scale of the dataset (more than 135,000 students in fall and roughly 148,000 in winter) supported stable estimation of reliability across grades and seasons and enabled detailed item-level checks. The report documented strong internal consistency across all grades and administrations, with alpha values spanning approximately .81 to .95 and a high median around .90. The report also included split-half indices for first-half and second-half test segments and correlations between halves, supporting the conclusion that items functioned cohesively. An additional contribution was item discrimination evidence based on upper and lower performance groups: nearly all items showed the expected pattern, with higher-performing students selecting correct responses at higher rates than lower-performing students. Technical Report 2602 extended the same analysis of internal-consistency reliability evidence for easyCBM® benchmark measures in Grades 3–8. Data included several student identifiers, for Benchmark administrations and then split into separate Fall and Winter benchmark files. Results are reported by grade and season.

Technical Report 2602 addressed internal-consistency reliability evidence for easyCBM® benchmark measures in Grades 3–8, using both Rasch and Classical Test Theory. Results support internal-consistency reliability across Grades 3–8 in Fall and Winter; Rasch marginal reliabilities were consistently above .80.

Technical Report 0908 focused on a single grade (7) progress monitoring and benchmark form development and complements classical reliability evidence with item-bank calibration and form comparability evidence. The study piloted a large item pool (912 items) with a national sample of approximately 2,800 students and analyzed item performance using a one-parameter logistic Rasch model

(Winsteps). Technical Report 0804 extended this analysis to Grades 1-8 to address alternate form reliability and reported strong inter-form correlations. The calibrated item bank supported the construction of multiple progress-monitoring forms and benchmark screeners with closely matched mean difficulties within focal-point groupings. The report also presented classical reliability evidence (including **Cronbach's alpha** for forms and **alternate-form** correlations), aligning the item-bank work with the practical need for equivalent forms in repeated measurement.

Technical Report 0916 addressed reliability while developing scalable mathematics screening measures. Cronbach's alpha was reported over various forms and within specific mathematical content domains. These findings are reported across Grades 1-8 with results are posted under Item-Test Development.

Technical Report 1312 investigated reliability for math measures used with students in Grades 6-8. Three types of reliability were considered: internal consistency, test-retest, and generalizability. The results confirmed sufficient reliability in the forms across all grades.

Finally, Technical Report 1804 documented the reliability of the slope for math measures used with students in grades K-8. The findings varied in the slope for each subset of math domains and their correlation as well as the student growth slopes (true score variance/total variance).

Across the technical reports, several cross-cutting themes emerge. First, reliability is strongest for full-length total scores and decreases when scores are subdivided into short subscales, underscoring the importance of prioritizing total scores for screening decisions. Second, the combination of standardized online administration and large samples supports stable reliability estimation and reduces procedural sources of error, but reliability still depends on how well items target the intended achievement range for a grade and season. Third, progress monitoring places special emphasis on form equivalence and growth sensitivity; therefore, evidence about alternate-form reliability and growth (slope) reliability adds meaning beyond a single alpha coefficient. Taken together, the set of reports summarized here provides converging evidence that easyCBM® mathematics measures can yield dependable total scores for benchmark screening and can support progress monitoring when forms are carefully constructed and monitored. As with any assessment system, reliability should be re-evaluated when measures are updated, used in new contexts, or applied to decisions beyond their original design.

Summary of Technical Report 0915: Internal Consistency of General Outcome Measures in Grades 1 – 8 (Anderson et al., 2009).

Technical Report 0915 examined the internal consistency of **mathematics general outcome measures** used for screening and progress monitoring across grades 1 through 8. The primary purpose of the study was to determine whether the grade-level mathematics screeners produced reliable total scores suitable for district-level decision-making.

Methods

Participants included large, grade-specific samples drawn from a mid-sized school district in Oregon. Sample sizes ranged from approximately 1,100 to over 1,350 students per grade, with representation across general education, special education, economically disadvantaged, and historically low-achieving populations. Demographic data were collected during the spring of 2009 and used as an approximation of student characteristics during fall 2009 testing. The mathematics screeners were computer-based assessments administered in group settings during the fall of the 2009 school year. Each grade-level test consisted of 48 dichotomously scored items aligned with National Council of Teachers of Mathematics (NCTM) Curriculum Focal Point Standards. Items were distributed evenly across three focal point domains per grade. Standardized administration procedures and automated scoring were used to reduce procedural error and ensure consistency across schools. Data were prepared by coding each item for response selection, correctness, and focal point domain. Internal consistency reliability was evaluated

using Cronbach's alpha for total scores and for subtests corresponding to individual focal point domains. Descriptive statistics, including item means, variances, inter-item correlations, standard deviations, and standard errors of measurement, were calculated for each grade level.

Results

Across grades 1 through 8, total test reliability was consistently strong. Cronbach's alpha values for the full 48-item tests ranged from .80 to .87, indicating adequate internal consistency for screening purposes. Item difficulty values were generally centered near the midpoint, with sufficient variability to differentiate student performance. Inter-item correlations were modest but positive, reflecting broad coverage of mathematical content. Subtest reliabilities were lower than total test reliabilities, largely due to the reduced number of items per focal point. The findings support the use of total scores for screening and progress monitoring, while suggesting caution when interpreting individual subtest scores.

Table 1. Example of Overall Statistics from Technical Report 0915

Table 4
Overall Statistics.

Grade	Cronbach's Alpha	SD	SEM
1	0.82	6.83	2.90
2	0.86	8.53	3.19
3	0.80	6.29	2.81
4	0.86	7.13	2.67
5	0.85	7.04	2.73
6	0.87	7.34	2.65
7	0.86	7.88	2.95
8	0.83	7.40	3.05

Table 2. Sample Inter-Item Correlations from Technical Report 0915

Table 3
Inter-Item Correlations.

Grade 1	Count	Mean	Min	Max	Cronbach's alpha
Number & operations	16	.12	-.11	.37	.69
Geometry	16	.01	-.12	.38	.64
Number & operations and algebra	16	.10	-.06	.41	.64
Total	48	.08	-.15	.41	.82
Grade 2					
Number & operations	16	.08	-.17	.41	.58
Geometry	16	.11	-.13	.51	.67
Number & operations and algebra	16	.09	-.52	.71	.61
Total	48	.12	-.72	.86	.86
Grade 3					
Number & operations	16	.09	-.06	.33	.60
Geometry	16	.07	-.05	.35	.56
Number & operations and algebra	16	.12	-.02	.68	.70
Total	48	.08	-.10	.69	.80
Grade 4					
Number & operations	16	.14	-.09	.62	.72
Measurement	16	.09	-.06	.70	.61
Number & operations and algebra	16	.13	-.02	.36	.70
Total	48	.11	-.08	.70	.86
Grade 5					
Number & operations	16	.14	-.03	.35	.72
Geometry, measurement, & algebra	16	.07	-.08	.65	.55
Number & operations and algebra	16	.17	-.02	.38	.76
Total	48	.11	-.07	.66	.85
Grade 6					
Number & operations	16	.11	-.03	.27	.66
Algebra	16	.17	-.01	.50	.77
Number & operations and algebra	16	.14	.01	.43	.71
Total	48	.12	-.07	.50	.87
Grade 7					
Number & operations	16	.22	.08	.42	.82
Geometry	16	.07	-.08	.24	.54
Number & operations and algebra	16	.13	-.12	.52	.70
Total	48	.11	-.12	.52	.86
Grade 8					
Number & operations	16	.07	-.07	.34	.56
Geometry	16	.10	-.03	.27	.65
Number & operations and algebra	16	.14	-.01	.37	.73
Total	48	.09	-.09	.37	.83

Note. Cronbach's alpha scores based on standardized item.

Table 3. Summary of Key Findings from Technical Report 0915

Category	Summary
Sample	Large district-wide samples from a mid-sized Oregon school district. Sample sizes ranged from approximately 1,115 to 1,359 students per grade (Grades 1–8), with representation of special education, economically disadvantaged, and historically low-achieving students.
Assessment Forms	Computer-based mathematics general outcome measures aligned to NCTM Curriculum Focal Point Standards. Each grade-level test contained 48 items total, with 16 items per focal point domain.
Analysis Method	Classical test theory methods were used, with internal consistency evaluated primarily using Cronbach's alpha. Inter-item correlations, score variance, and standard error of measurement (SEM) were also calculated.
Items Analyzed	A total of 48 dichotomously scored items per grade-level assessment, spanning three focal point domains (e.g., number and operations, geometry, algebra, measurement, or data analysis, depending on grade).
Problematic Items	No individual items were flagged as severely problematic. However, shorter subtests (individual focal point domains) showed reduced reliability compared to the full 48-item composite, reflecting fewer items rather than poor item quality.
Item Fit	Item-level descriptive statistics indicated appropriate difficulty ranges and variability across grades. Inter-item correlations were generally low to moderate, consistent with broad-domain outcome measures.
Overall Conclusion	Results demonstrated adequate to strong internal consistency for the full mathematics screener across Grades 1–8 (Cronbach's alpha ≈ .80–.87). The measures are appropriate for screening purposes, though caution is advised when interpreting subtest scores due to lower reliability associated with shorter item sets.

Reference:

Anderson, D., Tindal, G., & Alonzo, J. (2009). *Internal consistency of general outcome measures in grades 1–8 (Technical Report 0915).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1006: Technical Adequacy of the easyCBM® Primary-Level Mathematics Measures (Grades K–2), 2009–2010 version (Anderson et al., 2010).

This technical report evaluates the psychometric adequacy of the easyCBM® **primary-level mathematics measures** for Grades K–2 during the 2009–2010 academic year. The study focused on reliability, validity, and growth sensitivity to support use of the assessments for benchmark screening and progress monitoring within Response to Intervention frameworks.

Methods

Participants for the reliability analyses included students from three school districts in the Pacific Northwest. The kindergarten sample consisted of 3,511 students, Grade 1 included 3,785 students, and Grade 2 included 3,675 students. No additional demographic variables were available. Assessment data were collected using the online easyCBM® mathematics system. Cronbach's alpha and split-half estimates were computed for each grade and season. Slope reliability was estimated using two-level hierarchical linear models (HLM), partitioning variance into student-level growth and measurement error components, with analyses stratified by grade, ethnicity, and fall score quartile.

Results

Reliability of the benchmark measures was examined using Cronbach's alpha and split-half estimates. Across grades K–2, internal consistency was strong, with Cronbach's alpha values ranging from .78 to .89 across seasons. Internal consistency was strong across grades and seasons. Split-half reliability coefficients were generally moderate, reflecting the reduced number of items used in each split. Cronbach's alpha ranged from .78 to .89 across grades and seasons, while split-half estimates were moderate, generally in the .50–.80 range. Growth reliability was examined using two-level hierarchical linear modeling, with time nested within students. Slope reliability from the HLM analyses was adequate for the lower three performance quartiles (ranging approximately .42–.77) but consistently low for the top quartile (approximately .19–.24), suggesting less stable growth estimation for higher-performing students. Therefore, slope reliability was adequate for students in the lower three achievement quartiles but consistently lower for students in the top quartile.

Table 4. Illustrative Reliability Indices from Technical Report 1006

Table 2
Case Processing Summary

		N	%
Cases	Valid	802	22.8
	Excluded[a]	2709	77.2
	Total	3511	100.0

a. Listwise deletion based on all variables in the procedure.

Table 3
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.825	.830	45

Table 4
Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	.620	.253	.929	.676	3.670	.030	45
Item Variances	.206	.066	.250	.184	3.786	.003	45
Inter-Item Covariances	.020	-.011	.114	.125	-10.625	.000	45

Table 5
Scale Statistics

Mean	Variance	Std. Deviation	N of Items
27.92	47.918	6.922	45

Table 5. Additional Sample Reliability Indices from Technical Report 1006

Split-half Reliability: Fall

Table 6
Case Processing Summary

		N	%
Cases	Valid	802	22.8
	Excluded[a]	2709	77.2
	Total	3511	100.0

a. Listwise deletion based on all variables in the procedure.

Table 7
Reliability Statistics

Cronbach's Alpha	Part 1	Value	.658
		N of Items	23[a]
	Part 2	Value	.753
		N of Items	22[b]
	Total N of Items		45
Correlation Between Forms			.667
Spearman-Brown Coefficient	Equal Length		.800
	Unequal Length		.800
Guttman Split-Half Coefficient			.798

Table 6. Sample Item Statistics from Technical Report 1006

Table 8
Summary Item Statistics

		Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	Part 1	.610	.277	.915	.638	3.306	.029	23[a]
	Part 2	.631	.253	.929	.676	3.670	.033	22[b]
	Both Parts	.620	.253	.929	.676	3.670	.030	45
Item Variances	Part 1	.211	.078	.250	.172	3.220	.003	23[a]
	Part 2	.201	.066	.250	.184	3.786	.003	22[b]
	Both Parts	.206	.066	.250	.184	3.786	.003	45
Inter-Item Covariances	Part 1	.016	-.011	.054	.065	-5.031	.000	23[a]
	Part 2	.024	.000	.114	.114	-882.506	.000	22[b]
	Both Parts	.020	-.011	.114	.125	-10.625	.000	45

Table 7. Sample Scale Statistics from Technical Report 1006

Table 9
Scale Statistics

	Mean	Variance	Std. Deviation	N of Items
Part 1	14.03	13.054	3.613	23[a]
Part 2	13.88	15.738	3.967	22[b]
Both Parts	27.92	47.918	6.922	45

Figure 1. Example of Item-Measure Relations from Technical Report 1006

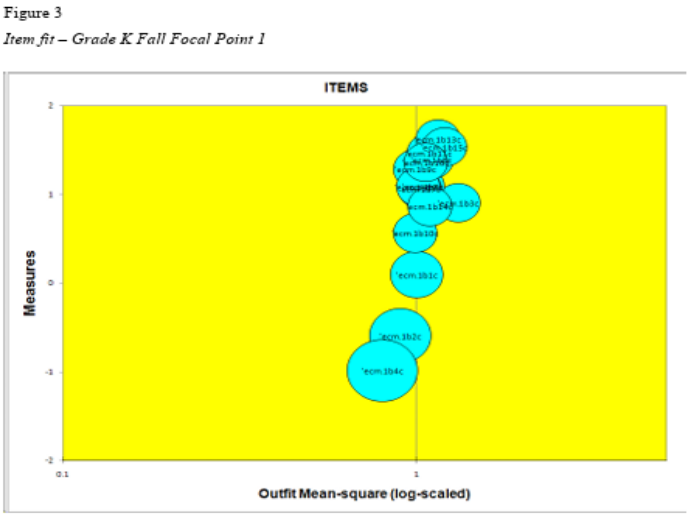


Table 8. Key Findings Summary for Technical Report 1006

Category	Summary
Participants	Over 10,900 K–2 students across regional and national samples
Reliability	Cronbach's alpha ranged from .78 to .89 across grades and seasons
Growth Analysis	Moderate slope reliability for lower three achievement quartiles
Criterion Validity	easyCBM® explained 39–66% of variance in TerraNova math scores
Construct Validity	Rasch and CFA analyses supported unidimensional measurement
Overall Conclusion	Measures demonstrated strong technical adequacy for screening and progress monitoring

Reference

Anderson, D., Lai, C.-F., Nese, J. F. T., Park, B. J., Sáez, L., Jamgochian, E., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM® primary-level mathematics measures (Grades K–2), 2009–2010 version (Technical Report 1006)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1405: Internal Consistency of the easyCBM® CCSS Math Measures (Grades K–8) (Wray et al., 2014).

This technical report examines the internal consistency and reliability of the easyCBM® **Common Core State Standards (CCSS) mathematics benchmark measures** for grades K–8. The study used extant data collected in 2013–2014 to evaluate score reliability across grade levels and seasonal administrations.

Methods

Data were drawn from fall and winter benchmark assessments completed by more than 135,000 students in the fall and approximately 148,000 students in the winter across grades K–8. Assessments were administered online through the easyCBM® platform in participating schools nationwide. Each grade-level measure consisted of multiple items aligned to CCSS mathematics standards, with total possible scores ranging from 30 points in kindergarten to 45 points in grades 6–8. Prior to analysis, cases with no item

responses and out-of-range scores were removed.

Internal consistency was evaluated using Cronbach's alpha, split-half reliability (first half and second-half forms), and correlations between test halves. Item-level analyses compared performance of students in the top and bottom 27th percentiles to assess discrimination. Reliability estimates were calculated separately for each grade and season.

Results

Results demonstrated strong internal consistency across all grades and administrations. Cronbach's alpha values ranged from .81 to .95, with a median of .90 across measures. Split-half reliability estimates showed median coefficients of .80 for the first half and .86 for the second half, with correlations between halves ranging from .52 to .73. Nearly all items performed as expected, with higher-performing students answering items correctly at higher rates than lower-performing students. These findings provide evidence that the easyCBM® CCSS math measures yield reliable scores suitable for benchmarking and progress monitoring purposes.

Table 9. Example Internal Reliability from Technical Report 1405

Table 5

Internal Reliability: CCSS Math

		Split-half Reliability		
Grade/Time	Cronbach's Alpha	1st Half	2nd Half	Correlation
K/F	.84	.68	.81	.52
K/W	.81	.70	.73	.53
1/F	.81	.65	.77	.53
1/W	.84	.72	.76	.62
2/F	.87	.77	.81	.67
2/W	.88	.78	.81	.66
3/F	.87	.71	.84	.64
3/W	.88	.73	.86	.65
4/F	.90	.81	.86	.67
4/W	.90	.82	.86	.68
5/F	.92	.79	.90	.68
5/W	.91	.80	.89	.69
6/F	.92	.80	.92	.65
6/W	.95	.86	.95	.69
7/F	.93	.82	.94	.62
7/W	.95	.87	.94	.70
8/F	.93	.83	.92	.71
8/W	.93	.83	.92	.73

Table 10. Summary of Key Findings from Technical Report 1405

Category	Summary
Participants	135,000+ fall and 148,000+ winter students (Grades K–8)
Administration	Online easyCBM® benchmark assessments
Reliability Metrics	Cronbach's alpha and split-half reliability
Internal Consistency	Median alpha = .90 across grades
Item Performance	Nearly all items discriminated effectively

Reference

Wray, K. A., Alonzo, J., & Tindal, G. (2013). *Internal consistency of the easyCBM® Common Core State Standards mathematics measures: Grades K–8 (Technical Report 1405).* 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

Technical Report 2602 covers the internal-consistency reliability evidence for easyCBM® benchmark measures in Grades 3–8, using large-scale server records from districts that permitted research use of their data. The master extract included site and student identifiers, grade, demographic indicators (e.g., disability, English learner status, race/ethnicity, gender), administration metadata (form check, start/finish, academic year), and item-level responses coded 1 (correct) and 0 (incorrect). Records were restricted to Benchmark administrations and then split into separate Fall and Winter benchmark files. Results are reported by grade and season.

Two complementary reliability approaches were used. First, a Rasch/IRT approach summarized score precision with marginal reliability based on expected a posteriori (EAP) ability estimates and their standard errors. A Rasch model was fit in R using the TAM package, and marginal reliability was computed as $1 - \text{mean}(SE^2)/\text{Var}(\theta)$. To quantify uncertainty, 500 bootstrap replications were used to produce a median reliability coefficient and a 95% confidence interval for each grade-by-season cell. Second, a classical test theory approach estimated Cronbach's alpha (KR-20 for dichotomous items) using the psych package, also summarized with a bootstrapped median and 95% interval. Descriptive statistics (n, mean, SD, min/max, median, kurtosis, skewness) accompanied reliability results to characterize score distributions.

Results: Proficient Math

Proficient Math showed strong and consistent internal-consistency reliability across Grades 3–8 in Fall and Winter benchmarks. Descriptively, students in Grades 3–5 completed 40 items and Grades 6–8 completed 45 items. Fall means were about 24–25 (Grades 3–5) and about 22–24 (Grades 6–8), with SDs near 6–8 points; winter means were higher (e.g., Grade 3 mean 28.07; Grade 6 mean 25.86), consistent with expected growth. Score ranges spanned scale, and distributions were mildly skewed.

Rasch marginal reliability medians ranged from 0.81 to 0.87. Values were at or above 0.81 in Fall and often higher in Winter (e.g., Grade 3 Winter 0.84; Grade 6 Winter 0.87; Grade 7 Winter 0.87; Grade 8 Winter 0.87). the lower bound of the 95% confidence interval around the marginal reliability exceeded 0.80 for every grade and season, supporting dependable total-score interpretations for screening and benchmarking. Cronbach's alpha converged with the IRT evidence: alphas ranged from 0.81 to 0.88 (e.g., Grade 3 Fall 0.81; Grade 4 Fall 0.87; Grade 6 Winter 0.88; Grade 7 Winter 0.88), and all alpha lower bounds met or exceeded 0.80. Overall, Proficient Math demonstrates internal consistency across Grades 3–8 in benchmark seasons.

Reference

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 0908: The Development of K–8 Progress Monitoring Measures in Mathematics for Use with the 2% and General Education Populations: Grade 7 (Lai et al., 2009).

This technical report documents the development, piloting, and psychometric evaluation of Grade 7 mathematics progress monitoring measures designed for use with both the general education population and the federally defined "2% population" of students with disabilities. The primary goal was to create universally designed, curriculum-aligned measures that are sensitive to short-term growth and appropriate for progress monitoring within Response to Intervention (RTI) frameworks. An important outcome was to also determine the **reliability** of alternate forms.

Methods

Item piloting involved approximately 2,800 Grade 7 students drawn from schools across the United States. Teachers were recruited through the easyCBM® and DIBELS websites, district partnerships, and professional networks. Data collection occurred in November and December 2008 using an online testing platform. Each student completed a 25-item test: 20 items randomly selected from the Grade 7 item pool and five fixed anchor items. Calculators were prohibited, scratch paper was permitted, and an "I don't know" option was included to reduce guessing. No identifying student or school information was collected to ensure confidentiality.

Items were aligned to National Council of Teachers of Mathematics (NCTM) Focal Point Standards and written with strong emphasis on universal design principles. Cognitive and linguistic complexity was intentionally minimized to improve accessibility for students with disabilities and English language learners while preserving alignment to grade-level content. Items were reviewed extensively by a multidisciplinary research team prior to piloting.

Item performance was analyzed using a one-parameter logistic (1PL) Rasch model implemented in Winsteps (v3.61). Key parameters examined included item difficulty (measure), standard error, and Mean Square Outfit statistics. Items with outfit values outside the recommended range (0.50–1.50) were examined further but retained when distractor analysis showed appropriate functioning. Distractor analyses confirmed that higher-ability students consistently selected correct responses while lower-ability students selected distractors.

Results

A total of 912 Grade 7 items were analyzed. Fifteen items exhibited overfit and 51 exhibited underfit, yet all were retained due to acceptable distractor functioning. The calibrated item bank supported the construction of multiple equivalent progress monitoring and benchmark forms. Thirty progress monitoring measures (10 forms per focal point grouping) and nine benchmark screeners were developed. Forms demonstrated strong comparability, with mean item difficulty values tightly clustered within focal point groupings. Measures aligned with Number and Operations, Algebra, and Geometry were the least difficult overall, while those aligned with Measurement, Geometry, and Algebra were the most challenging. Overall results support the technical adequacy of the Grade 7 measures for monitoring student progress across a broad ability range.

Table 11. Illustrative Results for Technical Report 0908

Table 4

Grade 8 Test Form Point-Biserial Correlations

Item	Form				
	6	7	8	9	10
1	.343**	.204**	.297**	.295**	.430**
2	.270**	.372**	.350**	.298**	.307**
3	.392**	.302**	.181**	.362**	.271**
4	.327**	.199**	.231**	.289**	.415**
5	.204**	.333**	.139*	.171*	.260**
6	.279**	.256**	.344**	.323**	.349**
7	0.068	.280**	.345**	.378**	0.13
8	.413**	.323**	.283**	.235**	0.101
9	.350**	0.112	.294**	.241**	.427**
10	.346**	.176*	.402**	.192**	0.117
11	.445**	.423**	.361**	.284**	.394**
12	.396**	.351**	.326**	.411**	.328**
13	.242**	.374**	.370**	0.131	.336**
14	.243**	.336**	.375**	.320**	.156*
15	.422**	0.124	.311**	.441**	.381**
16	.420**	.266**	.416**	.367**	.213**
17	.169*	.318**	.317**	.382**	.511**
18	.436**	.386**	.364**	.277**	.484**
19	0.094	.312**	.262**	.392**	.406**
20	.232**	.245**	.328**	.344**	.393**
21	.189**	.426**	.433**	.265**	.311**
22	.283**	.231**	.335**	.173*	.163*
23	.345**	.435**	.279**	.382**	.375**
24	.343**	.223**	0.137	.450**	.279**
25	.433**	.363**	.402**	.343**	.198**

Note. Items displayed in red font were removed prior to subsequent analyses.

* $p < .05$

Table 12. Example of Alternate Form Reliability Coefficients from Technical Report 0908

Table 25

Grade 6: Alternate Form Reliability Coefficients

Test form	6	7	8	9	10	n
6	-	.432	.601	.597	.465	.662
7	.376	-	.819	.641	.760	.572
8	.721	.525	-	.813	.744	.591
9	.492	.720	.426	-	.752	.522
10	.197	.784	.553	.728	-	.549
n	.806	.491	.665	.743	.569	-

Note. Coefficients below the diagonal represent correlations from the first testing occasion, while the coefficients above the diagonal represent correlations from the second testing occasion occurring one week later.

Table 13. Summary of Results for Technical Report 0908

Area	Summary
Item Functioning	Most items exhibited acceptable IRT fit and difficulty levels
Reliability	Scores demonstrated sufficient reliability for monitoring growth
Validity Evidence	Content alignment and score patterns supported validity
Population Coverage	Measure functioned across general and 2% populations

Reference

Lai, C. F., Alonzo, J., & Tindal, G. (2009). *The development of K–8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 7 (Technical Report No. 0908)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 0804: Examining Item Functioning of Math Screening Measures for Grade 1-8 Students (Liu et al., 2008).

Technical Report No. 804 describes the development and technical evaluation of mathematics computation curriculum-based measures designed for use in progress monitoring with students in Grades 3 through 8. The primary purpose of the study was to examine the reliability, comparability, and sensitivity of computation measures intended for repeated administration within a response-to-intervention framework. The measures focused on grade-appropriate number and operations skills aligned with curricular expectations. **Alternate form reliability** was addressed to ensure their use in progress monitoring.

Methods

Participants included more than one thousand students recruited from public schools across multiple grade levels. Data collection occurred during scheduled assessment windows, with students completing grade-specific computation forms under standardized testing conditions. Responses were scored using digits-correct procedures, yielding fluency-based scores commonly used in computation CBMs to support instructional decision making.

Item and form development emphasized broad coverage of grade-level computation content while minimizing construct-irrelevant variance. Multiple equivalent forms were constructed for each grade level to allow frequent reassessment without compromising score interpretability. Anchor items were embedded across forms to support equating and evaluation of form comparability.

Statistical analyses incorporated both Classical Test Theory and item response modeling approaches. Rasch (1PL) analyses were conducted to evaluate item difficulty, fit statistics, and measurement precision across grades. Complementary CTT analyses examined score distributions, **reliability coefficients**, and inter-form correlations. Items demonstrating misfit or unstable parameter estimates were reviewed and removed when appropriate.

Results

Results indicated that most items demonstrated acceptable fit to the Rasch model and contributed meaningfully to measurement precision. Inter-form correlations were strong, supporting the equivalence of alternate forms. Overall findings provide evidence that the mathematics computation measures are technically adequate and suitable for progress monitoring and instructional decision making across elementary and middle school grades.

Table 14. Illustrative Results for Technical Report 0804

Table 26.
Three Types of Problematic Items.

		Items with Incorrect Answer Keys	Items Should Be Deleted or Revised	Items Should Be Kept in spite of Noticeable Off-variable Noises
Grade 1	F	None	None	Q1, Q5, Q10, Q16, Q18
	W	None	Q29, Q30	Q2, Q3, Q26, Q27, Q28
	S	None	Q26, Q30	Q25
Grade 2	F	None	None	Q16
	W*	None	Q27	None
	S	None	None	None
Grade 3	F	None	None	Q21
	W	None	None	Q32, Q48
	S	Q 22, Q40	None	Q32, Q48
Grade 4	F	Q 21, Q32	None	Q35
	W	None	None	Q32, Q48
	S	None	None	None
Grade 5	F	None	None	None
	W	None	None	Q26
	S	None	Q13, Q41	None
Grade 6	F	Q32	None	Q28
	W	None	None	None
	S	None	None	Q25
Grade 7	F	None	None	None
	W	None	None	None
	S	Q18	None	Q22, Q36, Q37, Q48
Grade 8	F	None	Q38	Q36
	W	None	Q15	None
	S	None	None	Q43, Q49

Table 15. Key Findings Summary for Technical Report 0804

Category	Summary
Grades	Grades 3–8
Assessment Focus	Mathematics computation
Sample	Over 1,000 students across multiple grade levels
Statistical Models	CTT and 1PL Rasch
Form Equivalence	Strong inter-form correlations
Primary Outcome	Technically adequate computation CBMs for progress monitoring

Reference

Liu, K., Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2008). Examining item functioning of math screening measures for grades 1-8 students (Technical Report # 0804). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1312 Summary: Study of the Reliability of CCSS-aligned Math Measures (2012 research version): Grades 6–8 (Anderson et al., 2013).

Methods

This study examined the reliability of research-version easyCBM® mathematics measures aligned with the Common Core State Standards (CCSS) for grades 6–8. Participants were students from two middle schools in a Pacific Northwest district, totaling approximately 1,100 students across the three grade levels. The study evaluated five CCSS-aligned progress-monitoring test forms per grade along with an experimental form consisting of 25 items originally written to the National Council of Teachers of Mathematics (NCTM) focal point standards but rated as aligned with CCSS. Students completed two test forms during two testing occasions spaced one week apart, resulting in four total administrations per student within a factorial design used to control for test order effects. Tests were administered in paper-and-pencil format with standardized scripts and a 20-minute time limit per form. Analyses included point-biserial item analyses to identify poorly functioning items, Rasch modeling to compare item difficulty across CCSS and NCTM items, classical reliability statistics (Cronbach's alpha, test–retest correlations, and alternate-form correlations), and generalizability theory (G-theory) for variance to persons, items, forms, and occasions.

Results

Results from classical test theory analyses showed that internal consistency reliability coefficients for the revised 20-item CCSS forms were generally below ideal levels. Cronbach's alpha values ranged from approximately .63 to .79 in grade 6, .58 to .70 in grade 7, and .55 to .72 in grade 8. These results indicated moderate reliability for some forms but weak reliability for others. The experimental NCTM form generally demonstrated somewhat higher reliability than the CCSS forms.

Test–retest reliability estimates indicated moderate stability of scores across the one-week interval. Correlations for CCSS forms ranged from approximately .61 to .73 in grade 6, .57 to .78 in grade 7, and .52 to .66 in grade 8. Alternate-form reliability coefficients were more variable, with correlations ranging from about .20 to .82 depending on the form combination and testing occasion. These results suggested that equivalence across forms was inconsistent, likely influenced by relatively small sample sizes for each form.

Generalizability theory analyses provided additional insight into measurement error sources. In person-by-item-by-occasion models, only about 5–12% of variance was attributable to students, while a substantial proportion was associated with student-by-item interactions and residual error. Relative reliability (G-coefficients) under the study's testing conditions ranged approximately from .62 to .79 depending on grade level and test form. Decision studies indicated that reliability would increase substantially with more items or additional testing occasions. Overall, the findings suggested that the CCSS research-version test forms were somewhat more difficult and less reliable than desired, with researchers revising forms to with easier NCTM items and more difficult CCSS items in future assessments.

Table 16. Summary of Main Findings for Technical Report 1312

Evidence Area	Key Findings	Implication
Item Difficulty	CCSS items were more difficult than NCTM items across grades	Indicates higher cognitive demand of CCSS-aligned items
Internal Consistency	Cronbach's alpha ranged roughly from .55–.79 across grades	Reliability generally below desired screening standards
Test–Retest Reliability	Correlations ranged from .52–.78	Scores moderately stable over time
Alternate-Form Reliability	Wide variability (.20–.82) across forms	Forms were not consistently equivalent in difficulty
Generalizability Analyses	G-coefficients roughly .62–.79 under study conditions	Reliability improves with more items or testing occasions

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2013). *Study of the reliability of CCSS-aligned math measures (2012 research version): Grades 6–8 (Technical Report No. 1312)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1804: In-brief: Reliability of the Slope of the easyCBM® Math Measures (Nese et al., 2018).

Methods

Technical Report 1804 examined the reliability of growth slopes derived from easyCBM® mathematics progress monitoring measures for students in Grades K–8. The analytic sample included students identified by school districts as needing intensive intervention who completed easyCBM® math progress monitoring assessments between the 2014–2015 and 2016–2017 school years. Data were drawn from a larger national database of easyCBM® users. To ensure sufficient longitudinal data for estimating growth, students were included only if they had at least 10 assessment scores for a given math measure and a minimum span of 20 weeks between the first and last administration. Sample sizes varied considerably across grades and math domains due to differences in measure availability and data completeness.

The measures included CCSS Math assessments and domain-based measures such as Numbers and Operations, Geometry, Measurement, Algebra, and combinations of these domains depending on grade level. Two analytic approaches were used to estimate slope reliability. First, a Pearson split-test correlation procedure divided each student's assessment sequence into odd and even administrations, estimated ordinary least squares growth slopes for each subset, and correlated the two slopes. Second, reliability of slope was estimated using mixed-effects growth models in R (lme4 package), where reliability was defined as the ratio of true score variance in student growth slopes to the total variance of those slopes.

Results

Results showed that reliability of the slope for easyCBM® math measures varied substantially by grade, measure, and analytic method. Overall, slope reliability estimates were generally lower and more variable than those observed for many early reading measures reported in related studies. Across the elementary and middle grades, several measures demonstrated weak or unstable correlations between independently estimated growth slopes, suggesting limited consistency in growth estimates across repeated progress monitoring administrations.

In the earliest grades, results were mixed. For example, in Kindergarten the Numbers and Operations measure showed moderate slope reliability using both analytic approaches (Pearson $r \approx .62$; reliability-of-slope $\approx .51$), indicating some consistency in estimated growth trajectories. However, other Kindergarten measures such as the CCSS Math measure produced negative or near-zero split-test correlations, reflecting instability in slope estimates. Similarly, Grade 1 results generally indicated low reliability for CCSS Math and Numbers and Operations measures, although the Numbers Operations and Algebra measure demonstrated moderate slope reliability when using the mixed-effects approach ($\approx .52$).

By Grade 2, some measures displayed moderate reliability. Both CCSS Math and Numbers and Operations measures produced correlations ranging from approximately .34 to .57 depending on analytic approach. However, Measurement and Algebra-related measures showed highly variable results, largely due to small sample sizes. In Grades 3 through 5, reliability estimates were generally modest, with several measures showing correlations near zero or weak positive values. For example, Numbers and Operations reliability estimates were low in Grade 4, while certain combined domain measures in Grade 5 produced higher reliability using the mixed-effects method (e.g., Geometry–Measurement–Algebra reliability $\approx .84$).

Results in middle school grades were similarly inconsistent. In Grade 6 and Grade 7, some measures demonstrated moderate split-test correlations, but reliability-of-slope estimates were often small or unstable due to limited sample sizes. The most encouraging results appeared in Grade 8, where the CCSS Math measure demonstrated moderate reliability using the mixed-effects approach (≈ .65), with additional moderate estimates for combined domain measures involving data analysis, numbers, and algebra.

Overall, the study concluded that slope reliability for easyCBM® math progress monitoring measures varied widely and was often limited by small sample sizes and variability in the available longitudinal data. The authors note that additional research with larger samples and more controlled assessment conditions is needed to improve the precision and stability of growth estimates for mathematics progress monitoring.

Table 17. Summary of Main Findings for Technical Report 1804

Grade Range	Key Measures with Higher Slope Reliability	Interpretation
Kindergarten	Numbers and Operations	Moderate slope reliability for early math growth
Grades 1–3	CCSS Math, Numbers & Operations (varies)	Mixed and generally modest reliability estimates
Grades 4–5	Selected combined domain measures	Occasional moderate reliability but inconsistent patterns
Grades 6–7	Limited measures due to small samples	Estimates unstable and difficult to interpret
Grade 8	CCSS Math; Data/Numbers/Algebra composite	Moderate reliability using mixed-effects slope models

Reference

Nese, J. F. T., Anderson, D., Irvin, P. S., & Alonzo, J. (2018). *In-brief: Reliability of the slope of the easyCBM® math measures (Technical Report No. 1804)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Appendix A: Technical Report Table and Figure Titles

Appendix B: Guide to Spreadsheet Technical Report Value Displays

See Riverside Insights or BRT to access exact values for TR Summaries
2603_RK8M_ReliabilityMathTables.xlsx

- TR0915
- TR1006
- TR1405
- TR2602
- TR0908TD*
- TR0804TD*
- TR1312
- TR1804

*TD refers to reliability related to Test Development and therefore no spreadsheet of values.

Technical Report References

Anderson, D., Alonzo, J., & Tindal, G. (2013). *Study of the reliability of CCSS-aligned math measures (2012 research version): Grades 6-8 (Technical Report # 1312)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Lai, C. F., Nese, J. F. T., Park, B. J., Sáez, L., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM primary-level mathematics measures (Grades K-2), 2009-2010 version (Technical Report # 1006)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Tindal, G., & Alonzo, J. (2009). *Internal consistency of general outcome measures in grades 1-8 (Technical Report # 0915)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general populations: Grade 7 (Technical Report # 0908)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Liu, K., Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2008). *Examining item functioning of math screening measures for grades 1-8 students (Technical Report # 0804)*. 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® math measures (Technical Report # 1804)*. 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 CCSS math measures Grades K-8 (Technical Report # 1405)*. 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 documented across multiple validity frameworks. Rasch calibration (TR 1006, TR 1007) places item difficulty on a common logit scale across forms, supporting score comparability across administrations. Full-model regression analyses (TR 1010) account for 59–75% of Washington MSP variance, while TR 1006 explains 53–66% of TerraNova variance across K–2, providing criterion-related evidence that easyCBM® scores reflect the intended mathematics construct. CCSS alignment studies (TR 1208, TR 1228, TR 1229, TR 1230, TR 2101) confirm 93–99.6% of items align to CCSS standards through structured multi-phase expert review, ensuring consistency in item presentation and domain coverage across test events. DIF analyses (TR 1501) confirm that approximately 97% of items show negligible differential item functioning across gender, ELL, race, and special education subgroups, supporting equitable measurement. The research series spans 2009–2024, with TR 2401 using 8,000+ student records from 2023–2024 in two states, reflecting an active and ongoing validity research agenda.

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® mathematics scores are consistently articulated throughout. All TR summaries frame the measures as tools for universal screening, benchmark monitoring, and progress monitoring within RTI/MTSS frameworks and as interim assessments, with the explicit goal of documenting proficiency and identifying students at risk of failing state accountability assessments. The abstract notes these uses were examined "prior to and concurrent with implementation." Sufficient theoretical and empirical evidence supports these uses: TR 1006 (K–2, 76 schools in 26 states) and TR 1007 (Grades 3–8) establish psychometric foundations; criterion validity studies (TR 1010, TR 1011, TR 1402, TR 2401) document meaningful relationships with MSP, OAKS, SAT-10, and Smarter Balanced Math across multiple states; diagnostic efficiency studies (TR 1008, TR 1009) support screening use within RTI. TR 2401 (2024) provides the most recent evidence, with AUC values of .81–.94 across grades and seasons against Smarter Balanced, confirming continued adequacy for screening and benchmarking purposes.

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® math benchmarks explicitly supports prediction of performance on multiple external criterion measures. Construct similarity is established through CCSS alignment studies (TR 1208, TR 1228, TR 1229, TR 1230, TR 2101), confirming 93–99.6% of items align to CCSS standards, which closely mirror the content of Washington's MSP, Oregon's OAKS, and Smarter Balanced Math. Items are 45 multiple-choice per form targeting NCTM Focal Points or CCSS domains, scaled using 1PL Rasch procedures to ensure cross-form comparability. The intended use of easyCBM® as a CBM screening tool is consistent with—not contradictory to—its use for predicting state proficiency outcomes; CBM instruments are designed to be valid forecasters of end-of-year criterion performance. Although not designed to predict ACT or SAT, the measures are explicitly validated against MSP and OAKS (TR 1008–1011), TerraNova 3 (TR 1006), SAT-10 (TR 1402), and Smarter Balanced Math (TR 2401), with documented predictive evidence across each.

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. For classification predictions, ROC analyses in TR 1008, TR 1009, TR 1104, TR 1105, and TR 2401 report AUC with 95% confidence intervals, sensitivity, specificity, false positive and negative rates, and positive and negative predictive power—consistent with NCII screening guidance. For regression-based predictions, TR 1006, TR 1007, TR 1010, TR 1011, and TR 1402 report R, R^2, adjusted R^2, and model F statistics. These procedures are matched to their respective prediction types: ROC indices support binary at-risk classification; R^2 supports continuous predictive validity claims. Cross-validation studies (TR 1104 in Oregon; TR 1105 in Washington) provide direct cut score stability evidence: in Washington, AUC values ranged .82–.94 with overlapping 95% confidence intervals confirming statistically equivalent classification accuracy across independent split samples, and cut scores differed by no more than 1–2 points across groups, demonstrating reliable and replicable predictions.

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 appropriateness of predicted results. Data and procedures used to establish predictive relationships are documented and reasonable. Sample characteristics are described in each TR: TR 1010 used one Washington district, Grades 3–8 (n = 417–673 per grade); TR 1011 used Oregon districts with 1,262–1,357 students per grade; TR 1402 used approximately 65 students per grade from a Pacific Northwest middle school; TR 2401 used 8,000+ records from two states and four districts (Fall 2023–Spring 2024). Procedures are clearly articulated: regression models explain 53–66% of TerraNova variance (TR 1006), 59–75% of MSP variance (TR 1010), and 56–67% of SAT-10 variance (TR 1402); correlations range .68–.83 with MSP and .75–.82 with SAT-10. ROC analyses (TR 1009: AUC .86–.92; TR 2401: AUC .81–.94) confirm predictive accuracy. Cross-validation studies (TR 1104, TR 1105) demonstrate cut score stability across independent samples in both Oregon and Washington, directly supporting the appropriateness of predicted results.

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. Intended uses for predicted results are clearly and consistently articulated throughout the document. All TR summaries describe predictions as interim assessments as well as supporting RTI/MTSS decisions: identifying students at risk of failing state accountability assessments, supporting tier placement, and enabling timely intervention. The document notes that cut scores are grade- and season-specific and should not be generalized without validation, reflecting appropriate use boundaries. Sufficient theoretical and empirical evidence supports these uses. TR 1009 documents AUC of .86–.92 with negative predictive power above .95 for Oregon screening decisions; TR 2401 reports AUC .81–.94 and positive predictive power of .93–.98 against Smarter Balanced, confirming adequate precision for screening. TR 1501 DIF analyses (~97% negligible) support equitable use across gender, ELL, race, and special education subgroups. The convergence of criterion validity, classification accuracy, CCSS alignment, and fairness evidence across multiple states, grade levels, and a 15-year research span supports the validity of predicted results for their intended RTI/MTSS screening and monitoring purposes.

Abstract

In this series of studies on the validity of the easyCBM® Mathematics measures, the first set of reports summarizes a variety of analytical procedures: psychometric characteristics (difficulty and discrimination as well as fit), alignment with standards, reliability values, and criterion related associations with other measures using correlations along with regression and predictive modeling. The studies were conducted in several locations across the U.S., sampling from different populations, as part of the initial development of the measures, prior to and concurrent with implementation. **Note**: All tables and figures in this summary are examples of those presented in full within the individual Technical Reports: They are not exhaustive, but only illustrative.

Validity Evidence for easyCBM® Mathematics Measures

The Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014)[1] define validity as the degree to which evidence and theory support the interpretations of test scores for intended uses. Validity is not a property of a test itself, but of the inferences drawn from test scores. The Standards conceptualize validity as a unified construct supported by multiple complementary sources of evidence. Five primary sources are identified: evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing.

Evidence based on test **content** examines the extent to which assessment items represent the intended construct domain. This includes blueprint analyses, alignment to academic standards, depth-of-knowledge reviews, and expert judgment studies. In mathematics assessment, content evidence is often gathered through structured alignment studies evaluating item-to-standard correspondence and representativeness.

Evidence based on **response processes** evaluates whether examinees engage in the intended cognitive processes when responding to items. Methods such as cognitive interviews, scoring audits, and analysis of student work are used to determine whether responses reflect targeted mathematical reasoning.

Evidence based on **internal structure** evaluates dimensionality, reliability, and item functioning. Classical test theory indices (e.g., internal consistency), item response theory modeling, Rasch analysis, and differential item functioning studies provide statistical evidence regarding construct coherence and fairness across subgroups.

Evidence based on **relations to other variables** includes correlations, regression models, classification accuracy indices, and predictive validity analyses. Convergent validity is demonstrated when scores relate strongly to other measures of mathematics achievement; discriminant validity is supported when relations with unrelated constructs are weaker; predictive validity supports forecasting future outcomes.

Evidence based on **consequences** of testing addresses intended and unintended effects of score use, including decision accuracy, instructional implications, and fairness considerations. Together, these five sources provide a comprehensive framework for evaluating the validity of mathematics assessment interpretations and supporting responsible score use.

Summary of Technical Reports on Validity with easyCBM®

The series of technical reports summarized in Technical Report 2603-VK8M collectively provide comprehensive validity evidence for the easyCBM® mathematics measures across Grades K–8. Across studies, consistent methodological frameworks were applied, including classical test theory indices, Rasch modeling, alignment reviews, and criterion-related analyses linked to statewide accountability assessments. **Note**: Alignment-focused reports are summarized in Technical Report 2603-A38RM_AlignmentReadMath, documenting strong

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

correspondence between easyCBM® measures and the Common Core State Standards across grade bands. Structured expert review protocols confirmed substantial coverage of focal standards, with minor refinements identified for continuous improvement, providing evidence based on content.
Technical Reports 1006 and 1007 established foundational psychometric adequacy for primary (K–2) and upper-grade (3–8) mathematics measures. Internal consistency reliability coefficients were consistently within acceptable to strong ranges for screening and instructional decision-making. Item-level analyses demonstrated appropriate difficulty distributions and positive discrimination indices. Rasch analyses supported acceptable item fit statistics and stable parameter estimates, providing evidence of internal structure.

Criterion-related validity studies (1010, 1011, 1402) demonstrated moderate to strong correlations between easyCBM® benchmark scores and statewide mathematics assessments in Washington and Oregon. Regression and predictive modeling analyses indicated that benchmark scores contributed meaningful explanatory power in predicting proficiency outcomes. Receiver operating characteristic analyses yielded balanced sensitivity and specificity estimates, supporting risk classification decisions.

Diagnostic efficiency and cross-validation reports (1008, 1009, 1104, 1105) extended this evidence by examining cut score performance across states. Classification accuracy statistics demonstrated acceptable predictive values, supporting the consequences of score interpretation in screening contexts.

Technical Report 1501 examined differential item functioning across demographic subgroups using item response theory procedures. Results indicated minimal subgroup bias, supporting fairness and equitable measurement and Technical Report 2401 reported on classification accuracy.

Taken together, these integrated findings provide evidence across all five sources identified in the Standards (2014): content alignment, response process consistency, internal structure via reliability and Rasch modeling, relations to external variables through criterion and predictive studies, and consequential validity through classification accuracy and fairness analyses. The convergence of findings across multiple states, grade levels, and analytical methods supports the technical adequacy and validity of easyCBM® mathematics scores for screening, progress monitoring, and benchmark decision-making.

Summary of Technical Report 1002: The Alignment of easyCBM® Math Measures to Curriculum Standards (Nese, Lai, Anderson, Park, et al., 2010)

Methods

Participants included 13 certified teachers (one employed as a district curriculum specialist) recruited from Hillsboro and Springfield School Districts in Oregon. Raters were assigned to one of three grade bands (K–1, 3–5, or 6–8), and all had prior experience with easyCBM® math measures. The measures consisted of easyCBM® benchmark and progress monitoring math forms across grades K, 1, and 3–8 (grade 2 was excluded as items did not target NCTM Focal Points for that grade). Each form contained 16 items. Data were collected from November 2009 to January 2010 via online conferencing training sessions of 1.5–2 hours, followed by independent remote rating over 2–4 weeks. Raters judged item alignment to the NCTM Curriculum Focal Points on a 0–3 scale (0 = no link; 3 = clear, direct link) and assigned Depth of Knowledge (DOK) levels (1–3) to items and standards for grades 3–8. Analyses included descriptive statistics (frequency and percentage of ratings, dichotomized linked/not-linked), and inter-rater reliability was estimated using intraclass correlation coefficients (ICC) derived from a two-level hierarchical cross-classified model (HLM) that partitioned variance attributable to items, raters, and residual error.

Results

Alignment to Standards. Across grades, focal points, and test forms, teacher ratings of easyCBM® mathematics items showed generally strong alignment to the NCTM Curriculum Focal Points. Excluding one focal point in grade 8, the percentage of items rated as linked to standards ranged from 65% to 100%. At the kindergarten level,

benchmark form alignment was 88% for Numbers/Operations, 98% for Geometry, and 67% for Measurement—with Measurement being the weakest focal point. Progress monitoring forms showed a range of 25%–100%, with Measurement again the lowest (65%). Grade 1 benchmark alignment ranged from 83% to 94% across focal points, and progress monitoring forms ranged from 84% to 95%. Grades 3 through 7 generally showed strong alignment, with most focal points exceeding 75%–88% for both benchmark and progress monitoring forms. Grade 8 showed the weakest overall alignment: benchmark forms ranged from 42% (Geometry/Measurement) to 77% (Data Analysis/Numbers & Operations/Algebra), and progress monitoring forms from 58% to 73%. The Geometry/Measurement focal point at grade 8 was the single outlier below 65% across the entire study.

Across grades and focal points, teacher DOK ratings of easyCBM® items were predominantly at levels 1 (Recognition and Reproduction) and 2 (Skill and Concept). Consensus among raters on DOK item ratings within any given form ranged widely (0%–100%), though more typically between 13% and 50%. Consensus on the highest DOK level (Strategic Thinking, level 3) was rarely achieved. Geometry and Algebra standards were consistently rated at higher DOK levels than Number and Operations standards, suggesting these domains are conceptually more demanding. DOK ratings across raters showed relatively low inter-rater agreement overall, indicating that DOK classification is inherently subjective, limiting the strength of conclusions that can be drawn from these data. Inter-rater reliability for standard alignment ratings was high, with ICCs ranging from .80 to 1.0. Reliability for item-level and standard-level DOK ratings was moderately high, with ICCs ranging from .50 to .80. These results indicate that raters were dependably consistent in judging alignment to content standards, but more variable in their DOK classifications.

This study marked the first application of Webb's alignment model to a curriculum-based measurement (CBM) system aligned to modified state content standards. The authors noted that the Webb model was originally designed for large-scale summative assessments, and applying it to formative measures with extensive item banks (over 11,000 items) posed logistical and methodological challenges. Notably, the standard panel consensus process was not fully implemented, as raters worked independently rather than collectively calibrating ratings. Despite these limitations, the results provide meaningful evidence of content validity for the easyCBM® math assessment system, demonstrating that the measures are generally well-aligned to nationally recognized mathematics standards across grade levels.

Table 1. Illustrative Results from Technical Report 1002

Table 76

Grade 8 Benchmark Measures: Individual Raters' Ratings on Standard Depth of Knowledge

			Raters		
Focal point	Term	Ratings	J	K	L
Algebra	Fall	Recognition and Reproduction (1)	62.5	6.3	--
		Skill and Concept (2)	37.5	56.3	--
		Strategic Thinking (3)	0	37.5	--
		Extended Thinking (4)	0	0	--
	Winter	Recognition and Reproduction (1)	81.3	18.8	--
		Skill and Concept (2)	18.8	62.5	--
		Strategic Thinking (3)	0	18.8	--
		Extended Thinking (4)	0	0	--
	Spring	Recognition and Reproduction (1)	62.5	12.5	--
		Skill and Concept (2)	37.5	50.0	--
		Strategic Thinking (3)	0	37.5	--
		Extended Thinking (4)	0	0	--
Geometry	Fall	Recognition and Reproduction (1)	87.5	--	18.8
		Skill and Concept (2)	12.5	--	56.3
		Strategic Thinking (3)	0	--	25.0
		Extended Thinking (4)	0	--	0
	Winter	Recognition and Reproduction (1)	68.8	--	18.8
		Skill and Concept (2)	31.3	--	62.5
		Strategic Thinking (3)	0	--	18.8
		Extended Thinking (4)	0	--	0
	Spring	Recognition and Reproduction (1)	87.5	--	25.0
		Skill and Concept (2)	12.5	--	62.5
		Strategic Thinking (3)	0	--	12.5
		Extended Thinking (4)	0	--	0

Reference

Nese, J. F. T., Lai, C.-F., Anderson, D., Park, B. J., Tindal, G., & Alonzo, J. (2010). *The alignment of easyCBM® math measures to curriculum standards (Technical Report No. 1002*). Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1228: The Alignment of the easyCBM® Grades K–2 Math Measures to the Common Core Standards (Irvin et al., 2012b).

Methods

This study examined the degree to which the easyCBM® mathematics benchmark assessments for kindergarten through grade 2 align with the Common Core State Standards (CCSS). Because formative assessments are often used to guide instructional decisions in a Response to Intervention (RTI) framework, the authors emphasized that these measures must closely reflect the academic standards used to guide classroom instruction. The purpose of the study was therefore to determine whether the existing easyCBM® math items adequately represented the CCSS domains and standards that students are expected to master at each grade level.

The alignment analysis focused on the seasonal benchmark assessments included in the easyCBM® system. These assessments contain items designed to measure students' understanding of core mathematics concepts across the school year. Although the assessments were originally written to align with the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Points, the introduction of the CCSS required a systematic evaluation of how well those items corresponded with the new standards.

A panel of mathematics experts participated in the alignment study. These reviewers had expertise in mathematics instruction, standards-based curriculum, and the CCSS. Participants were recruited through an online screening process designed to confirm their qualifications and experience with standards-based mathematics education. After completing a training webinar, reviewers independently examined the easyCBM® benchmark assessment items and identified the CCSS domain and standard that each item most closely represented.

Each item was evaluated for alignment at two levels. First, reviewers determined whether an item aligned with a specific CCSS domain such as Operations and Algebraic Thinking, Number and Operations in Base Ten, or Measurement and Data. Second, reviewers identified the CCSS standard addressed by the item. Items were classified as aligning with an on-grade standard, a prior-grade standard, or not aligning with any CCSS standard. The results of these ratings were aggregated across reviewers to produce alignment summaries for kindergarten, first grade, and second grade benchmark assessments.

Results

The analysis found that the easyCBM® K–2 mathematics benchmark assessments demonstrate strong overall alignment with the Common Core State Standards. Across grades, most items aligned with either on-grade or prior-grade CCSS standards. Specifically, approximately 94% of kindergarten items, 99% of first grade items, and 96% of second grade items were judged to align with CCSS expectations. These findings suggest that the easyCBM® math assessments generally measure content that is consistent with the skills outlined in the standards.

At the domain level, alignment across grade levels was generally strong. The benchmark assessments included items representing multiple CCSS domains, providing coverage of major mathematics content areas. However, when alignment was examined at the individual standard level, several gaps became apparent. Some CCSS standards were represented by few or no assessment items, while others were represented multiple times.

For example, in grade 2 the Measurement and Data domain was broadly represented, but several standards within that domain were not addressed by any benchmark assessment items, including standards 2.MD.2, 2.MD.4, 2.MD.5, 2.MD.9, and 2.MD.10. Additional gaps were observed in the Numbers and Operations in Base Ten domain (e.g., standards 2.NBT.3, 2.NBT.6, 2.NBT.8, and 2.NBT.9) and in Operations and Algebraic Thinking (e.g., standards

2.OA.3 and 2.OA.4). Conversely, some standards such as 2.MD.1 and 2.NBT.5 appeared frequently and were therefore somewhat overrepresented within the existing item pool.

These results indicate that while the easyCBM® math measures provide broad coverage of CCSS domains, the distribution of items across specific standards is uneven. The findings were used to guide subsequent assessment development. In particular, the authors recommended writing new items to address underrepresented CCSS standards to strengthen alignment and ensure balanced coverage of mathematics content in the K–2 assessments.

Table 2. Illustrative Table of Key Findings from Technical Report 1228

Table 4
Item level alignment results for the easyCBM® first grade fall benchmark in mathematics.

Item	PS1	PS1 N	PS1 Ave N	PS1 Ave	PS2	PS2 N	PS2 Ave N	PS2 Ave	SS1	SS1 N	SS1 Ave N	SS1 Ave	SS2	SS2 N	SS2 Ave N	SS2 Ave	SS3	SS3 N	SS3 Ave N	SS3 Ave	Total n	Req Skills
1	K.CC.2	2	1	2					1.NBT.1	1		2	1.OA.1	1		2	1.OA.5	1		2	5	0
2	1.NBT.1	4	3	2					K.CC.1	1		2									5	0
3	NS	3		**					1.MD.1	1		2	1.MD.4	1		1					5	1
4	1.NBT.2	3	2	2					1.NBT.4	2		2									5	0
5	NS	4		**					1.NBT.2	1		1									5	1
6	1.NBT.3	2		1	NS	2		**	K.CC.2	1		*									5	2
7	NS	3		**					1.MD.4	2		1									5	1
8	NS	4		**					1.NBT.1	1		*									5	1
9	1.NBT.1	5	4	2																	5	0
10	1.NBT.2	5	4	1.5																	5	0
11	1.NBT.1	4	3	2					1.OA.4	1		1									5	0
12	NS	4		**					1.NBT.1	1		*									5	1
13	1.NBT.2	4	3	1.67					NS	1		**									5	1
14	1.NBT.2	4	3	2					1.NBT.4	1		2									5	0
15	1.NBT.3	4		2					1.OA.8	1		*									5	0
16	1.NBT.1	5	4	2																	5	0
17	K.G.4	5	4	1.5																	5	0
18	K.G.6	5	4	2																	5	0
19	K.G.4	5	4	1.5					K.G.2	1		*									6	0
20	K.G.4	4	3	2					1.G.4	1		2									5	0
21	1.G.2	3	2	1.5					K.G.5	1		2	K.G.6	1		2					5	0
22	K.G.4	4	3	1.67					1.G.1	1		1									5	0
23	K.G.4	3	2	1.5					1.G.1	1		1	NS	1		**					5	1
24	K.G.4	4	3	1.67					1.G.1	1		1									5	0
25	K.G.6	3		2					1.G.2	2		2									5	0
26	K.G.2	3	2	2					1.G.2	1		2	NS	1		**					5	0
27	NS	3		**					1.G.3	2		1.5									5	1

Reference

Irvin, P. S., Bitnara J. P., Alonzo, J., & Tindal, G. (2012). *The Alignment of the easyCBM® grades K–2 math measures to the Common Core Standards (Technical Report 1228).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1229: The Alignment of the easyCBM® Grades 3–5 Math Measures to the Common Core Standards (Park et al., 2012).

Methods

This study evaluated how well the easyCBM® seasonal mathematics benchmark assessments for grades 3 through 5 align with the Common Core State Standards (CCSS). The easyCBM® system includes formative benchmark assessments administered three times per year to monitor student progress. The mathematics items used in these assessments were originally developed to align with the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Points. With the adoption of the CCSS, the researchers conducted a systematic alignment study to determine how well the existing items correspond to the newer standards.

The study was conducted in two phases. In Phase 1, one educator per grade level (K–8) conducted an initial review of easyCBM® benchmark items and their alignment with CCSS standards. Reviewers were selected based on their experience teaching mathematics and their familiarity with the Common Core. The nine Phase 1 reviewers included general education teachers, special education teachers, teachers who taught both settings, and district-level mathematics specialists. Participants averaged approximately 12 years of mathematics teaching experience, with experience ranging from 3 to 31 years. The reviewers represented multiple states including Washington, Ohio, South Carolina, New Jersey, Indiana, Kansas, and Arizona.

Phase 2 expanded the review process by adding four additional educators per grade level. These reviewers were again selected for their expertise in mathematics instruction and familiarity with CCSS. Each reviewer examined grade-level benchmark assessment items and identified the CCSS domain and specific standard that best aligned with each item.

The alignment analysis included all seasonal benchmark mathematics items for grades 3–5. Each grade-level benchmark assessment contains three testing periods (fall, winter, and spring), with 45 items per assessment. This resulted in 135 items per grade being evaluated for alignment. Reviewers determined whether each item aligned with an on-grade CCSS standard, a prior-grade standard, or did not align with any standard. These classifications were aggregated to examine alignment patterns across grade levels, domains, and individual standards.

Results

Overall, the analysis found strong alignment between easyCBM® benchmark mathematics items and the CCSS across grades 3–5. Approximately 98% of third-grade items, 100% of fourth-grade items, and 97% of fifth-grade items were aligned with either grade-level or prior-grade CCSS standards. These results indicate that the benchmark assessments largely measure mathematical content consistent with the Common Core standards.

Although overall alignment was high, the analysis revealed uneven representation across domains and individual standards. At the domain level, the assessments generally covered the major CCSS content areas. However, certain domains and standards were either underrepresented or overrepresented within the item pool.

For example, in grade 5 the Number and Operations in Base Ten domain was highly represented, with more than 70 items across the three benchmark assessments aligning with standards from that domain either as primary or secondary standards. In contrast, several standards within the Number and Operations—Fractions domain were underrepresented, including standards 5.NF.3 through 5.NF.7. Additional gaps were identified in several domains, including Operations and Algebraic Thinking (5.OA.1–5.OA.3), Number and Operations in Base Ten (5.NBT.1–5.NBT.3 and 5.NBT.5), Measurement and Data (5.MD.1–5.MD.2), and Geometry (5.G.1–5.G.4).

The results indicate that while easyCBM® benchmark assessments broadly align with CCSS expectations, the distribution of items across standards is uneven. The findings were used to guide subsequent assessment development, particularly the creation of new items targeting underrepresented standards to strengthen alignment between easyCBM® mathematics measures and the Common Core standards.

Table 3. Illustrative Table of Key Findings from Technical Report 1229

Table 1
Item level alignment results for the easyCBM® third grade fall benchmark in mathematics.

Item	PS1	PS1 N	PS1 Ave N	PS1 Ave	PS2	PS2 N	PS2 Ave N	PS2 Ave	SS1	SS1 N	SS1 Ave N	SS1 Ave	SS2	SS2 N	SS2 Ave N	SS2 Ave	SS3	SS3 N	SS3 Ave N	SS3 Ave	Total n	Req Skills
1	3.NF.1	2		2					2.G.3	1		1	3.G.2	1		1	NS	1	**			
2	3.NF.1	4	3	2					3.G.2	1		2										
3	3.NF.1	4	3	2					3.G.2	1		2										
4	3.NF.1	4	3	2					3.G.2	1		2										
5	3.NF.1	4	3	2					3.G.2	1		2										
6	3.NF.1	4	3	2					3.G.2	1		2										
7	3.NF.1	4	3	2					3.G.2	1		1										
8	3.NF.1	4	3	1.67					3.G.2	1		2										
9	3.NF.1	4	3	1.67					3.G.2	1		1										
10	3.NF.3	5	4	2																		
11	3.NF.3	4	3	2					3.NF.2	1		2										
12	3.NF.3	5	4	1.5																		
13	3.NF.3	5	4	2																		
14	3.NF.3	4	3	2					3.NF.1	1		1										
15	NS	3		**					2.G.1	1		1	3.G.2	1		1						
16	NS	3		**					2.G.1	1		1	3.G.2	1								
17	NS	3		**					2.G.1	1		1	3.G.2	1								
18	3.G.2	2		1.5	NS	2	1	**	2.G.1	1		2										
19	NS	2	1	**					2.OA.1	1		2	2.MD.1	1		1	3.MD.6	1		1		
20	2.G.1	2		1.5	NS	2	1	**	4.G.1	1		1										
21	2.G.1	2		2	2.G.1	2	1	1	4.G.2	1		2										
22	NS	3		**					2.G.1	2		1.5										
23	2.G.1	2		2					3.G.1	1		2	3.G.2	1		1	NS	1		**		
24	NS	4		**					2.G.1	1		1										
25	***								2.G.2	1		2	2.OA.1	1		2	2.OA.2	1		2	3.G.2	1
26	NS	4		**					3.G.1	1		2										
27	3.G.1	2		2					4.G.1	1		2	2.G.1	1		*	NS	1		**		

Reference

Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The Alignment of the easyCBM® grades 3–5 math measures to the Common Core Standards (Technical Report 1229)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1230: The Alignment of the easyCBM® Grades 6–8 Math Measures to the Common Core Standards (Irvin et al., 2012a).

Methods

This technical report examined the alignment between easyCBM® seasonal benchmark mathematics assessments and the Common Core State Standards (CCSS) for grades 6 through 8. The easyCBM® system provides formative benchmark and progress-monitoring assessments designed to support instructional decision-making within Response to Intervention (RTI) frameworks. Because instructional decisions depend on the degree to which assessments reflect current academic standards, the study evaluated how closely existing easyCBM® mathematics items corresponded to CCSS expectations.

The study used a structured alignment review conducted in two phases. In Phase 1, a single educator for each grade level (K–8) conducted an initial review of easyCBM® mathematics benchmark items and their correspondence to CCSS standards. Reviewers were selected based on their experience teaching mathematics and familiarity with the Common Core. Phase 2 expanded the process by adding four additional educators per grade level to review the items and confirm alignment judgments. Participants included general education teachers, special education teachers, and district mathematics specialists with substantial classroom experience. Reviewers evaluated each benchmark item and identified the CCSS domain and specific standard to which it aligned. The analysis included all seasonal benchmark assessments in fall, winter, and spring, with ~135 items per grade level evaluated for alignment.

Results

Results indicated strong overall alignment between easyCBM® mathematics benchmark items and CCSS standards for grades 6–8. Approximately 99% of sixth-grade items, 93% of seventh-grade items, and 96% of eighth-grade items aligned with either on-grade or prior-grade CCSS standards.

While the overall alignment rates were high, the analysis revealed uneven representation of individual standards across domains. Some CCSS standards were heavily represented by multiple items, whereas others were minimally represented or absent. These findings provided guidance for future item development within the easyCBM® system, particularly for creating new assessment items that address underrepresented standards and improve alignment with the Common Core.

Table 4. Illustrative Table of Key Findings from Technical Report 1230

Table 1

Item level alignment results for the easyCBM® sixth grade fall benchmark in mathematics.

Item	PS1	PS1 N	PS1 Ave N	PS1 Ave	PS2	PS2 N	PS2 Ave N	PS2 Ave	SS1	SS1 N	SS1 Ave N	SS1 Ave	SS2	SS2 N	SS2 Ave N	SS2 Ave	SS3	SS3 N	SS3 Ave N	SS3 Ave	SS4	SS4 N	SS4 Ave N	SS4 Ave	SS5	SS5 N	SS5 Ave N	SS5 Ave	Total n	Req Skills
1	NS	4		**					5.NBT.3	1		1																	5	3
2	5.NF.4	2	1	2	5.NF.6	2	1	1	NS	1		**																	5	1
3	5.NBT.7	2		2					6.NS.2	1		2	6.NS.3	1		*	NS	1		**									5	1
4	NS	3		**					5.NBT.3	1		1	5.NF.3	1		1													5	2
5	5.NF.6	3		2					5.NF.4	2	1	2																	5	0
6	5.NF.4	4	3	1.67					5.NF.6	1		2																	5	0
7	5.NF.4	2	1	2	5.NF.6	2		1	5.NF.5	1		2																	5	0
8	5.NF.4	3	2	1.5	5.NF.7	2		1.5	5.NF.7	2		1.5																	7	0
9	5.NBT.2	3		1.67					6.EE.1	2	1	2																	5	0
10	NS	3		**					5.NBT.3	1		2	5.NF.3	1		1													5	2
11	NS	3		**					5.NBT.7	1		1	6.EE.3	1		2													5	2
12	NS	4		**					5.NF.5	1		1																	5	3
13	NS	3		**					5.NBT.3	1		1	5.NF.3	1		2													5	2
14	NS	3		**					5.NBT.3	2		1																	5	2
15	5.NF.2	3	2	2					NS	2		**																	5	1
16	6.EE.2	3	2	2					6.EE.5	1		2	6.EE.6	1		1													5	0
17	6.EE.2	4	3	1.67					6.EE.5	1		2																	5	0
18	6.EE.2	2		1.5					6.EE.5	1		2	6.EE.6	1		1	NS	1		**									5	0
19	6.EE.2	2		2					6.EE.6	1		*	6.EE.7	1		2	NS	1		**									5	0
20	NS	4		**					6.RP.2	1		2																	5	2
21	6.EE.7	2	1	2	6.EE.2	2		2	6.EE.5	1		1																	5	0
22	6.EE.6	2	1	2	6.EE.2	2		1.5	6.EE.7	1		2																	5	0
23	6.EE.5	2		1					6.EE.2	1		2	6.EE.3	1		2	NS	1		**									5	0
24	6.EE.1	5	4	2					6.EE.2	1		*	6.EE.3	1		*													7	0
25	6.EE.1	5	4	2																									5	0
26	6.EE.1	2	1	2	6.EE.4	2		2	6.EE.2	1		2	6.EE.3	1		*													6	0
27	6.EE.2	3		2					6.EE.3	1		2	NS	1		**													5	0

Reference

Irvin, P. S., Bitnara, J. P., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM® grades 6–8 math measures to the Common Core Standards (Technical Report No. 1230)*. Behavioral Research and Teaching, Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1208: The Alignment of the easyCBM® Middle School Mathematics CCSS Measures to the Common Core State Standards (Anderson et al., 2012).

Methods

Overall alignment was high. Of the 1,345 items reviewed, 1,180 (87.73%) received an adjusted MFRM rating at or above 2.0, indicating alignment to their corresponding CCSS standard. Of the 165 items rated as not aligned, 160 (97.00%) were judged by consensus to address a requisite skill to the standard. Combined, 99.6% of all items aligned either directly to a standard or to a requisite skill. Rater severity varied substantially across the 15 raters, with the most severe rater scoring items nearly a full category lower on average than the most lenient. Despite this range, all raters fit the MFRM expectations well, with mean square outfit statistics ranging from 0.76 to 1.16, indicating consistent rating behavior. The exploratory analysis found minimal variation in ratings attributable to item domain (5% of variance) or grade level (0.25% of variance), suggesting raters applied the alignment criteria consistently across content areas and grades.

Results

Overall alignment was high. Of the 1,345 items reviewed, 1,180 (87.73%) received an adjusted MFRM rating at or above 2.0, indicating alignment to their corresponding CCSS standard. Of the 165 items rated as not aligned, 160 (97.00%) were judged by consensus to address a requisite skill to the standard. Combined, 99.6% of all items aligned either directly to a standard or to a requisite skill. Rater severity varied substantially across the 15 raters, with the most severe rater scoring items nearly a full category lower on average than the most lenient. Despite this range, all raters fit the MFRM expectations well, with mean square outfit statistics ranging from 0.76 to 1.16, indicating consistent rating behavior. The exploratory analysis found minimal variation in ratings attributable to item domain (5% of variance) or grade level (0.25% of variance), suggesting raters applied the alignment criteria consistently across content areas and grades.

Table 5. Illustrative Table of Key Findings from Technical Report 1208

						Fit Statistics			
Item	ScoreTot	Obs. Avg.	Adj. Avg.	Endorsability	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
6EE1003	9	3	2.91	-3.8	1.85	1	0	1	0
6EE1004	9	3	2.9	-3.64	1.85	1	0	1	0
6EE1005	9	3	2.86	-3.28	1.85	1	0	1	0
6EE1007	8	2.67	2.51	-1.93	1.06	.49	-.28	.43	-.3
6EE1011	9	3	2.9	-3.64	1.85	1	0	1	0
6EE1012	9	3	2.83	-3.12	1.85	1	0	1	0
6EE1013	9	3	2.91	-3.8	1.85	1	0	1	0
6EE1014	9	3	2.85	-3.21	1.85	1	0	1	0
6EE1017	9	3	2.83	-3.12	1.85	1	0	1	0
6EE1020	9	3	2.86	-3.28	1.85	1	0	1	0
6EE1023	9	3	2.91	-3.8	1.85	1	0	1	0
6EE1024	9	3	2.9	-3.64	1.85	1	0	1	0
6EE2002	7	2.33	2.11	-1.18	.82	.17	-1.68	.2	-1.41
6EE2003	8	2.67	2.51	-1.93	1.06	.49	-.28	.43	-.3
6EE2007	8	2.67	2.47	-1.85	1.05	.54	-.2	.48	-.24
6EE2008	9	3	2.9	-3.64	1.85	1	0	1	0
6EE2009	9	3	2.85	-3.21	1.85	1	0	1	0
6EE2012	9	3	2.91	-3.8	1.85	1	0	1	0
6EE2015	8	2.67	2.47	-1.85	1.05	.54	-.2	.48	-.24
6EE2016	8	2.67	2.54	-2.01	1.06	.68	-.01	.61	-.04
6EE2017	9	3	2.91	-3.8	1.85	1	0	1	0
6EE2019	9	3	2.9	-3.64	1.85	1	0	1	0
6EE2022	8	2.67	2.54	-2.01	1.06	.68	-.01	.61	-.04
6EE2024	9	3	2.85	-3.21	1.85	1	0	1	0
6EE3001	9	3	2.9	-3.64	1.85	1	0	1	0
6EE3003	7	2.33	2.02	-1.02	.81	1.73	1.09	1.65	.99
6EE3006	9	3	2.91	-3.8	1.85	1	0	1	0
6EE3008	9	3	2.85	-3.21	1.85	1	0	1	0
6EE3011	9	3	2.86	-3.28	1.85	1	0	1	0
6EE3012	9	3	2.83	-3.12	1.85	1	0	1	0
6EE3014	9	3	2.91	-3.8	1.85	1	0	1	0
6EE3016	9	3	2.9	-3.64	1.85	1	0	1	0
6EE3017	9	3	2.85	-3.21	1.85	1	0	1	0

Note. Endorsability reported on logit scale, with higher values indicating a "harder" to endorse item.

Reference

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The Alignment of the easyCBM® middle school mathematics CCSS measures to the Common Core State Standards (Technical Report 1208).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

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

Methods

All 50 U.S. states' K–8 Mathematics content standards served as the subject of analysis, using the Common Core State Standards (CCSS) as the primary referent. Data were drawn from individual state department of education websites, corestandards.org, easyCBM® math item development files, easyCBM® test items (lite and district versions), and the easyCBM® user manual. Both easyCBM® Basic Math (formerly NCTM Math) and easyCBM® Proficient Math (formerly CCSS Math) were included. The Proficient Math measures were developed in 2012, after the release of the CCSS, and were explicitly designed to reflect CCSS expectations across grades K–8. Mathematics standards were gathered from November 2020 through March 2021. States were classified into three groups: CCSS Adopted (20 states), CCSS Revised (28 states), and State Unique (2 states: Texas and Virginia). Standards text was input into a multi-tabbed Excel file covering all 11 CCSS mathematics domains.

Using archived math development alignment files, an alignment dataset was constructed to document item coverage across all CCSS clusters and topics. A minimum threshold required coverage on at least 3 of 3 benchmark forms. For standards with only 1–2 benchmark forms covered, follow-up analysis of progress monitoring (PM) coverage was conducted. Alignment was rated on a four-level scale: Strong (all 3 benchmark forms), Moderate (2 benchmark forms + ≥50% of PM forms), Limited (2 benchmark forms + <50% of PM forms), or Insufficient (<2 benchmark forms). A second reviewer verified alignment classifications, and disagreements were resolved through consensus discussion.

Results

The easyCBM® mathematics assessments demonstrated strong and broad alignment with state mathematics standards, particularly for states that had fully or partially adopted the CCSS. Because the Proficient Math measures were explicitly developed with CCSS alignment in mind, coverage across the 11 mathematics domains was systematically documented and generally robust.

For the 20 states that fully adopted the CCSS (e.g., Colorado, Oregon, Washington, Maryland), alignment results were uniform and applied equally across all states in this group. The Proficient Math measures showed Strong to Moderate alignment across all 11 CCSS domains, with the strongest alignment found in domains most heavily represented in the item development process, such as Operations and Algebraic Thinking, Numbers & Operations in Base 10, and Measurement & Data. The 28 CCSS Revised states showed largely consistent alignment with the CCSS-based findings, given the substantial overlap between their standards and the CCSS. State-specific ADDITIONAL standards—those extending beyond the CCSS framework—were rated separately. Most ADDITIONALS reflected off-grade associations with CCSS clusters (e.g., first-grade Counting and Cardinality content appearing in kindergarten standards) and were generally rated Limited, as easyCBM® items partially but not fully captured these extended competencies. Texas and Virginia, whose standards substantially deviate from the CCSS, required individual alignment analyses. Given the limited overlap between their frameworks and the CCSS, alignment ratings for non-CCSS standards were more frequently Limited to Insufficient. Educators in these states should consult the state-specific Google Sheets datasets to identify which easyCBM® items and domains best correspond to their standards.

Across all states, the 11 CCSS mathematics domains were reviewed: Geometry; Measurement & Data; Counting & Cardinality; Operations and Algebraic Thinking; Numbers & Operations in Base 10; Numbers & Operations–Fractions; The Number System; Ratios & Proportional Relations; Expressions & Equations; Statistics & Probability;

and Functions. Domains assessed across multiple grade levels (e.g., Geometry, Measurement & Data) showed more robust item coverage, while upper-grade domains (e.g., Functions, Statistics & Probability) showed more variability in alignment strength given the narrower grade band in which they appear.

These findings indicate that easyCBM® mathematics assessments—particularly the Proficient Math measures—provide well-aligned measurement tools for educators in CCSS-aligned states. The explicit CCSS-based development of these measures supports their use as progress monitoring and benchmark tools in states with full or modified CCSS adoption. Users in State Unique contexts should consult the detailed Google Sheets results to determine which domains and clusters offer adequate alignment support.

Reference

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.

Summary of Technical Report 1006: Technical Adequacy of the easyCBM® Primary-Level Mathematics Measures (Grades K–2), 2009–2010 Version (Anderson, Lai, et al., 2010).

Methods

This report examined the technical adequacy of the easyCBM® mathematics benchmark assessments across Kindergarten, Grade 1, and Grade 2 using data collected throughout the 2009–2010 school year. Criterion and construct validity analyses used a nationally stratified random sample drawn from 76 schools across 26 states.

For criterion validity, a national stratified random sample across 76 schools in 26 states was used. The TerraNova 3 mathematics assessment, administered in May, served as the criterion. Regression and correlation analyses were run in four models at each grade: a full model (all three seasonal scores as predictors) and three seasonal models. Fall and winter models were interpreted as predictive validity evidence, the spring model as concurrent validity evidence. Construct validity was assessed through Rasch item fit analyses and confirmatory factor analysis (CFA), testing a unidimensional model against a three-factor alternative.

Results

Criterion-related validity was evaluated using regression and correlation analyses with the TerraNova 3 mathematics assessment as the external criterion. Predictive validity was assessed using fall and winter easyCBM® scores, while concurrent validity was examined using spring scores. Full regression models were significant at all three grade levels, explaining 53% (K), 59% (Gr. 1), and 66% (Gr. 2) of TerraNova variance. Individual seasonal models were also all significant. Concurrent validity (spring model) accounted for approximately 52–53% of TerraNova variance across grades. Predictive validity (fall and winter models) explained between 27% and 54% of TerraNova variance. Rasch and CFA results supported the unidimensional structure of the assessments, with items generally fitting the model well.

Construct validity was investigated through Rasch item-fit analyses and confirmatory factor analysis. Item outfit statistics largely fell within acceptable ranges, supporting unidimensionality. CFA results further indicated that a unidimensional model fit the data as well as or better than competing three-factor models aligned with NCTM focal points. Overall, the findings provide strong evidence supporting the technical adequacy of the easyCBM® K–2 mathematics measures for educational decision-making

Table 6. Example Summary of Key Findings from Technical Report 1006

Table 82
Correlations

		TerraNova3 Math Scale Score	Fall09Tot	Wint10Tot	Sprng10Tot
Pearson Correlation	TerraNova3 Math Scale Score	1.000	.594	.509	.653
	Fall09Tot	.594	1.000	.529	.478
	Wint10Tot	.509	.529	1.000	.661
	Sprng10Tot	.653	.478	.661	1.000
Sig. (1-tailed)	TerraNova3 Math Scale Score	.	.000	.000	.000
	Fall09Tot	.000	.	.000	.000
	Wint10Tot	.000	.000	.	.000
	Sprng10Tot	.000	.000	.000	.
N	TerraNova3 Math Scale Score	153	153	153	153
	Fall09Tot	153	153	153	153
	Wint10Tot	153	153	153	153
	Sprng10Tot	153	153	153	153

Reference

Anderson, D., Cheng-Fei, L., Nese, J. F. T., Bitnara, J. P., Sáez, L., Jamgochian, E., Alonzo, J., & Tindal, G. (2010). *Technical Adequacy of the easyCBM® Primary-Level Mathematics Measures (Grades K–2), 2009–2010 Version (Technical Report 1006)* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1007: Technical Adequacy of the easyCBM® Mathematics Measures: Grades 3–8, 2009–2010 version. (Nese, Lai, Anderson, Jamgochian, et al., 2010).

Methods

This report examined the technical adequacy of the easyCBM® mathematics progress monitoring assessments for students in grades 3 through 8. The study evaluated the reliability, validity, and practical utility of the 2009–2010 version of the easyCBM® mathematics measures, which are designed to support instructional decision-making within Response to Intervention (RTI) frameworks. The system includes seasonal benchmark assessments administered in fall, winter, and spring as well as alternate forms of progress monitoring measures used throughout the academic year. The study used data collected from four school districts located in two states: three districts in Oregon and one in Washington. All students present on the testing days were included in the analyses, providing large samples across grades 3–8. Researchers conducted several statistical analyses to evaluate different aspects of the measures' technical properties. These analyses included estimation of minimum acceptable within-year growth rates, determination of minimum acceptable end-of-year benchmark performance levels, and examination of internal consistency and split-half reliability. Additional analyses evaluated the reliability of slope estimates derived from repeated progress-monitoring assessments. Validity evidence was examined through construct, concurrent, and predictive validity analyses, including correlations and regression analyses comparing easyCBM® scores with year-end state mathematics test results.

Results

Results indicated that the easyCBM® mathematics measures demonstrate strong reliability and validity across grades and samples. Internal consistency and split-half reliability estimates supported the stability of the measures, while slope reliability analyses suggested that growth estimates derived from repeated administrations were dependable indicators of student progress. Validity analyses showed moderate to strong correlations between easyCBM® scores and state mathematics assessment outcomes across grades and demographic groups. Regression analyses further demonstrated that fall and winter easyCBM® scores were strong predictors of year-end state mathematics performance, while spring scores showed strong concurrent validity with the same outcomes.

Overall, the findings support the technical adequacy of the easyCBM® mathematics measures for monitoring student progress and predicting performance on state mathematics assessments.

Table 7. Illustrative Table of Key Findings from Technical Report 1007

Oregon Predictive Validity for White Students in Grade 4, Regressing Winter easyCBM® Math Benchmark on Year-End State Math Test

Descriptive Statistics

	Mean	Std. Deviation	N
OAKSMathTot	220.13	9.918	1265
wint_tot	32.6798	6.04462	1265

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	.733[a]	.538	.537	6.745	.538	1469.772	1	1263	.000

a. Predictors: (Constant), wint_tot

Coefficients[a]

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.	95.0% Confidence Interval for B: Lower Bound	95.0% Confidence Interval for B: Upper Bound	Correlations: Zero-order	Correlations: Partial	Correlations: Part
1	(Constant)	180.804	1.043		173.336	.000	178.757	182.850			
	wint_tot	1.203	.031	.733	38.338	.000	1.142	1.265	.733	.733	.733

a. Dependent Variable: OAKSMathTot

Reference

Nese, J. F. T., Cheng-Fei, L., Anderson, D., Jamgochian E. M., Kamata, A., Sáez, L., Bitnara J. P., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM® mathematics measures: grades 3–8, 2009–2010 version (Technical Report 1007).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1008: Diagnostic Efficiency of easyCBM® Math: Washington State (Anderson, Alonzo, et al., 2010b).

Methods

This technical report examined the diagnostic efficiency of the easyCBM® mathematics benchmark assessments for predicting student performance on the Washington State mathematics assessment. The purpose of the study was to determine optimal cut scores on the easyCBM® measures that could accurately classify students as likely to meet or not meet proficiency standards on the state test. The analysis focused on students in grades 3 through 8 and evaluated how effectively seasonal easyCBM® benchmark scores predicted outcomes on the statewide assessment.

Data for the study were collected from three school districts in Washington State that had implemented district-wide Response to Intervention (RTI) models. All students present during the scheduled assessment periods were included in the analyses, and the sample included students across grades 3–8 as well as various demographic subgroups, including English language learners and students receiving special education services.

Two primary measures were used. The predictor variable was the easyCBM® mathematics benchmark assessment, a computer-administered test consisting of 45 multiple-choice items per form. The system includes 13 equivalent forms per grade level that were calibrated using a one-parameter Rasch model to ensure comparable difficulty across administrations. Three of these forms are used for seasonal benchmark screening (fall, winter, and spring), while the remaining forms are used for progress monitoring throughout the year. The outcome measure was the Washington Measures of Student Progress (MSP) mathematics assessment, the statewide accountability test used to determine student proficiency.

Receiver Operating Characteristic (ROC) curve analyses were conducted to evaluate the diagnostic accuracy of easyCBM® scores in predicting MSP proficiency outcomes. These analyses were used to estimate classification accuracy statistics and determine optimal cut scores for each grade level and benchmark period. Results were reported separately by grade and season and were also examined across demographic subgroups.

Results

The results indicated that the easyCBM® mathematics benchmark assessments demonstrated strong diagnostic efficiency for predicting student performance on the Washington state mathematics test. ROC analyses showed that easyCBM® scores were effective at distinguishing between students who ultimately met proficiency standards and those who did not. Optimal cut scores were identified for each grade and seasonal benchmark, allowing educators to classify students into risk categories with acceptable levels of sensitivity and specificity.

Overall, the findings suggest that easyCBM® mathematics benchmarks can serve as effective screening tools within RTI systems, enabling educators to identify students at risk of failing the state assessment early in the school year and to monitor their progress toward proficiency over time.

Table 8. Illustrative Table of Key Findings from Technical Report 1008

Table 4

Resulting Statistics for Each Chosen Cut Score: Full Sample

Grd	Season	Meets Score	*n*	Failure Base Rate	False Positive Rate	False Negative Rate	Sensitivity	Specificity	Positve Predictive Power	Negative Predictive Power	Overall Correct Classification	AUC
	Fall	31	594	.39	.25	.21	.79	.75	.67	.85	.77	.84
3	Winter	36	721	.33	.29	.15	.85	.71	.59	.90	.75	.87
	Spring	39	923	.36	.35	.13	.88	.65	.59	.90	.73	.88
	Fall	33	671	.37	.16	.17	.83	.84	.75	.89	.84	.90
4	Winter	36	857	.37	.20	.16	.84	.80	.71	.89	.81	.90
	Spring	39	911	.38	.25	.12	.88	.75	.69	.91	.80	.93
	Fall	33	640	.35	.19	.16	.84	.81	.71	.91	.82	.91
5	Winter	37	776	.37	.16	.13	.87	.84	.76	.91	.85	.93
	Spring	42	1042	.39	.27	.11	.89	.73	.68	.91	.79	.92
	Fall	32	829	.36	.27	.13	.87	.73	.65	.91	.78	.90
6	Winter	33	830	.36	.18	.14	.86	.82	.72	.91	.83	.92
	Spring	38	1668	.39	.22	.11	.89	.78	.73	.92	.83	.94
	Fall	29	753	.36	.18	.20	.80	.82	.72	.88	.82	.90
7	Winter	29	772	.35	.17	.20	.80	.83	.71	.89	.82	.91
	Spring	34	1590	.41	.22	.11	.89	.78	.74	.91	.82	.93
	Fall	32	513	.29	.20	.14	.86	.80	.63	.94	.82	.92
8	Winter	35	636	.35	.25	.10	.90	.75	.66	.93	.81	.92
	Spring	35	1462	.40	.23	.20	.80	.77	.70	.86	.78	.91

Note. AUC = Area Under the ROC Curve

Reference

Anderson, D., Alonzo, J., & Tindal. G. (2009). *Diagnostic efficiency of easyCBM® math: Washington state (Technical Report 1008).* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1009 Summary: Diagnostic efficiency of easyCBM® math: Oregon **(Anderson, Alonzo, et al., 2010a)**.

Methods

The study examined the diagnostic efficiency of the easyCBM® mathematics benchmark assessments for predicting student performance on the Oregon Assessment of Knowledge and Skills (OAKS) state test. Participants included students in grades 3 through 8 from three Oregon school districts. Two districts implemented a district-wide Response to Intervention (RTI) framework and administered easyCBM® benchmark assessments to all students, including English language learners and students with disabilities. A third district administered the benchmark assessments to a representative subset of classes designed to mirror district demographics.

Two primary measures were used. The predictor measure was the easyCBM® mathematics benchmark assessment, a computer-administered test consisting of 45 multiple-choice items with multiple alternate forms equated using a Rasch (1PL) model. The criterion measure was the mathematics portion of OAKS, a computer adaptive statewide assessment reported on a Rasch scale and classified into performance levels. For analysis, OAKS performance was dichotomized as meeting or not meeting expectations. Receiver Operating Characteristic (ROC) curve analyses were conducted to evaluate diagnostic efficiency. Sensitivity, specificity, predictive power, and classification accuracy were calculated for all possible cut scores across fall, winter, and spring administrations. Optimal cut scores were determined using decision rules based on Silberglitt and Hintze (2005), with emphasis on maximizing sensitivity while maintaining acceptable specificity.

Results

Results demonstrated that the easyCBM® mathematics benchmarks provided strong predictive accuracy for determining whether students would meet expectations on the Oregon state test. Across grades 3–8 and across seasonal benchmark administrations (fall, winter, and spring), the Area Under the ROC Curve (AUC) ranged from approximately .86 to .92, indicating high diagnostic accuracy. These values suggest that the easyCBM® measures were effective in distinguishing students who would meet or exceed state standards from those who would not. Optimal cut scores were established for each grade and season. These cut points balanced sensitivity and specificity while prioritizing the identification of students at risk of failing the state test. Sensitivity values were generally high, often exceeding .80 and in some cases approaching .90 or higher, meaning the measures were effective at correctly identifying students likely to meet state standards. Specificity values were also acceptable, typically above .70, indicating that most students identified as at risk were correctly classified.

The study also reported classification accuracy statistics including false positive and false negative rates, positive predictive power, negative predictive power, and overall correct classification. Overall classification accuracy ranged approximately from the low .70s to the mid .80s depending on grade and season. Negative predictive power values were particularly high, often above .95, indicating that students identified as not at risk were very likely to meet state expectations. Seasonal differences in cut scores were also observed. In lower grades, cut scores increased more noticeably across the school year. For example, in grade 3 the optimal cut score increased substantially from fall to spring, reflecting expected student learning growth. In contrast, upper grades showed smaller seasonal shifts in cut scores. This pattern may reflect slower growth rates in later grades or differences in scaling between easyCBM® and the state test.

Subgroup analyses were conducted by ethnicity and English language learner status. Although sample sizes varied, the general pattern of diagnostic efficiency remained consistent across groups. The findings therefore support the use of easyCBM® mathematics benchmarks as an early screening and progress monitoring tool within RTI frameworks, particularly for identifying students at risk of failing the state accountability assessment.

Table 9. Summary of Main Findings from Technical Report 1009

Finding	Evidence	Interpretation
Diagnostic accuracy	AUC values ranged from .86 to .92	easyCBM® strongly predicts performance on the Oregon state test
Sensitivity	Typically, .80–.93 across grades and seasons	Measures effectively identify students likely to meet standards
Specificity	Generally, above .70	Students flagged as at risk are usually correctly identified
Seasonal cut score change	Lower grades showed larger fall–spring increases	Reflects greater academic growth in early grades
Predictive power	Negative predictive power often above .95	Students identified as safe are very likely to meet expectations

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2010). *Diagnostic efficiency of easyCBM® math: Oregon (Technical Report No. 1009)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1010: easyCBM® Mathematics Criterion Related Validity Evidence: Washington State Test (Anderson, Alonzo, et al., 2010d).

Methods

This report examined the criterion-related validity of the easyCBM® mathematics benchmark assessments in Grades 3–8 using Washington State's Measures of Student Progress (MSP) as the external criterion. The study was conducted in one medium-sized Washington school district. Sample sizes ranged from approximately 417 (Grade 8) to 673 (Grade 4) students per grade, with demographic data detailed by grade including percentages of English Language Learners, students receiving special education services, sex, and ethnicity.

The easyCBM® mathematics tests consisted of 45 multiple-choice items aligned to NCTM Focal Point Standards, administered in fall, winter, and spring. The MSP, newly implemented for 2009–2010, included multiple-choice and short-answer items and was administered at the end of the school year. Because the MSP was administered only once at year-end, fall and winter easyCBM® scores were analyzed as predictive validity evidence, while the spring administration served as concurrent validity evidence.

Four regression models were run at each grade level: a full model including all three seasonal easyCBM® scores simultaneously, and three individual seasonal models. Scatterplots were also produced for each grade and season, with vertical lines marking the 20th and 50th easyCBM® percentiles and a horizontal line marking the MSP proficiency cut score, allowing visual examination of classification accuracy.

Results

Results indicated a strong relationship between easyCBM® and the MSP across all grades. The full regression model accounted for between 59% and 75% of MSP variance depending on grade. Individual seasonal models explained between 48% and 67% of MSP variance. Pearson correlations between easyCBM® seasonal scores and the MSP were consistently high, generally ranging from approximately .68 to .83 across grades and seasons, with the strongest relationships observed in Grades 6–8. Examination of the scatterplots showed that even in fall, very few students scoring below the 20th easyCBM® percentile reached the MSP proficiency level, while most students

above the 50th percentile did reach proficiency. The authors note that easyCBM® demonstrated stronger predictive accuracy for identifying students who would not reach proficiency than for those who would.

Table 10. Illustrative Table of Key Findings from Technical Report 1010

Correlations

		Washington State Assessment Scale Score	Fall09TotMath	Wint10TotMath	Spr10TotMath
Pearson Correlation	Washington State Assessment Scale Score	1.000	.682	.705	.689
	Fall09TotMath	.682	1.000	.735	.682
	Wint10TotMath	.705	.735	1.000	.710
	Spr10TotMath	.689	.682	.710	1.000
Sig. (1-tailed)	Washington State Assessment Scale Score	.	.000	.000	.000
	Fall09TotMath	.000	.	.000	.000
	Wint10TotMath	.000	.000	.	.000
	Spr10TotMath	.000	.000	.000	.
N	Washington State Assessment Scale Score	463	463	463	463
	Fall09TotMath	463	463	463	463
	Wint10TotMath	463	463	463	463
	Spr10TotMath	463	463	463	463

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2009). *easyCBM® Mathematics Criterion Related Validity Evidence: Washington State Test (Technical Report 1010)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1011: easyCBM® Mathematics Criterion Related Validity Evidence: Oregon State Test (Anderson, Alonzo, et al., 2010c).

Methods

This technical report examined the criterion-related validity of the easyCBM® mathematics benchmark assessments by evaluating their relationship with the Oregon statewide mathematics assessment. The study investigated whether easyCBM® benchmark scores could predict student performance on the state test under the revised proficiency standards introduced for the 2010–2011 school year. The analyses focused on students in grades 3 through 8 and explored the strength of the relationship between seasonal benchmark scores and outcomes on the state assessment.

Data were collected from school districts in Oregon that administered the easyCBM® mathematics benchmark assessments as part of a Response to Intervention (RTI) framework. Students completed seasonal benchmark assessments (fall, winter, and spring) through the online easyCBM® system. These computer-based assessments include multiple equivalent forms designed for progress monitoring and screening within RTI systems.he sample included students across grades 3–8. Demographic information was reported by district and grade level. For example, in one district the grade-level samples included 1,311 students in Grade 3, 1,299 in Grade 4, 1,357 in Grade 5, 1,329 in Grade 6, and 1,262 in Grade 7. Student demographic variables included English language learner status, economic disadvantage (free or reduced lunch), special education participation, gender, and ethnicity.

To examine criterion-related validity, the study used both correlation and regression analyses to evaluate the relationship between easyCBM® mathematics scores and performance on the Oregon statewide mathematics assessment. The analyses focused on determining how well benchmark scores predicted whether students would meet the state proficiency standard.

Results

Results indicated meaningful relationships between easyCBM® benchmark scores and the Oregon statewide mathematics test. Correlation analyses demonstrated moderate to strong associations between the seasonal easyCBM® scores and the state test outcomes. Regression analyses further showed that easyCBM® scores provided significant predictive information regarding student proficiency status on the state assessment.

Overall, the findings provided evidence that the easyCBM® mathematics benchmarks function as valid indicators of student performance relative to Oregon's statewide mathematics standards. The results support the use of easyCBM® benchmark assessments as screening tools within RTI systems for identifying students who may be at risk of not meeting state proficiency expectations.

Table 11. Illustrative Table of Key Findings from Technical Report 1011

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	180.471	.580		311.140	.000
	fall_tot	1.065	.019	.694	55.366	.000

a. Dependent Variable: OAKSMathTot

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2009). *easyCBM® Mathematics criterion related validity evidence: Oregon state test. Technical Report No. 1011.* Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1402 Summary: Criterion Validity Evidence for the easyCBM® CCSS Math Measures Grades 6-8 (Anderson et al., 2014).

Methods

This study investigated the criterion validity of the easyCBM® CCSS Math benchmark assessments for Grades 6–8 by examining their relationship with the Stanford Achievement Test, Tenth Edition (SAT-10).

Participants were randomly selected from one middle school in the Pacific Northwest. Researchers selected approximately 65 students per grade to ensure adequate statistical power. The final samples included 67 students in Grade 6, 63 in Grade 7, and 64 in Grade 8. Demographic information included gender, special education status, English language learner status, eligibility for free or reduced lunch, and ethnicity. All students in the school completed the winter easyCBM® CCSS Math benchmark assessment. Within one week of this benchmark administration, the randomly selected sample completed the SAT-10 mathematics test. Both assessments were administered online. The easyCBM® CCSS Math measures were developed to align with the Common Core State Standards and are used within Response to Intervention (RTI) systems for screening and progress monitoring. Each test form included multiple-choice items designed to measure mathematical skills aligned with the standards.

The SAT-10 forms administered were Intermediate 3 for Grade 6, Advanced 1 for Grade 7, and Advanced 2 for Grade 8. Each SAT-10 mathematics test contained 80 multiple-choice items, including 48 problem-solving items and 32 procedural mathematics items. Items were scored dichotomously and scaled using a Rasch model. To

evaluate the relationship between the assessments, researchers conducted bivariate correlation analyses and simple linear regression analyses. Exploratory regression models were also estimated to examine whether demographic variables contributed additional explanatory power.

Results

The results indicated strong relationships between the easyCBM® CCSS Math benchmark and the SAT-10 mathematics scores across all grade levels. Correlation analyses showed coefficients of .82 for Grade 6, .77 for Grade 7, and .75 for Grade 8, indicating strong positive associations between the two measures.

Simple linear regression analyses demonstrated that easyCBM® scores were significant predictors of SAT-10 performance for all grades ($p < .001$). The easyCBM® measure explained 67% of the variance in SAT-10 scores for Grade 6, 59% for Grade 7, and 56% for Grade 8. Regression coefficients indicated that a one-point increase in easyCBM® score corresponded to approximately a 3.5 to 4.5 point increase in SAT-10 scale score, depending on grade level. Exploratory regression models that included demographic variables produced similar results. Most demographic variables were not significant predictors of SAT-10 performance, although special education status was significant for Grade 7. Overall, the findings indicated that easyCBM® CCSS Math scores were strongly related to SAT-10 scores, providing evidence that the assessments measure similar underlying mathematics constructs.

Table 12. Illustrative Table of Key Findings from Technical Report 1402

Table 3

Multiple Regression Results: Grade 6

Parameter	Estimate		95% CI		Sr^2
	Standardized	Raw	Lower	Upper	
Intercept	-	564. 53	537.81	591.25	-
FRL	-0.07	-5.34	-17.34	6.66	0.004
Female	0.14	10.46	-0.89	21.80	0.017
ELL	0.06	8.05	-13.42	29.53	0.003
SPED	0.02	2.67	-16.58	21.93	0.001
easyCBM©	0.82	3.98	3.18	4.79	0.494

Note. The overall model was significant, $F = 27.60$ (5, 61), $R^2 = 0.69$.

Reference

Anderson, D., Rowley, B., Alonzo, J., & Tindal, G. (2014*). Criterion validity evidence for the easyCBM® CCSS math measures: Grades 6-8 (Technical Report # 1402).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Summary of Technical Report 1104: A Cross-Validation of easyCBM® Mathematics Cut Scores in Oregon: 2009–2010 (Anderson et al., 2011a).

Methods

This study evaluated the diagnostic efficiency of the easyCBM® mathematics benchmark assessments for predicting performance on the Oregon statewide mathematics assessment. The analyses focused on determining cut scores that could accurately classify students as likely to meet or not meet the state proficiency standard. Participants were students in Grades 3 through 8 from school districts in Oregon that had implemented the easyCBM® assessment system within a Response to Intervention (RTI) framework. Students completed the easyCBM® mathematics benchmark assessments during the fall, winter, and spring screening periods. Each benchmark form consisted of 45 multiple-choice items designed to measure grade-level mathematics skills. The assessments were computer-administered and drawn from a larger pool of calibrated items developed using Rasch measurement procedures, which ensured comparable difficulty across forms.

The outcome measure was the Oregon statewide mathematics assessment, which served as the criterion measure for determining student proficiency. To evaluate the predictive accuracy of the easyCBM® benchmarks, the researchers used Receiver Operating Characteristic (ROC) curve analyses. These analyses allowed the researchers to estimate classification accuracy statistics and determine optimal benchmark cut scores. Diagnostic efficiency was evaluated through measures such as sensitivity, specificity, and overall classification accuracy across grades and benchmark seasons.

Results

The analyses showed that the easyCBM® mathematics benchmarks demonstrated strong diagnostic efficiency for predicting performance on the Oregon state mathematics assessment. ROC analyses indicated that the benchmark scores were effective at distinguishing between students who would meet proficiency standards and those who would not. Across grades and testing periods, Area Under the Curve (AUC) values were consistently high, indicating strong classification performance.

Optimal cut scores were identified for each grade and seasonal benchmark period. These cut scores allowed students to be classified into risk categories with acceptable levels of sensitivity (correctly identifying students at risk) and specificity (correctly identifying students not at risk). The results also showed that predictive accuracy generally improved across the school year, with the spring benchmark demonstrating the strongest relationship with the state assessment.

Overall, the findings indicate that easyCBM® mathematics benchmark scores provide meaningful information about students' likelihood of meeting state mathematics proficiency standards, supporting their use as screening tools within RTI systems.

Table 13. Illustrative Table of Key Findings from Technical Report 1104

Independent Samples Test (continued)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
EconDsvntg	Equal variances assumed	.279	.597	-.266	2722	.791	-.005	.019	-.043	.032
	Equal variances not assumed			-.266	2718.968	.791	-.005	.019	-.043	.032
OAKS Math Tot	Equal variances assumed	.157	.692	.300	3702	.764	.096	.322	-.535	.728
	Equal variances not assumed			.300	3700.198	.764	.096	.322	-.535	.728
Fall easyCBM	Equal variances assumed	1.077	.299	1.508	3923	.132	.308	.204	-.093	.709
	Equal variances not assumed			1.508	3922.475	.132	.308	.204	-.093	.709
Wint easyCBM	Equal variances assumed	.496	.481	-.067	2717	.947	-.017	.251	-.509	.475
	Equal variances not assumed			-.067	2713.431	.947	-.017	.251	-.509	.476
Spring easyCBM	Equal variances assumed	2.834	.092	.654	3710	.513	.126	.193	-.252	.504
	Equal variances not assumed			.654	3709.839	.513	.126	.193	-.252	.504
PLC	Equal variances assumed	.865	.352	-.465	3739	.642	-.006	.013	-.031	.019
	Equal variances not assumed			-.465	3738.656	.642	-.006	.013	-.031	.019

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2009). *A cross-validation of easyCBM® mathematics cut scores in Oregon: 2009–2010. Technical Report No. 1104*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1105: A Cross-Validation of easyCBM® Mathematics Cut Scores in Washington State: 2009–2010 Test (Anderson et al., 2011b).

Methods

The study drew on data from three Washington State school districts that had implemented district-wide Response to Intervention (RTI) programs, enrolling students in grades 3 through 8, including English language learners and students with learning disabilities. Two assessments provided the data: the easyCBM® mathematics benchmark (fall, winter, and spring administrations), and the Measures of Student Progress (MSP), Washington's statewide accountability test. easyCBM® forms were aligned to NCTM Focal Point Standards and scaled using a 1PL Rasch model; MSP scores classified students as below basic, basic, proficient, or advanced, then collapsed dichotomously into "meeting" (proficient/advanced) versus "not meeting" (below basic/basic) for analysis. The full sample was randomly split into two roughly equal groups using a Bernoulli random-value function in SPSS 18.0 ($p = 0.50$). Independent-samples t-tests verified demographic and achievement comparability across groups on ten subgroup variables. Receiver Operating Characteristic (ROC) curve analyses were then conducted at each grade for each group, with area under the curve (AUC) statistics compared via 95% confidence intervals.

Results

The random split produced two groups with closely matched demographic profiles. T-tests revealed few statistically significant differences in student subgroup composition or achievement between groups, supporting the assumption that the two samples were equivalent and that any differences in outcomes could be attributed to sampling or measurement error rather than systematic bias.

Across all grades and measurement occasions (fall, winter, and spring), the optimal cut scores derived independently for each group were strikingly similar, typically differing by no more than one to two points. In Grade 3, for example, fall benchmark meeting scores were 32 and 31 for Groups 1 and 2 respectively, converging to an identical score of 39 by spring. Grade 4 showed similarly small divergence: fall scores were 34 and 33, rising to an identical spring score of 39. Grades 5 through 8 followed comparable patterns, with most within-season differences between groups being one point or less. Grade 7 was a slight exception, with a two-point difference observed on the winter benchmark (29 vs. 31), though this was still within a narrow range. The general trajectory across all grades was an increase in the optimal meeting score from fall to spring, reflecting expected growth in mathematics achievement over the school year.

The most significant finding was that in no case did the AUC statistics differ significantly between the two randomly selected groups at any grade or measurement occasion. The AUC values themselves were consistently high across all grades, ranging from approximately 0.82 to 0.94 across grades and seasons, indicating strong overall discriminative accuracy. The overlapping 95% confidence intervals for AUC comparisons at every grade confirm that the two groups' ROC curves were statistically equivalent, lending strong validity evidence to the identified cut scores. ROC curve figures displayed visually similar curve shapes for Groups 1 and 2 at every grade level.

The study also examined whether strictly following the Silberglitt and Hintze (2005) decision rules versus the modified rules would have affected stability. In most cases, the modified rules (which prioritized maximizing sensitivity while keeping specificity above 0.70) produced stable or more stable results. The Grade 6 fall benchmark was a notable case: for Group 1, no cut score achieved both sensitivity and specificity above 0.80, and the modified rules selected a meeting score of 32; for Group 2, a score of 30 met the higher threshold. Had the unmodified rules been applied uniformly, the two groups would have been only one point apart (31 vs. 30), but the authors argued the modified rules were appropriate given the RTI framework's priority in minimizing false negatives.

Overall, these findings provide compelling cross-validation evidence that the easyCBM® mathematics cut scores recommended for use in Washington State are stable across student samples, and that the diagnostic efficiency of the instrument is consistent and replicable.

Table 14. Illustrative Table of Key Findings from Technical Report 1105

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
AmerInd/ AkNative	Equal variances assumed	30.605	.000	-2.744	1930	.006	-.020	.007	-.034	-.006
	Equal variances not assumed			-2.762	1726.160	.006	-.020	.007	-.034	-.006
Asian/ PacIslnder	Equal variances assumed	.699	.403	-.418	1930	.676	-.006	.015	-.035	.023
	Equal variances not assumed			-.418	1929.682	.676	-.006	.015	-.035	.023
Black	Equal variances assumed	.452	.502	-.336	1930	.737	-.003	.010	-.023	.016
	Equal variances not assumed			-.336	1929.968	.737	-.003	.010	-.023	.016
Hispanic	Equal variances assumed	8.528	.004	1.458	1930	.145	.020	.014	-.007	.047
	Equal variances not assumed			1.456	1901.980	.146	.020	.014	-.007	.047
White	Equal variances assumed	.115	.735	.169	1930	.866	.004	.022	-.040	.048
	Equal variances not assumed			.169	1927.770	.866	.004	.022	-.040	.048
Multiethnic	Equal variances assumed	.008	.928	.045	1930	.964	.001	.014	-.027	.028
	Equal variances not assumed			.045	1927.239	.964	.001	.014	-.027	.028
Decline	Equal variances assumed	2.779	.096	.833	1930	.405	.005	.006	-.006	.016
	Equal variances not assumed			.831	1867.408	.406	.005	.006	-.006	.016
SPED	Equal variances assumed	4.101	.043	-1.011	1930	.312	-.016	.016	-.047	.015
	Equal variances not assumed			-1.012	1929.659	.312	-.016	.016	-.047	.015
Female	Equal variances assumed	3.195	.074	1.298	1930	.195	.030	.023	-.015	.074
	Equal variances not assumed			1.298	1927.373	.195	.030	.023	-.015	.074
ELL	Equal variances assumed	7.802	.005	1.394	1930	.163	.014	.010	-.006	.034
	Equal variances not assumed			1.391	1881.700	.164	.014	.010	-.006	.035

Reference

Anderson, D., Alonzo, J., & Tindal, G. (2009). *A cross-validation of easyCBM® mathematics cut scores in Washington state: 2009–2010 test (Technical Report 1105)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 1501: An Exploration of Differential Item Functioning with the easyCBM® Middle School Mathematics Tests: Grades 6–8 (Anderson et al., 2015).

Data were drawn from a large, extant dataset collected during the 2013–2014 school year across multiple states. The sample included middle school students in Grades 6–8 participating in seasonal (fall, winter, spring) easyCBM® mathematics benchmark assessments. Each test form contained 45 multiple-choice items aligned to the Common Core State Standards. Subgroups examined included gender (female/male), English language learners (ELL/non-ELL), race (non-White/White), ethnicity (Latino/non-Latino), and special education status (received/did not receive services). Data collection occurred through operational administration of benchmark assessments, with student demographic data linked to test performance. Differential item functioning (DIF) analyses were conducted using the Mantel-Haenszel (MH) procedure with iterative purification. The raw total test score served as the matching criterion, allowing comparisons of item performance across focal and reference groups at equivalent ability levels.

Statistical analyses included calculation of MH odds ratios and log transformations, with confidence intervals. Items were classified using ETS criteria into A (negligible DIF), B (moderate DIF), or C (large DIF). Purification removed flagged items from the matching criterion and re-estimated DIF iteratively. Over 2,000 DIF evaluations were conducted across forms and groups. Overall findings indicated that most items functioned equivalently across groups, supporting fairness and validity.

Results

Results showed minimal differential item functioning across grades, seasons, and subgroup comparisons. Approximately 97% of items were classified as “A” (negligible DIF), while about 3% were “B” (moderate DIF), and very few were “C” (large DIF).

Patterns were consistent across gender, ELL status, race/ethnicity, and special education status. Many test forms showed no DIF items, and others contained only one or two flagged items. Some variability appeared in specific Grade 8 winter comparisons, but overall stability remained high. DIF does not imply bias but indicates items for review. The results suggest easyCBM® mathematics measures are largely free from systematic bias and appropriate for diverse populations. Flagged items were recommended for monitoring and potential revision rather than removal.

Conclusions

- Overall DIF Rates were ~97% negligible DIF; ~3% moderate; few were large. These results indicate strong evidence of fairness.
- Subgroup Comparisons were consistent across gender, ELL, race, and SPED indicating valid across populations.
- Form-Level Results demonstrated that many forms had zero DIF items, thus reflecting stable test construction.
- Flagged Items were small in the number of B/C items implying little need to monitor and revise.

Table 15. Illustrative Table of Key Findings from Technical Report 1501

Table 5: g6fall_ell

item	DIF.Grade	Alpha	Alpha_LB	Alpha_UB	Beta	Beta_LB	Beta_UB
I6EE5014	A	1.12	1.02	1.22	0.11	0.02	0.20
I6EE5006	A	1.07	0.97	1.18	0.07	-0.03	0.17
I6EE7007	A	1.00	0.91	1.10	0.00	-0.09	0.09
I6EE1019	A	1.19	1.09	1.30	0.18	0.09	0.27
I6EE7015	A	1.03	0.94	1.12	0.03	-0.07	0.12
I6G1005	A	0.85	0.76	0.94	-0.17	-0.27	-0.07
I6G2024	A	0.96	0.88	1.04	-0.05	-0.13	0.04
I6G3010	A	1.00	0.91	1.09	-0.01	-0.09	0.08
I6G1033	A	0.89	0.81	0.98	-0.12	-0.21	-0.02
I6G4023	A	0.96	0.87	1.05	-0.05	-0.14	0.05
I6G3038	A	1.15	1.06	1.26	0.14	0.05	0.23
I6NS2017	A	0.96	0.87	1.06	-0.04	-0.14	0.06
I6NS4010	A	0.96	0.87	1.06	-0.04	-0.13	0.06
I6NS8016	A	1.12	1.02	1.23	0.11	0.02	0.21
I6NS5011	A	1.00	0.91	1.10	0.00	-0.09	0.09
I6RP1006	A	1.03	0.93	1.13	0.03	-0.07	0.12
I6RP3006	A	0.93	0.85	1.03	-0.07	-0.17	0.03
I6RP1028	A	0.90	0.83	0.99	-0.10	-0.19	-0.01
I6RP3029	A	1.03	0.95	1.13	0.03	-0.06	0.12
I6RP2035	A	0.95	0.87	1.04	-0.05	-0.14	0.04
I6SP2025	A	1.00	0.92	1.10	0.00	-0.09	0.09
I6SP2034	A	0.97	0.89	1.06	-0.03	-0.12	0.06
I6SP1001	A	0.87	0.80	0.95	-0.14	-0.23	-0.05
I6SP4012	A	0.98	0.89	1.07	-0.03	-0.12	0.06
I6SP5008	A	1.07	0.98	1.17	0.07	-0.02	0.16
60097.	A	0.82	0.72	0.93	-0.20	-0.33	-0.07
60265.	A	1.07	0.97	1.19	0.07	-0.03	0.17
60300.	A	1.12	1.01	1.23	0.11	0.01	0.21
60353.	A	1.05	0.95	1.17	0.05	-0.05	0.15
60366.	A	0.92	0.85	1.01	-0.08	-0.17	0.01
I6NS7011	A	1.05	0.94	1.17	0.05	-0.06	0.16
I6NS5015	A	0.91	0.83	1.00	-0.09	-0.19	0.00
I6NS5006	A	0.89	0.81	0.98	-0.11	-0.21	-0.02
I6RP1023	A	0.88	0.81	0.96	-0.13	-0.22	-0.04
I6EE2010	A	0.99	0.90	1.10	-0.01	-0.11	0.10
I7G3019	A	1.02	0.90	1.15	0.02	-0.10	0.14
I7RP2003	A	0.88	0.79	0.97	-0.13	-0.23	-0.03
I7RP2005	A	0.96	0.88	1.05	-0.04	-0.13	0.05
I7NS3059	A	1.03	0.94	1.12	0.03	-0.06	0.12
I7RP1047	A	1.04	0.94	1.14	0.04	-0.06	0.13
@5OA34	A	1.10	0.95	1.26	0.09	-0.05	0.23
@5NBT25	A	1.11	0.96	1.29	0.11	-0.04	0.26
@5OA11	A	1.07	0.95	1.19	0.06	-0.05	0.17
@5G412	A	0.91	0.83	0.99	-0.10	-0.19	-0.01
@5MD415	A	0.91	0.83	1.00	-0.10	-0.19	-0.01

Reference

Anderson, D., Park, S., Alonzo, J., & Tindal, G. (2015). *An exploration of differential item functioning with the easyCBM® middle school mathematics tests: Grades 6–8 (Technical Report 1501)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Summary of Technical Report 2401: Criterion Validity and Classification Accuracy of easyCBM®: Grades 3-8 (Tindal & Nese, 2024).

Methods

Technical Report 2401 evaluates criterion validity and classification accuracy for the easyCBM® Math Benchmark in Grades 3–8, using the Smarter Balanced (SB) Mathematics test as the external criterion. The study uses well over 8,000 anonymized student records from two states and four school districts. Students completed easyCBM® Math benchmarks in Fall 2023, Winter 2024, and Spring 2024, and SB Math was administered in spring. The report first presents grade-by-season descriptive statistics for easyCBM® Math and SB Math.

For predictive/concurrent validity, the authors report grade-specific correlations between easyCBM® Math and SB Math for each season (fall and winter as predictive; spring as near-concurrent). To support comparisons across grades, an easyCBM® Math composite is also computed by converting each student's Math score to a grade-level z score (subtract the grade mean and divide by the grade SD) in each season; correlations and regression plots are then presented between the Math composite and SB Math.

For classification accuracy, SB Math serves as the criterion and "risk" is defined primarily as performance at the 20th percentile. For each grade and season, the report provides ROC evidence (AUC with 95% confidence intervals) and grade/season cut scores, along with confusion-matrix counts (true/false positives and negatives). Summary tables report base rates, overall classification rates, sensitivity, specificity, false positive/negative rates, and positive/negative predictive power aligned to NCII screening guidance.

Results

Validity evidence indicates that easyCBM® Math benchmarks relate strongly to SB Math across Grades 3–8 and seasons. In the grade-level correlation tables, the Math–SB Math correlations are consistently high (generally in the upper .60s to high .80s). Specifically, the reported coefficients range from about .69 (Grade 3 fall) up to .88 (Grade 6 spring), with many values in the .74–.86 range. The composite summary further supports these relations: the easyCBM® Math composite correlates .57 (fall), .59 (winter), and .61 (spring) with SB Math, and the corresponding regression plots show clear positive linear trends, indicating that higher benchmark Math performance is associated with higher SB Math scores.

Descriptive statistics show expected seasonal growth. For example, Grade 3 mean Math scores increase from 24.52 (fall) to 28.76 (winter) to 31.82 (spring), and similar upward shifts appear across grades (with score ranges spanning much of each grade's possible scale). SB Math distributions by grade show substantial variability (SDs roughly mid-80s to ~120 scale-score points), supporting the use of continuous models and ROC analyses. Classification accuracy results for Math (risk defined at the SB 20th percentile) are generally strong. In fall, AUC values range from .81 to .92 across grades (highest in Grade 4 at .92), with specificity often very high (.79–.93) and sensitivity typically moderate (.63–.77). In winter, AUC values improve and cluster tightly between .88 and .92 across grades; sensitivity is about .76–.81 and specificity about .83–.91, yielding overall classification rates around .79–.82. In spring, AUC values are again very strong (.88–.94) with several grades at or above .92; sensitivity ranges from .71 to .91 and specificity from .81 to .93, and overall classification rates rise as high as .89 (Grade 5) and .86 (Grades 3 and 6).

Operating characteristics show the typical screening tradeoffs. Base rates for risk are modest (about .15–.24 depending on grade/season). False positive rates are usually low (often .07–.19), while false negative rates vary more (roughly .09–.37), with the larger false negative values concentrated in some upper grades or earlier seasons. Positive predictive power is consistently very high—most often .93–.98 across grades and seasons—meaning students flagged "at risk" are very likely to fall below the SB cut. Negative predictive power is lower (often about .35–.66), reflecting the combination of base rates and the sensitivity/specificity balance. Overall, TR2401 concludes that easyCBM® Math benchmarks provide credible criterion validity with SB Math and strong classification accuracy for screening when grade- and season-specific cut scores are used.

Cut scores reflect growth: for example, Grade 3 uses an easyCBM® Math cut of 23 (fall), 27 (winter), and 30 (spring) for the SB 20th-percentile criterion, while Grade 8 uses 21, 24, and 25 across the same seasons. SB Math cut scores also rise with grade (2362 in Grade 3 to 2434 in Grade 8). Overall classification rates are typically in the mid-.70s to high-.80s (about .68–.89). AUC confidence intervals are extremely tight for large within-grade samples.

Reference

Tindal, G. & Nese, J. F. T. (2024). *Criterion validity and classification accuracy of easyCBM®: Grades 3-8. (Technical Report 2401)*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Appendix A: Technical Report Table Titles

Appendix B: Guide to Spreadsheet Technical Report Value Displays

See Riverside Insights or BRT to access exact values for TR Summaries
2603-VK8M_ValidityMathTables.xlsx

- TR1002… See Technical Report 2603-AK8RM...Page 9
- TR1228…See Technical Report 2603-AK8RM...Page 13
- TR1229…See Technical Report 2603-AK8RM...Page 17
- TR1230…See Technical Report 2603-AK8RM...Page 20
- TR1208…See Technical Report 2603-AK8RM...Page 24
- TR2101…See Technical Report 2603-AK8RM...Page 6
- TR1006
- TR1007
- TR1008
- TR1009
- TR1010
- TR1011
- TR1402
- TR1104
- TR1105
- TR1501
- TR2401

Technical Report References

Anderson, D., Alonzo, J., & Tindal, G. (2010a). *Diagnostic efficiency of easyCBM math: Oregon (Technical Report # 1009)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010b). *Diagnostic efficiency of easyCBM mathematics: Washington state (Technical Report # 1008)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010c). *easyCBM mathematics criterion related validity evidence: Oregon state test (Technical Report # 1011)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010d). *easyCBM mathematics criterion related validity evidence: Washington state test (Technical Report # 1010)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2011a). *A cross-validation of easyCBM mathematics cut scores in Oregon: 2009-2010 (Technical Report # 1104)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2011b). *A cross-validation of easyCBM mathematics cut scores in Washington state: 2009-2010 Test (Technical Teport # 1105)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM middle school mathematics CCSS measures to the Common Core State Standards (Technical Report # 1208)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Lai, C. F., Nese, J. F. T., Park, B. J., Sáez, L., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM primary-level mathematics measures (Grades K-2), 2009-2010 version (Technical Report # 1006)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Park, S., Alonzo, J., & Tindal, G. (2015). *An exploration of differential item functioning with the easyCBM middle school mathematics tests: Grades 6-8 (Technical Report # 1501)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Rowley, B., Alonzo, J., & Tindal, G. (2014). *Criterion validity evidence for the easyCBM CCSS math measures: Grades 6-8 (Technical Report # 1402)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012a). *The alignment of the easyCBM Grades 6-8 math measures to the Common Core Standards (Technical Report # 1230)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012b). *The alignment of the easyCBM Grades K-2 math measures to the Common Core Standards (Technical Report # 1228)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nese, J. F. T., Lai, C. F., Anderson, D., Jamgochian, E. M., Kamata, A., Saez, L., Park, B. J., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM mathematics measures: Grades 3-8: 2009-2010 version (Technical Report # 1007)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nese, J. F. T., Lai, C. F., Anderson, D., Park, B. J., Tindal, G., & Alonzo, J. (2010). *The alignment of easyCBM math measures to curriculum standards (Technical Report #1002)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). *The alignment of the easyCBM grades 3-5 math measures to the Common Core Standards (Technical Report # 1229)*. 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.

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.

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 (e.g., at the reportable category, content strand or objective level).

This evaluation applies Criterion 2.3 (Indicators 2.3.a–2.3.d) to all sections of the MGCI_2.3 documentation and finds strong evidence that easyCBM® mathematics sub-scores support valid, reliable, and appropriate inferences about students' specific areas of strength and need.

Indicator 2.3.a: The assessment design clearly supports reporting mathematics sub-scores at meaningful levels of granularity. Sub-scores are intentionally embedded within the measure design to reflect distinct content strands (e.g., number operations, algebraic reasoning, geometry, measurement, and data). This structure enables educators to identify targeted instructional needs rather than relying solely on a global mathematics score. The documentation consistently links design features to intended interpretations of strengths and weaknesses within the mathematics domain.

Indicator 2.3.b: Substantial reliability and precision evidence is provided for reported sub-scores. Reliability estimates and classification accuracy results are documented using defensible psychometric methods appropriate for curriculum-based measures, including analyses of sensitivity, specificity, and base-rate effects. External technical reviews by the National Center on Intensive Interventions (NCII) further corroborate the adequacy of score precision for screening and progress-monitoring purposes. Collectively, the evidence supports the use of mathematics sub-scores for their intended educational decisions.

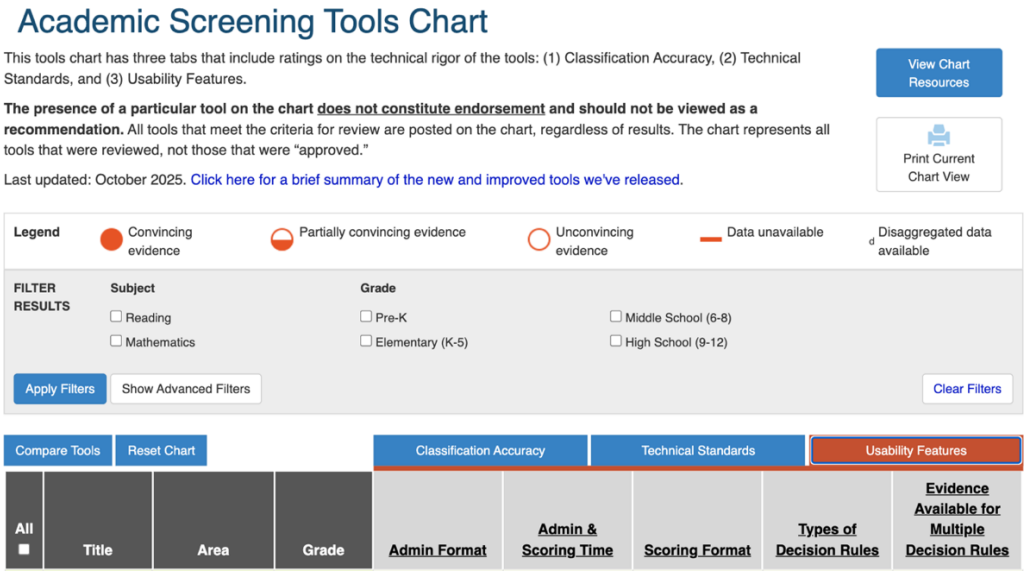
Indicator 2.3.c: Empirical and theoretical evidence supports the interpretation of sub-scores as representing meaningful sub-domains rather than arbitrary score partitions. Published research demonstrates that mathematics sub-scores differentiate student performance across content areas and are sensitive to instructional effects, particularly for low-performing students. Factor-analytic and predictive studies indicate that sub-scores capture both shared and distinct variance, justifying their separate reporting.
Under **Indicator 2.3.d**, the intended uses of sub-scores are clearly articulated and consistently bounded. Sub-scores are positioned to support screening, instructional planning, progress monitoring, and evaluation of intervention response within MTSS frameworks—not for high-stakes diagnostic decisions in isolation. A robust body of peer-reviewed research and technical documentation supports these uses.

Overall, MGCI_2.3 demonstrates strong alignment between assessment design, psychometric evidence, and intended instructional uses, supporting defensible and meaningful interpretation of mathematics sub-scores.

National Center on Intensive Interventions Results on Screening with easyCBM Reading Measures
https://intensiveintervention.org/tools-charts/overview

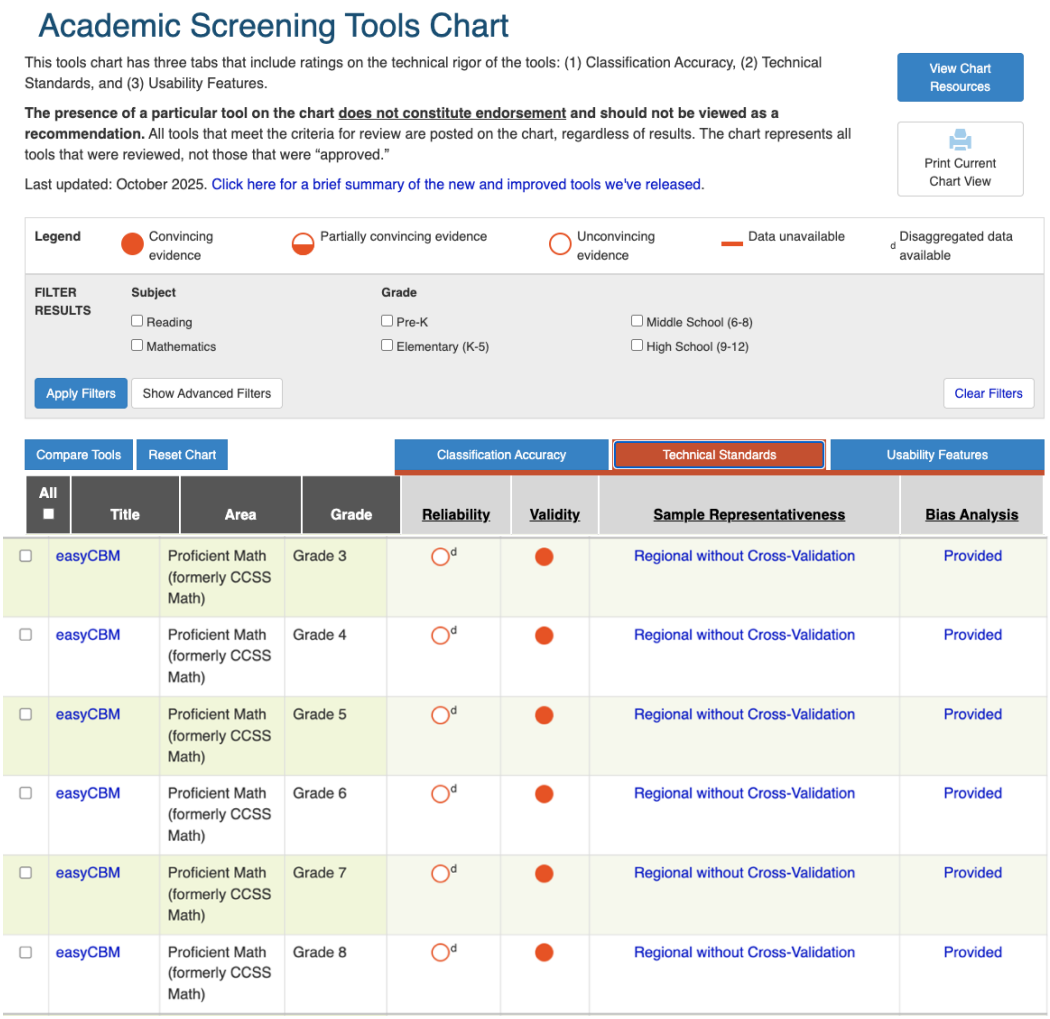
- Tools charts display expert ratings on the technical rigor of assessments and interventions.
- Products are reviewed by an external Technical Review Committee of experts.
- The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation from 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



All	Title	Area	Grade	Admin Format	Admin & Scoring Time	Scoring Format	Types of Decision Rules	Evidence Available for Multiple Decision Rules
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 3	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 4	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 5	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 6	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 7	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 8	Individual Group Computer-administered	30 minutes	Automatic	Risk Levels	No

Technical Standards



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.

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: October 2025. 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

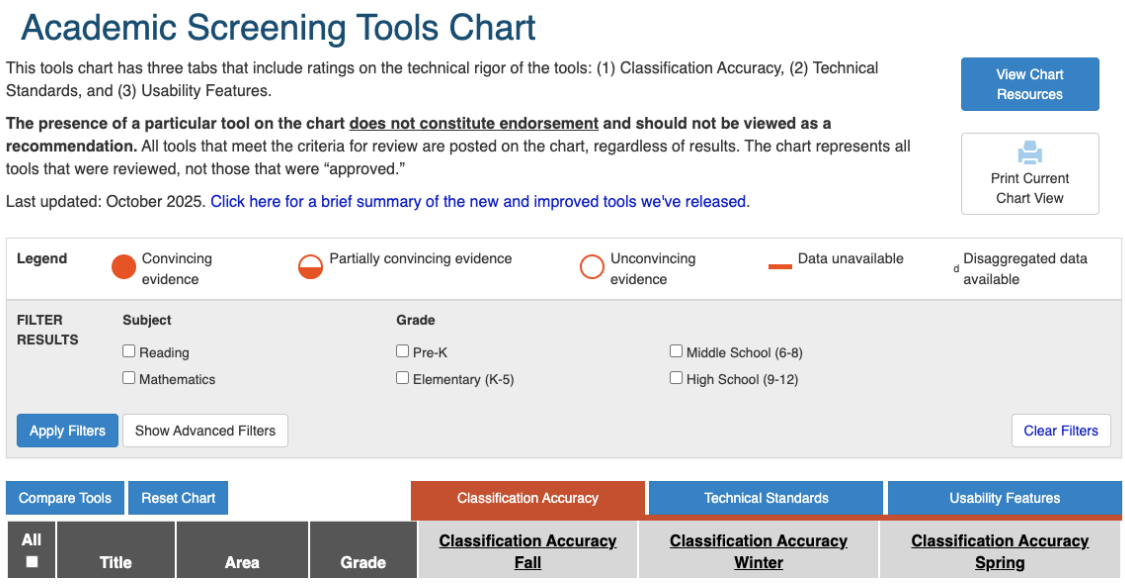
Grade: Pre-K, Elementary (K-5), Middle School (6-8), High School (9-12)

Apply Filters | Show Advanced Filters | Clear Filters

Compare Tools | Reset Chart | Classification Accuracy | Technical Standards | Usability Features

All	Title	Area	Grade	Reliability	Validity	Sample Representativeness	Bias Analysis
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 3	○d	●	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 4	○d	●	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 5	○d	●	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 6	○d	●	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 7	○d	●	Regional without Cross-Validation	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	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.

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: October 2025. 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

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	Proficient Math (formerly CCSS Math)	Grade 3	◒	●	●
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 4	◒	●	●
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 5	◒	●	●
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 6	◒	●	●
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 7	◒	◒	◒
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 8	◒	◒	●

Published Research Literature on Progress Monitoring with easyCBM Math 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.

Clarke, B., Nese, J. F. T., Alonzo, J., Smith, J. L. M., Tindal, G., Kame'enui, E. J., & Baker, S. K. (2011). Classification accuracy of easyCBM first-grade mathematics measures: Findings and implications for the field. *Assessment for Effective Intervention, 36(4)*, 243–255. https://doi.org/10.1177/1534508411414153

Abstract (≈100 words). This article evaluated how well easyCBM Grade 1 mathematics measures classify students for screening purposes. Using a year-long dataset, the authors examined decision accuracy under different cut scores and summarized the trade-offs educators face when identifying students at risk (e.g., false positives and false negatives). Results provide evidence about screening performance for Grade 1 math CBM and discuss practical implications for RTI/MTSS implementation, including how base rates and local context influence optimal decision rules. The paper frames findings in terms of improving early identification and aligning screening decisions with available instructional resources.

Anderson, D., Lai, C.-F., Alonzo, J., & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment, 16(1),* 15–34. https://doi.org/10.1080/10627197.2011.551084

Abstract (≈100 words). The study investigated a grade-level mathematics curriculum-based measure (CBM) intended to be informative for persistently low-performing students—an area where traditional measures can show floor effects. The authors examined technical features relevant to repeated use, including score distributions and the measure's ability to differentiate performance among very low achievers. Findings address whether grade-level CBM can provide interpretable information about student status and change for students far below grade expectations. The paper discusses implications for selecting progress-monitoring tools and for designing measures that remain sensitive to growth across a wide range of achievement.

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.

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 MGCI_2.4 documentation and finds strong evidence that easyCBM mathematics assessments provide valid, reliable, and appropriate information about student progress.

For **Indicator 2.4.a**, the mathematics assessment system is explicitly designed to support growth measurement. Test content and design specifications demonstrate vertical coherence within and across grades, ensuring that repeated administrations sample stable constructs over time. Both Proficient and Basic Math measures support progress monitoring, with reportable scales and benchmark structures appropriate for capturing instructional change across benchmark periods and within shorter progress-monitoring intervals.

Regarding **Indicator 2.4.b**, substantial evidence supports the reliability of student growth scores. The documentation describes appropriate procedures for estimating standard errors around growth estimates and evaluates growth reliability across the ability continuum. Empirical studies show that growth precision improves with increased measurement density and that reliability varies predictably by student performance level—findings that are explicitly acknowledged and incorporated into guidance for use. External reviews by the National Center on Intensive Interventions further corroborate the adequacy of growth score reliability for intended uses.

For **Indicator 2.4.c**, procedures for calculating growth are clearly documented and grounded in CBM theory. Growth is operationalized through trend lines, slope estimates, and benchmark-to-benchmark comparisons that are appropriate for interim assessment contexts. When potential disruptions to trend lines may occur (e.g., changes in grade-level expectations or assessment design), the documentation emphasizes empirical verification to preserve interpretive continuity. Published research confirms that observed growth corresponds meaningfully to instructional exposure and intervention effects in mathematics.

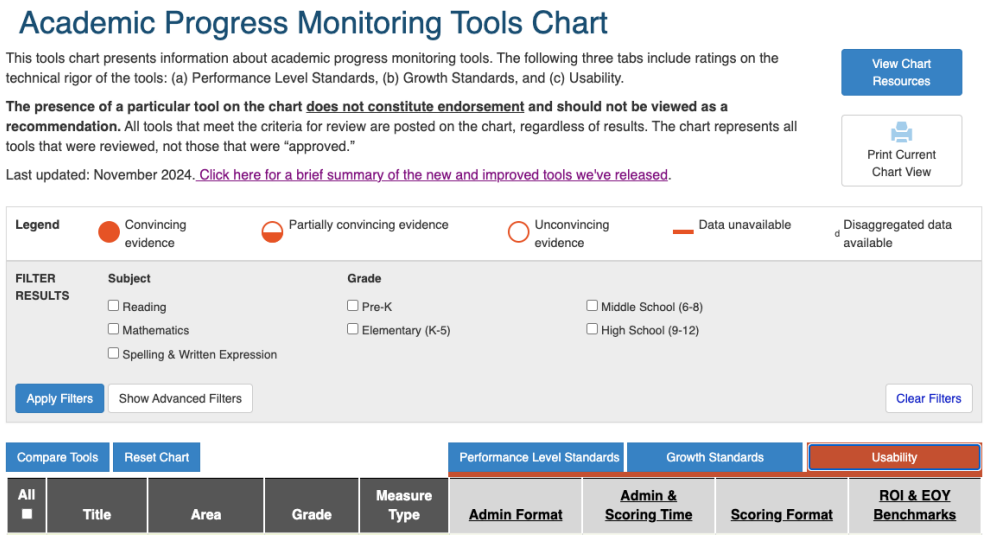
Under **Indicator 2.4.d**, the intended uses of growth scores are clearly articulated and consistently bounded. Growth scores are positioned to support progress monitoring, goal setting, instructional adjustment, and evaluation of intervention response within MTSS frameworks, rather than high-stakes decision making in isolation. Theoretical and empirical evidence strongly supports these uses across general and special education populations.

Overall, MGCI_2.4 demonstrates strong alignment between assessment design, psychometric evidence, and intended instructional uses, supporting defensible and meaningful interpretation of mathematics student progress.

National Center on Intensive Interventions Results on Screening with easyCBM Reading Measures
https://intensiveintervention.org/tools-charts/overview

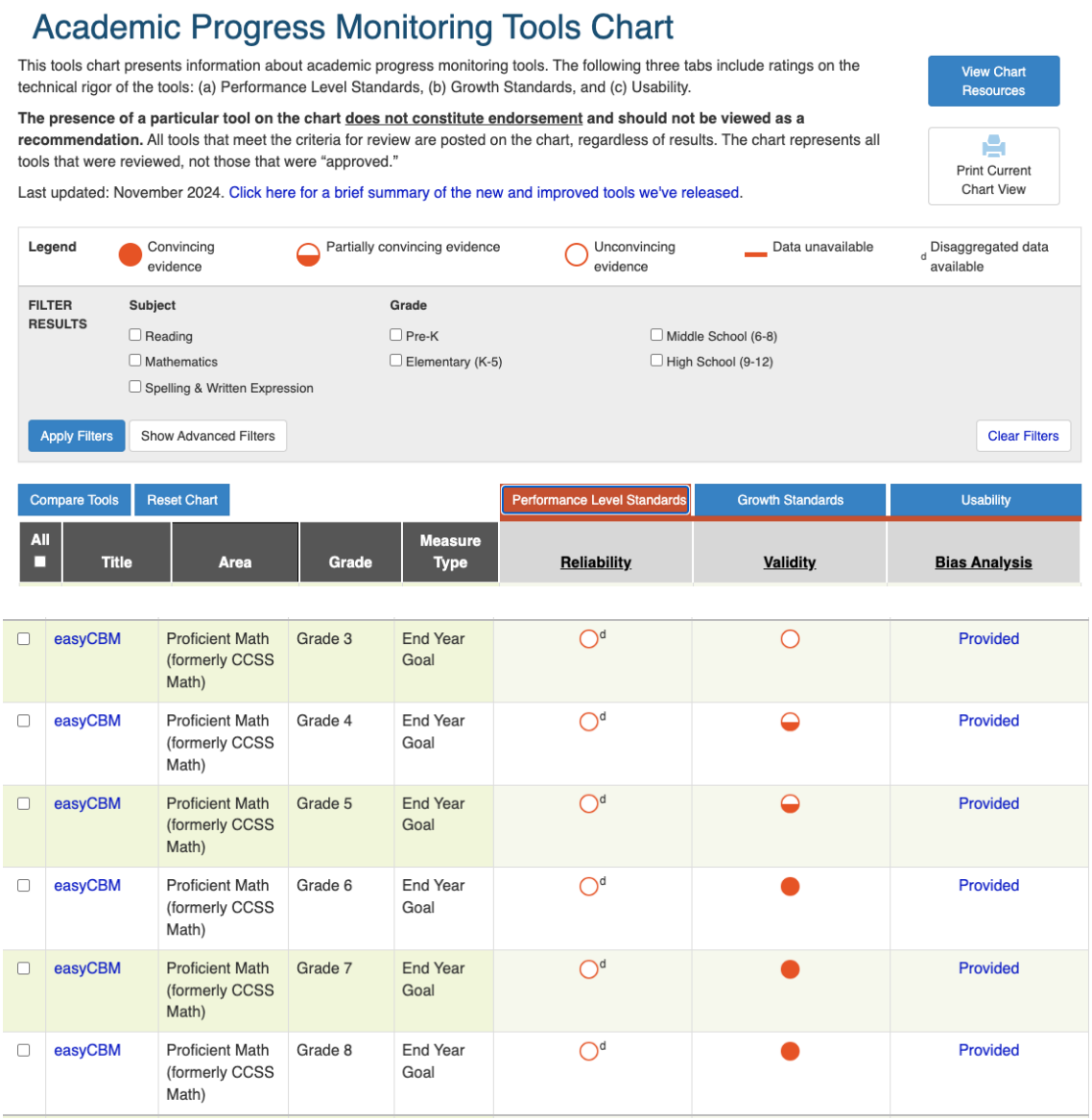
- Tools charts display expert ratings on the technical rigor of assessments and interventions.
- Products are reviewed by an external Technical Review Committee of experts.
- The presence of a particular tool on the chart does not constitute endorsement and should not be viewed as a recommendation from 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	Proficient Math (formerly CCSS Math)	Grade 3	End Year Goal	Individual Group Computer-administered Other	20 minutes	Automatic	Benchmarks Available
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 4	End Year Goal	Individual Group Computer-administered Other	20 minutes	Automatic	Benchmarks Available
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 5	End Year Goal	Individual Group Computer-administered Other	20 minutes	Automatic	Benchmarks Available
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 6	End Year Goal	Individual Group Computer-administered Other	20 minutes	Automatic	Benchmarks Available
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 7	End Year Goal	Individual Group Computer-administered Other	20 minutes	Automatic	Benchmarks Available
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 8	End Year Goal	Individual Group Computer-administered Other	20 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.

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

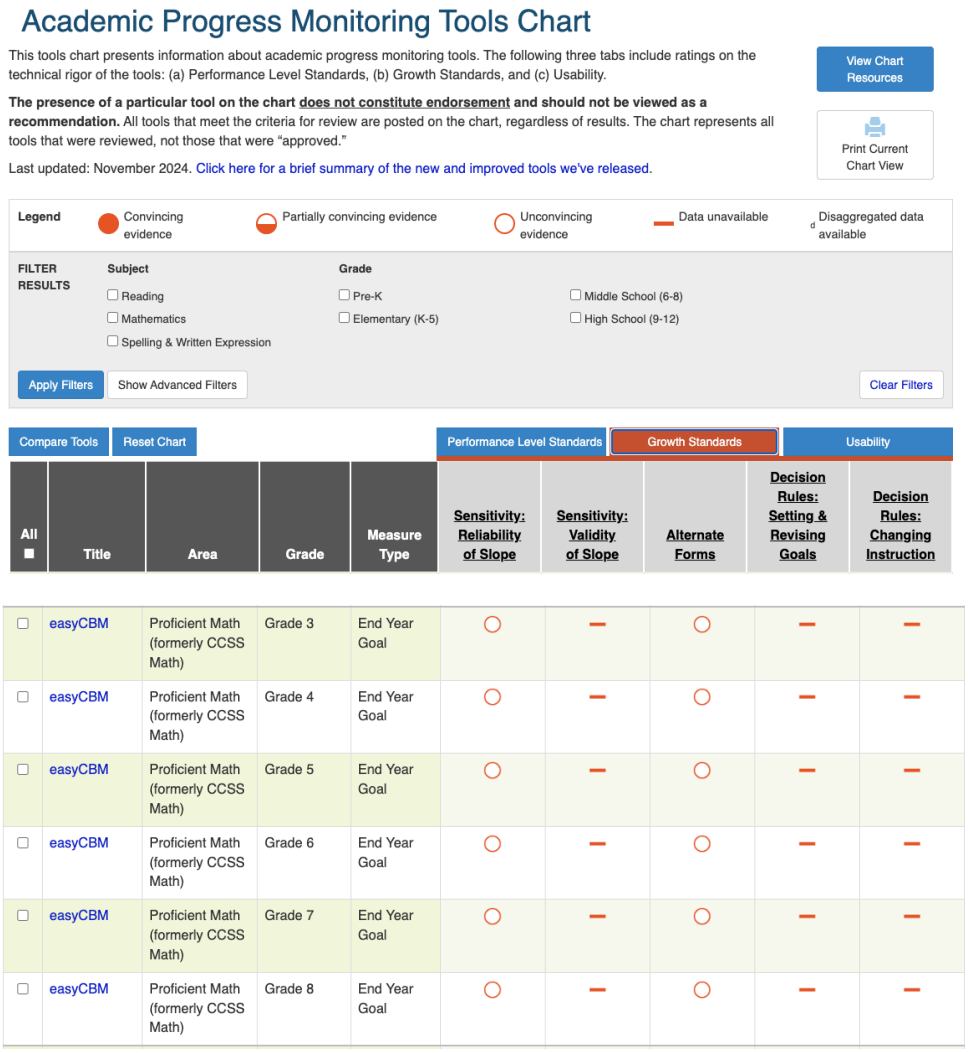
Grade: Pre-K; Elementary (K-5); Middle School (6-8); High School (9-12)

Apply Filters | Show Advanced Filters | Clear Filters

Compare Tools | Reset Chart | Performance Level Standards | Growth Standards | Usability

All ■	Title	Area	Grade	Measure Type	Reliability	Validity	Bias Analysis
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 3	End Year Goal	○d	○	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 4	End Year Goal	○d	◒	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 5	End Year Goal	○d	◒	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 6	End Year Goal	○d	●	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 7	End Year Goal	○d	●	Provided
☐	easyCBM	Proficient Math (formerly CCSS Math)	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; 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	Proficient Math (formerly CCSS Math)	Grade 3	End Year Goal	○	–	○	–	–
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 4	End Year Goal	○	–	○	–	–
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 5	End Year Goal	○	–	○	–	–
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 6	End Year Goal	○	–	○	–	–
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 7	End Year Goal	○	–	○	–	–
☐	easyCBM	Proficient Math (formerly CCSS Math)	Grade 8	End Year Goal	○	–	○	–	–

Published Research Literature on Progress Monitoring with eacyCBM Math Measures

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 MGCI_3.1 documentation and concludes that easyCBM® mathematics score reports and supporting materials appropriately support valid interpretations and uses of overall achievement.

Indicator 3.1.a: The design of mathematics score reports is clearly aligned with intended users and purposes. Benchmark and progress-monitoring reports support educators, administrators, parents, and students through multiple formats and levels of aggregation. Overall mathematics achievement is represented through Proficient Math benchmark scores and complementary Basic Math measures, allowing both universal screening and instructional adjustment. The documentation demonstrates explicit attention to user needs, including cautions for students performing well below grade level and guidance on when alternative measures should be used. Conditions that may compromise interpretation (e.g., reliance on Basic Math for screening) are clearly articulated, reducing the risk of misuse.

Indicator 3.1.b: The documentation provides defensible information about measurement error and score variability. Overall achievement interpretations are grounded in a Multi-Skill, Multi-Method (MS-MM) framework, integrating multiple domains (e.g., numbers and operations, algebra, geometry, measurement, and data analysis) assessed with varied item formats. Norm-referenced percentile ranks contextualize raw scores, and guidance explains how convergent and divergent patterns across domains should be interpreted cautiously. Although numeric confidence intervals are not always displayed, the combination of distributional information, benchmarks, and multiple measures supports accurate understanding of uncertainty in achievement scores.

Indicator 3.1.c: Extensive guidance is provided to support appropriate use of mathematics achievement results. Interpretive guidance is aligned with MTSS decision making and grounded in national standards (CCSS and NCTM Focal Points) and CBM research. Teachers are guided in using percentile ranks and benchmark distributions to support decisions across the full performance range, from enrichment to intensive intervention. The integration of overall achievement with sub-scores and progress monitoring further supports coherent, defensible instructional planning.

Overall, MGCI_3.1 demonstrates strong coherence among report design, interpretive guidance, and intended uses, supporting valid, transparent, and responsible interpretation of overall mathematics achievement.

3.1a The Design of the Score Reports to Document Overall Achievement in 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.

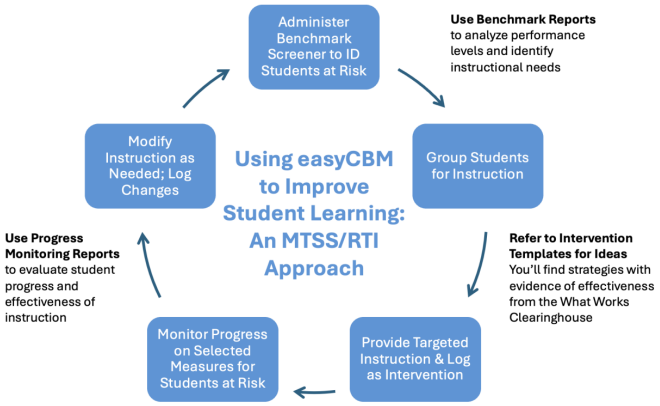
Table 1. Usage and Reporting for easyCBM® (User's Manual, page 4)

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 'Mathematics'. 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 mathematics score for students in Grades Kindergarten through 8. Many different 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).

In **Mathematics**, a total score on Proficient Math is expressed as a raw score: In Kindergarten through Grade 2, the total is 25; in Grades 3 through 8, the total is 30. This measure is used for screening purposes and consideration of risk successful in learning math problem solving. The Proficient Math measures were developed using the Common Core State Standards (CCSS) as an initial framework. In addition to items aligned with the respective grade level, the Proficient Math benchmark measures also include a small number of items from prior and subsequent grade levels to enhance the test's accuracy as a universal screener, thereby extending the population of students whom it reliably measures. The Proficient Math measures were designed to be more challenging, in line with high expectations of grade-level performance.

[1] https://conjointly.com/kb/multitrait-multimethod-matrix/

Most students should receive benchmark testing on Proficient Math, but for students performing well below grade level (i.e., recently scored at or below the 10th percentile on other Proficient Math forms), Basic Math might provide a more accurate assessment. It is important to remember that these students should still be considered at high risk even if their performance on Basic Math places them at a higher percentile rank. This Basic Math measure is available for teachers that is oriented toward progress monitoring, using items designed to be more basic (hence the name Basic Math Measures) and with sub scores available for progress monitoring in the following areas, with each domain presenting 16 items (which is addressed in more detail in Section 3.3 and 3.4).

- Kindergarten: Numbers/Operations, Geometry, and Measurement
- Grade 1: Numbers Operations, Geometry, and Numbers Operations/Algebra
- Grade 2: Numbers Operations, Measurement, Numbers Operations/Algebra
- Grade 3: Numbers Operations, Geometry, and Numbers Operations/Algebra
- Grade 4: Numbers/Operations, Measurement, and Numbers Operations/Algebra
- Grade 5: Numbers Operations, Geometry/Measurement/Algebra, and Numbers Operations/Algebra
- Grade 6: Numbers Operations, Algebra, and Numbers Operations/Ratios
- Grade 7: Numbers Operations/Algebra/Geometry, Measurement/ Geometry/Algebra, and Numbers Operations/Algebra
- Grade 8: Algebra, Geometry/Measurement, and Data Analysis/Numbers Operations/Algebra

MS-MM Math Measures (User's Manual, pages 15-16)

The Proficient Math measures were developed using the Common Core State Standards (CCSS) as an initial framework. In addition to items aligned with the respective grade level, the Proficient Math benchmark measures also include a small number of items from prior and subsequent grade levels to enhance the test's accuracy as a universal screener, thereby extending the population of students whom it reliably measures. The Proficient Math measures were designed to be more challenging, in line with high expectations of grade-level performance. Most students should receive benchmark testing on Proficient Math, but for students performing well below grade level (i.e., recently scored at or below the 10th percentile on other Proficient Math forms), Basic Math might provide a more accurate assessment. It is important to remember that these students should still be considered at high risk even if their performance on Basic Math places them at a higher percentile rank.

The Basic Math measures were developed using the National Council of Teachers of Mathematics (NCTM) Focal Point Standards as an initial framework, with benchmark forms including test items from all three focal point standards at each respective grade level. The Basic Math measures were designed to be more easily accessible (fewer cognitive demands for processing what is being asked) and to assess a more foundational understanding of math, making them most appropriate for students who are performing substantially below their grade-level peers.

Table 2. Measures Defining Basic and Proficient Constructs of Math Computation and Application (User's Manual, page 14).

easyCBM Math Measures by Content Area

	Basic Math						Proficient Math
	The content areas assessed by Basic Math were based on the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Point Standards in Mathematics. For each grade, the content areas below are grouped into three focal point standards. For example, in Grade 8, the focal point standards are 'Algebra,' 'Geometry and Measurement,' and 'Data Analysis, Numbers and Operations, and Algebra.'						The content areas for Proficient Math were based on Common Core State Standards (CCSS).
Grade	Numbers and Operations	Geometry	Measurement	Algebra	Data Analysis	Ratios	Common Core
K	✓	✓	✓	*	*		✓
1	✓	✓	*	✓	*		✓
2	✓	*	✓	✓	*		✓
3	✓	✓	*	✓	*		✓
4	✓		✓	✓	*		✓
5	✓	✓	✓	✓	*		✓
6	✓	*		✓	*	✓	✓
7	✓	✓	✓	✓	*	*	✓
8	✓	✓	✓	✓	✓	*	✓

** Note: Asterisks indicate Connections to Focal Points as identified by NCTM. Within the constructs of mathematics, elements are woven in to build the foundation and progress a student to the next level or next topic. For example, as a Kindergarten student identifies, duplicates, and extends simple number patterns and sequential growing patterns, they are receiving foundational preparation for creating rules that describe relationships in algebra (adapted from NCTM Focal Points).*

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 based on a single score. Rather it is 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 math 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.

An analysis of overall achievement in math can be conducted when item and focal points are considered. With low overall achievement in numbers and operations may converge with similar low achievement in measurement or data analysis, which typically require operational proficiency in applied problem-solving. Or algebraic relations may provide swing information for understanding performance in data analysis.

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, in mathematics with operational proficiency and other skill areas that depend upon it: When students' overall achievement is low but their mathematical problem solving in data analysis or ratios is high, this divergent pattern requires follow up assessments, perhaps using different methods. 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 10^{th} or 15^{th} 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].

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)[3].

Table 3. 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 4. Validation Supporting Decision Making

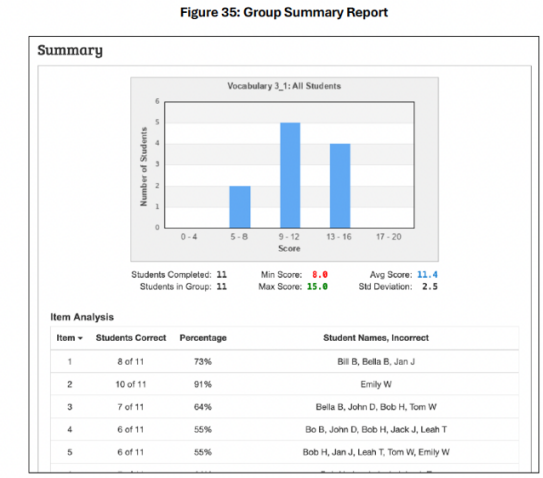
Figure Number and Display	Information Presented	Decision/Interpretation to Make
Figure 2. Bar chart	Distribution of students in classrooms (with skills)	Grouping students with minimum variance
Figure 3. Time series	Progress of students (in groups) progress over time	Confirm/disconfirm grouping composition

[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.

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

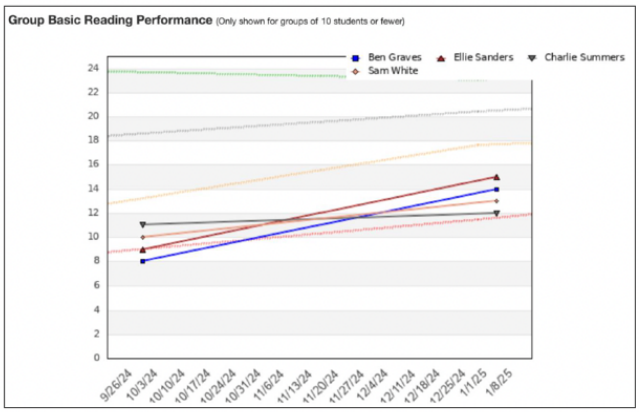
This performance on BMs can also be used to help teachers **group students** and display strengths and weaknesses among students based on an item analysis. **Figure 2** illustrates how benchmark performance combined with item-level skill analysis can be used to group students for instruction. Rather 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.

Figure 2. Using Benchmark Performance to Group Students with Skill Analysis (User's Manual, page 95)



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)



In summary, the *Standards* set the stage for state and local educational agencies (SEAs and LEAs, respectively) to adopt and deploy testing programs ranging in use from accountability systems (Center for Assessment and Peer Review) to Multi-tier Support Systems (MTSS).

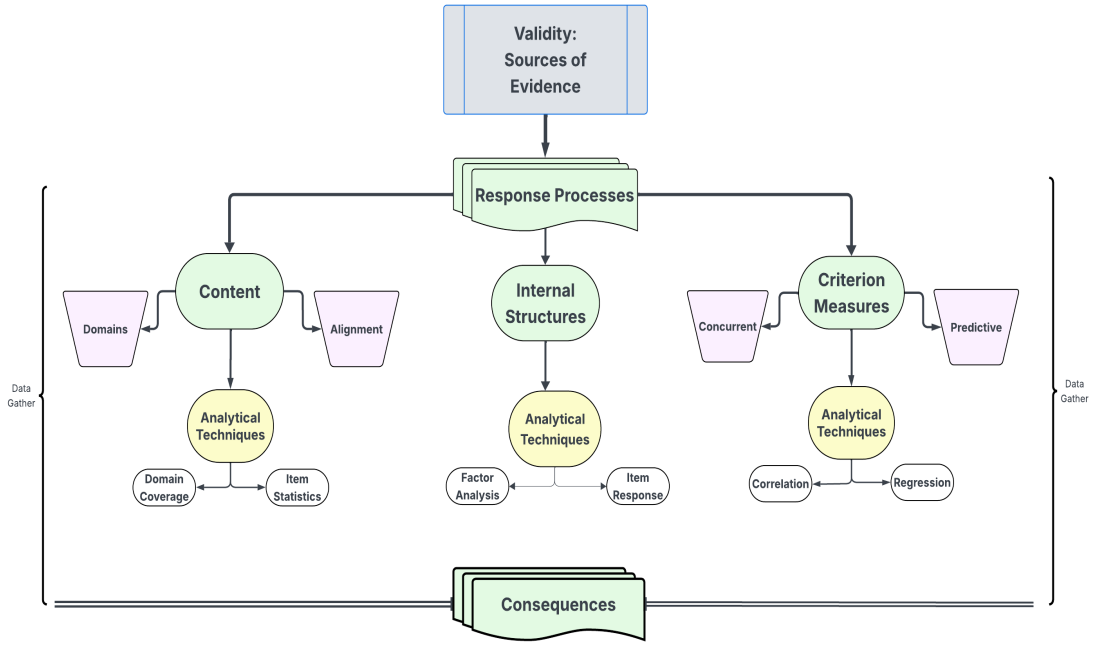
As of 2026, numerous vendors have been listed with NCII with evaluations of their MTSS measurement programs, most of them listed in a publication by the (Oregon Department of Education, 2025) in their survey of academic tests (the District Assessment Inventory, or DAI) that Oregon school districts required schools to administer…"Based on this survey's findings, ODE has provided recommendations and best practices that include observations of areas where current practices are to be commended and suggestions for improvement. The recommendations and best practices are organized in 10 categories:

1. Clarify assessment purpose and goals.
2. Align tests with learning objectives.
3. Use a variety of assessment methods.
4. Integrate assessment into instruction.
5. Leverage staff resources and technology.
6. Prioritize essential tests.
7. Optimize assessment timing.
8. Collaborate with students, families, and educators.
9. Provide professional development.
10. Regularly evaluate and adjust.

Immediately following publication of this inventory, the Oregon state legislature passed House Bill 141 (Oregeon Legislature: 83rd OREGON LEGISLATIVE ASSEMBLY–2025 Regular Session, 2025) requiring all districts to adopt one of four interim assessment that had been reviewed by a panel of measurement experts. The evaluation was based on the Gateways-Criteria-Indicators published by the Center for Assessment. In the end, the assessment landscape in Oregon is now cluttered with multiple assessments being used for both MTSS and state accountability (which also includes the state summative test used to pass peer review, as noted above). We argue, however, that many (most) of these MTSS assessment programs are quite applicable as Interim Assessments, given that they can satisfy the criteria promulgated by the Center for Assessment. At the same time, it is possible for an interim assessment to be used in the context of MTSS. However, this equivalence in function depends upon changing the validation process. And this transformation in validation needs to mover from a traditional perspective of validity to a more comprehensive view.

In a traditional view of validity, measurement systems are evaluated rather holistically and categorically. For example, a test is developed with specific content and administered under standardized conditions. Thus, two major sources of evidence are considered, though the response processes are rarely overtly considered in the context of universal design and accommodations (Tindal, 2025). Then, other sources of evidence are used to establish that the measure in question has not only integrity as a measure (with coherent internal structures) but also relates to other measures. In the end, this series of validation inquiries provide a holistic view of consequential validity.

Figure 4. Diagram of Traditional View of Reliability and Validity



This traditional view, however, is insufficient in expanding the view of validity to the full range of differentiated decision-making for MTSSs. Not only do the NCII criteria (as they currently exist) fail to satisfy a deep validation of MTSS but they also fail to satisfy the broader criteria developed for interim assessments. Furthermore, the criteria for interim assessments also fail to satisfy the criteria needed for a comprehensive validity argument applied to a Multi-Tiered System of Supports (MTSS) (Oregon Department of Education, 2025). In this last section, the *Standards* are applied in a multi-variate manner across the range of measures used, from the initial benchmarks to the eventual progress measures. We use easyCBM® as an example.

Application of a Comprehensive Validity Argument

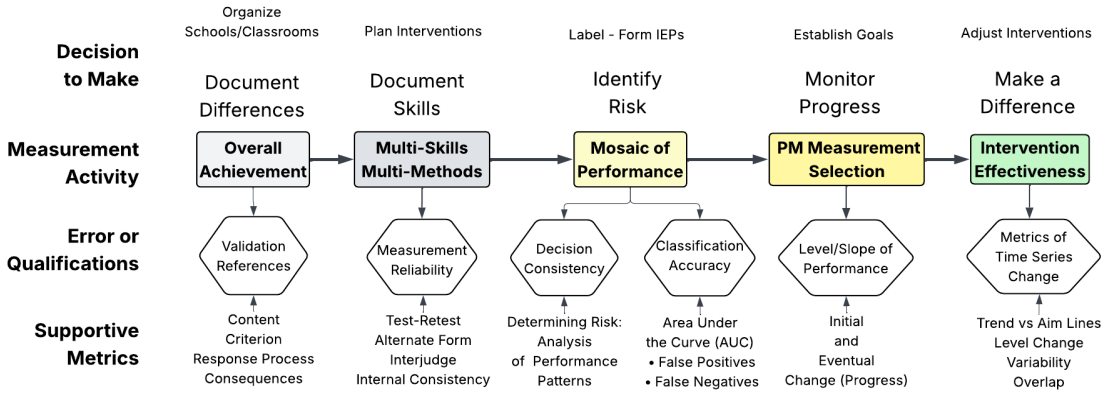
easyCBM® reports both an Overall Outcome in each subject area (Reading, Spanish literacy, and Math) as well as a matrix reflecting Multi-Skill Multi Method (MS-MM) approach: multiple skills are assessed using multiple methods (selection and production responses as well as computer-based and paper-pencil 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 **Comprehensive 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. 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*. 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 5** 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. In the end, all scientific findings are tentative ad infinitum—never absolute (Attribution to Dr. Ed Kame'enui).[4]

To complete this comprehensive validity argument, data need to be collected and used to vindicate or modify interpretations and decisions made with attention to consequential validity. "Evidence for Validity and Consequences of Testing: Some consequences of test use follow directly from the interpretation of test scores for uses intended by the test developer. The validation process involves gathering evidence to evaluate the soundness of these proposed interpretations for their intended uses. Other consequences may also be part of a claim that extends beyond the interpretation or use of scores intended by the test developer" (American Educational Research Association et al., 2014).

Figure 5. Comprehensive Validity Argument in easyCBM®



In supporting the claims, evidence would address the differences among students in their overall achievement. This evidence ideally would attend to not only averages but perhaps more importantly, the variance among different aggregates and would be documented at all levels. Teachers would use this information to group students so instruction could be efficient. Principals could use this information to streamline staffing and allocation of support resources (like instructional assistants and parent volunteers). Central office personnel (curriculum coordinators and elementary/secondary coordinators as well as school psychologists and teachers on special assignment serving as consultants) could use this information to organize their schedules and address targeted support.

Overall achievement, however, is unlikely to be sufficient in developing actionable interventions. Rather, a more fine-grained analysis is warranted using the MS-MM Matrix. Again, at all levels, information across and within subject domains (e.g., reading and mathematics) would warrant a refined proactive analysis and reactive response that is targeted. Such analyses can then be skill-specific and timely. Importantly, these analyses set up a responsive time-series mechanism for interventions being implemented 'just in time'.

Given the assumption that diagnoses are limited and not always correctly made, the next set of interpretations and decisions simply increase the stakes. In a MS-MM system for identifying risk, two classes of information are necessary. The first is at the *teacher* level, individually or as part of a Response-to-Intervention (RTI) team, as well as at the *individual* student level. Risk can be operationalized not only in terms of ranking students on a normative

[4] Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge – 2nd Edition*. New York: Routledge.

basis overall (low percentile ranks), but also in the pattern of skill deficits. Given that all skills are not equal, attention can be devoted to necessary pre-requisite skills as the basis for classifying students in need of support. Ideally, the measurement system has established validation data using classification accuracy, which can serve as the basis for judging sensitivity and specificity.

Once students have been grouped into more intensive tiers (Tier 2 and Tier 3), the somewhat ambiguous task remains to monitor progress with the appropriate measures. Assuming the previous interpretations and decisions have been carefully wrought, measures can be selected as 'a bird in the mine shaft' where generalizations can be made to the larger constructs of grade appropriate reading and mathematics problem-solving. Verification can be determined by attending to initial level of performance and immediate gains made, with changes made before too much time has been taken. Obviously, the skills for monitoring progress would also map tightly into the interventions being deployed.

The final interpretations and decisions focus on the effects of interventions and answering the following question: Is change in performance and progress being made? Answers to this critical question can be more specifically answered by addressing the following four questions.

1. Is level of performance sufficient, reflecting neither a floor nor ceiling effect?
2. Is change over time (slope or rise over run) being made?
3. Is variation of performance minimal, particularly relative to the slope?
4. Are goals being met, particularly relative to the slope?

If these basic questions are not affirmatively answered, individually and collectively, intervention changes are warranted. At this point, the validation process moves to an interrupted time-series design and the student's graph is punctuated with a vertical line that separates the data series into pre- and post. Three new questions can then be answered:

5. Does a change in level occur (a comparison between the last data value of the previous intervention and the first data value of the new intervention)?
6. Does the slope increase and the variation around it decrease?
7. Is overlap minimal: A horizontal line demarcating the difference between the highest data value in previous intervention and the lowest data value in the subsequent intervention.

In summary, a comprehensive validity argument involves tying claims to evidence and iteratively addressing specific interpretations. And once begun, " it is commonly observed that the validation process never ends, as there is always additional information that can be gathered to more fully understand a test and the inferences that can be drawn from it. In this way an inference of validity is similar to any scientific inference" (American Educational Research Association et al., 2014).

Appendix A: Technical Report Table and Figure Titles

Table 1. Usage and Reporting for easyCBM® (User's Manual, page 4)

Table 2. Measures Defining Basic and Proficient Constructs of Math Computation and Application (User's Manual, page 14).

Table 3. Cross-Tabulation of Perspectives and Time

Table 4. Validation Supporting Decision Making

Table 5. Cross-Tabulation of Perspectives and Time

Table 6. Validation Supporting Decision Making

References

American Educational Research Association, American Psychological Association, & Education, N. C. o. M. i. (2014). *Standards for Educational and Psychological Testing*. Author.

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

Oregeon Legislature: 83rd OREGON LEGISLATIVE ASSEMBLY–2025 Regular Session. (2025). *House Bill 141*. Salem, OR

Oregon Department of Education. (2025). *HB 4124 Legislative Report: District Assessment Inventory*. Salem, OR: Author

Tindal, G. (2025). *Rethinking "Standardization" for NAEP to Increase Equity and Access (Technical Report 2510)*. 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 MGCI_3.2 documentation and finds strong evidence that easyCBM® mathematics score reports and supporting materials appropriately facilitate valid interpretations and uses of predicted student performance.

Indicator 3.2.a: The design of benchmark score reports and interpretive materials is well aligned with intended users and purposes. Mathematics benchmark reports clearly support universal screening by presenting grade-level performance relative to end-of-year expectations across fall, winter, and spring administrations. The distinction between Proficient Math (used for risk classification) and Basic Math (used diagnostically and for students performing well below grade level) is explicitly documented, reducing the likelihood of misinterpretation. The type and grain size of reported information—scores, percentile ranks, and benchmark windows—are appropriate for classroom, school, and district decision making. Clear warnings are provided regarding common misuses, including interpreting Basic Math percentiles as indicating reduced risk.

Indicator 3.2.b: The documentation provides defensible representations of uncertainty in predicted performance. While numeric confidence intervals are not always displayed on reports, percentile ranks derived from large, nationally representative samples provide a clear probabilistic interpretation of student standing. Supporting text explains how percentile variability should be interpreted across seasons and cautions against over-interpreting small differences, particularly near cut points. These features support accurate understanding of score variability in applied decision contexts.

Indicator 3.2.c: Extensive, research-based guidance supports appropriate interpretation across the full performance range. The 2025 norms technical report documents a rigorous norming methodology, including large, stratified samples, seasonal benchmarking, and non-parametric percentile estimation well suited to CBM score distributions. Guidance explicitly links predicted performance interpretations to intended uses such as early risk identification, tiered intervention placement, and instructional planning. The documentation reflects consultation with experienced educators and aligns with national testing standards.

Overall, MGCI_3.2 demonstrates strong coherence between report design, normative evidence, and interpretive guidance, supporting transparent, defensible, and equitable use of predicted mathematics performance data.

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 and mathematics 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.

3.2b Score Reports Include Information on the Degree of Error

Basic Math is an optional benchmark assessment. For universal screening, most students should be given Proficient Math. However, for students performing well below grade level, it may make sense to administer Basic Math in place of or in addition to the Proficient measures. If a student is unable to complete the Proficient test (e.g., due to low accuracy or fluency), consider giving Basic instead. If a student has recently scored at or below the 10th percentile on other Proficient Math forms, you may need to administer Basic Math to get a better idea of the student's skill. It is important to remember that these students should still be considered at high risk even if their performance on Basic Math places them at a higher percentile rank.

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 very 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 1. 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 mathematics 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 mathematics 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 mathematics 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.

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 math 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 mathematics measures. Mathematics norms are provided through the Proficient Math measure. Results illustrate clear seasonal growth patterns across all mathematics measures.

Proficient Math norms provide stable benchmarks across Grades K–8 for interpreting student performance relative to national expectations in mathematical reasoning and computation. 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 mathematics 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 2. Summary of Major easyCBM 2025 Norms Findings

Domain	Key Findings	Educational Implications
Proficient Math (K–8)	Stable developmental increases across all grades.	Enables math screening and progress monitoring using national benchmarks.

Appendix A: Technical Report Table Titles

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 MGCI_3.3 documentation and finds strong evidence that mathematics sub-score reports and associated resources support valid, defensible interpretations and uses of sub-scores within instructional and MTSS frameworks.

Indicator 3.3.a: The design of mathematics score reports is closely aligned with intended interpretations and users. Reports present sub-scores at multiple, coherent grain sizes—individual student, classroom, grade, school, and district—supporting decisions about instructional grouping, tier placement, and resource allocation. Visual displays such as heat maps, stacked bar charts, and benchmark-to-benchmark change analyses explicitly link reported information to actionable decisions. Documentation demonstrates clear attention to audience needs, particularly educators and administrators, and includes explicit cautions against common misuses (e.g., treating risk as fixed or global). Conditions that may compromise interpretation, including benchmark window selection and cut-score choices, are clearly articulated in interpretive guidance.

Indicator 3.3.b: The documentation provides meaningful representations of uncertainty associated with mathematics sub-scores. Although numeric confidence intervals are not always displayed, uncertainty is communicated through percentile ranks, color-coded risk bands, and longitudinal displays showing movement across benchmarks and just noticeable differences. Guidance on sensitivity, specificity, and false-positive/false-negative trade-offs in setting cut scores further supports accurate interpretation of error and classification risk.

Indicator 3.3.c: Extensive, research-based guidance supports appropriate use of mathematics sub-scores across the full performance continuum. Guidance is grounded in CBM theory, MTSS practice, and national mathematics standards, and reflects consultation with experienced educators. Sub-score patterns are explicitly linked to targeted instructional responses and to evaluation of instructional effectiveness over time, integrating both nomothetic (risk classification) and idiographic (change and response) perspectives.

Overall, MGCI_3.3 demonstrates strong coherence among sub-score design, treatment of uncertainty, and interpretive guidance, supporting transparent, equitable, and instructionally meaningful use of mathematics sub-scores.

3.3a Documenting Risk using Sub Scores in Achievement (from User's Manual, pages 91-92)

easyCBM District Edition offers a variety of reports designed to provide useful information to guide decision-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).

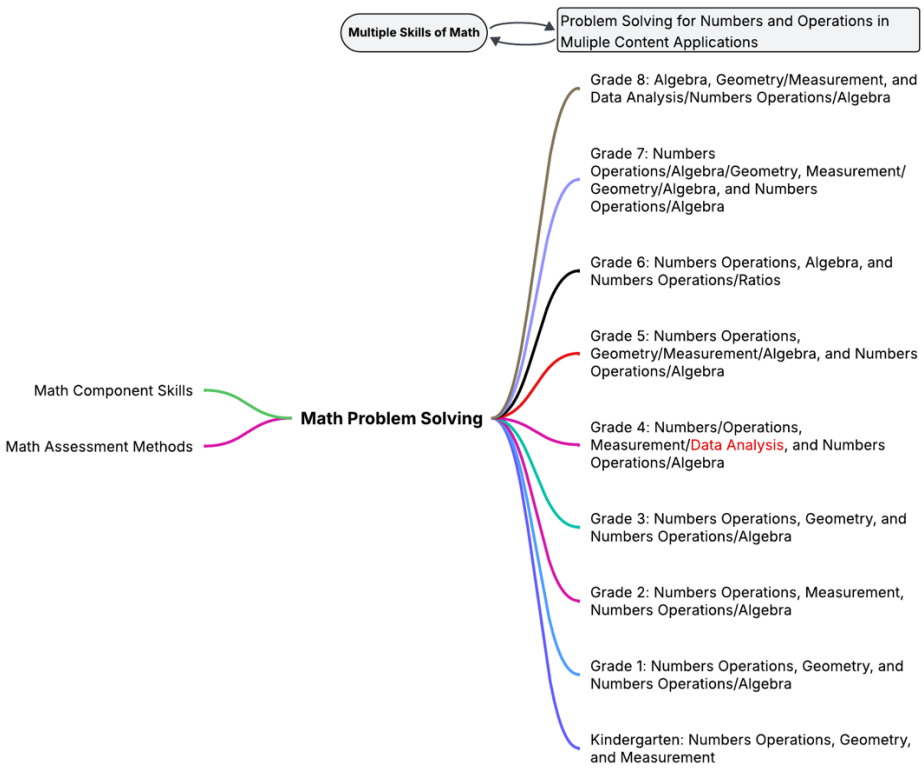
Basic Math measures are available for teachers that are oriented toward progress monitoring, using items designed to be more basic (hence the name Basic Math Measures) and with sub scores available for progress monitoring in the following areas, with each domain presenting 16 items (which is addressed in more detail in Section 3.3 and 3.4). The multiple skills in math are grade-specific and braided alternately over time.

- Kindergarten: Numbers/Operations, Geometry, and Measurement
- Grade 1: Numbers Operations, Geometry, and Numbers Operations/Algebra
- Grade 2: Numbers Operations, Measurement, Numbers Operations/Algebra
- Grade 3: Numbers Operations, Geometry, and Numbers Operations/Algebra
- Grade 4: Numbers/Operations, Measurement/Data Analysis, and Numbers Operations/Algebra
- Grade 5: Numbers Operations, Geometry/Measurement/Algebra, and Numbers Operations/Algebra
- Grade 6: Numbers Operations, Algebra, and Numbers Operations/Ratios
- Grade 7: Numbers Operations/Algebra/Geometry, Measurement/ Geometry/Algebra, and Numbers Operations/Algebra
- Grade 8: Algebra, Geometry/Measurement, and Data Analysis/Numbers Operations/Algebra

In **Figure 1** below, latent **Math** skills and methods are operationalized into a single construct of Math Problem Solving. In easyCBM, overall achievement can be delivered with two measures: Proficient and Basic, 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 1. Math Sub Scores Critical to the Overall Measure of Math Problem Solving



Sub Score Risk Reports

- 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.
- 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 2. Benchmark Windows and Levels	When to activate BMs and set levels of risk	Allocation of Tiers of Support to provide
Figure 3. Heat map	Risk Analysis for students by measures in classrooms	Tier of Support to provide
Figure 4. Heat Map	Risk Analysis Reading Composite by classroom	Tier of Support to provide
Figure 5. Vertical stacked bar chart	Individual standing in risk bands	Continue or change Tier of Support to provide
Figure 6. Heat map	Change in risk by BMs and measures over grades	Confirm/disconfirm risk and Tiers of Support
Figure 7. Heat Map/Bar Chart	Risk Analysis for students across measures for BMs and grades	Tier of Support to provide and confirmation/disconfirmation
Figure 8. Heat map	Change in risk by BMs and measures over grades	Confirm/disconfirm risk and Tiers of support
Figure 9. Table of percentages	Changes in risk level across BMs and grades	School allocation of resources to assign
Figure 10. 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.

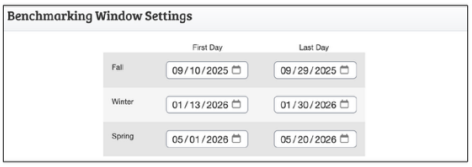
Districts have options in determining risk using percentile ranks as part of MTSS. The rationale for this option is that districts may differ in their student populations and available resources. Because assignment of special education

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 2 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 2. Adjusting Benchmark (BM) Test Windows and Risk Values (User's Manual, pages 76-77)

Figure 24: Benchmark Window Settings



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

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 3 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 3. Display of Risk for Different Benchmark easyCBMs (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	

This risk analysis can also be used to measure change across successive BMs to determine if a just noticeable difference (JND) is occurring. **Figure 4** 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 4. Displaying Risk for Individuals across Measures and Benchmarks (User's Manual, page 101)

Figure 42: Benchmark Performance Report for Individual Students

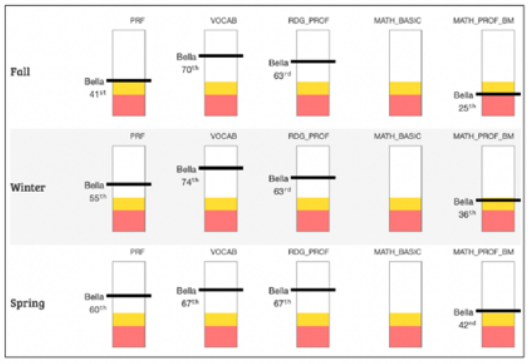
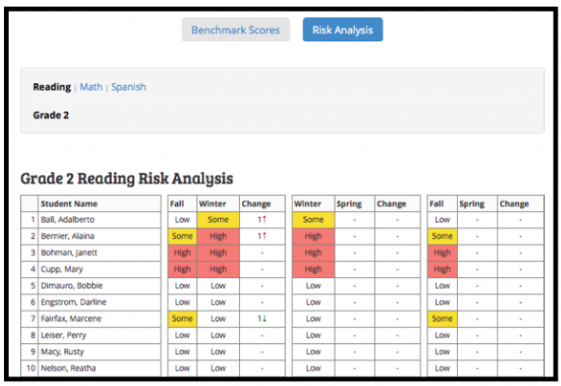


Figure 5 focuses on changes in student risk status across successive benchmark periods. Rather than emphasizing static classification, the display highlights movement over time, allowing educators to evaluate whether instruction is producing a just noticeable difference (JND). This visualization supports 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 5. Classroom Analysis of Risk and Change in Risk across Benchmarks (User's Manual, page 103)

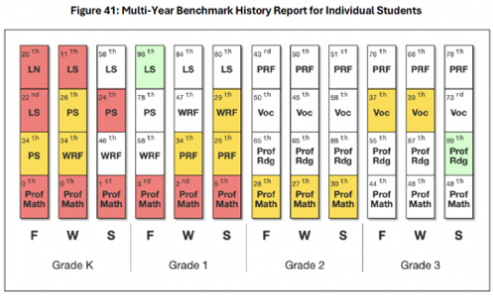


	Student Name	Fall	Winter	Change	Winter	Spring	Change	Fall	Spring	Change
1	Ball, Adalberto	Low	Some	↑↑	Some	-	-	Low	-	-
2	Bernier, Alaina	Some	High	↑↑	High	-	-	Some	-	-
3	Bohman, Janett	High	High	-	High	-	-	High	-	-
4	Cupp, Mary	High	High	-	High	-	-	High	-	-
5	Dimauro, Bobbie	Low	Low	-	Low	-	-	Low	-	-
6	Engstrom, Darline	Low	Low	-	Low	-	-	Low	-	-
7	Fairfax, Marcene	Some	Low	↑↓	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 6)**, for individual measures across Benchmark and Grades numerically documenting change **(Figures 7)**, counts and percentages of students Changing Risk across BMs and Grades **(Figure 8),** and Comparing Schools **(Figure 9)** 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 6, 7, and 8**, the analysis at the grade level can also focus on different instructional programs in the aggregate aimed to ameliorate specific skill deficits.

Figure 6. 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 7**. These data can be displayed within a grade level using actual numerical values, which provide a more sensitive view of outcomes.

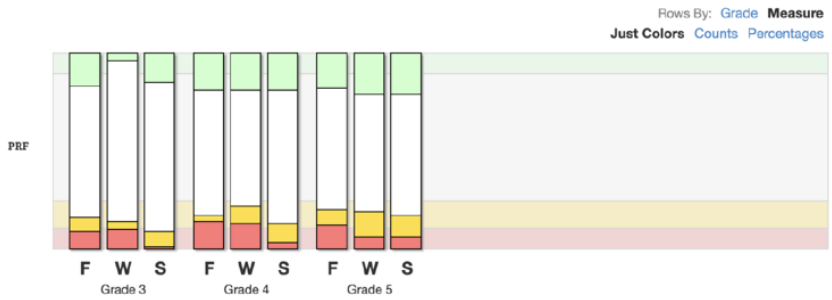
Figure 7. Displaying Risk within Measures and Grade Benchmarks (User's Manual, page 104)

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%	-

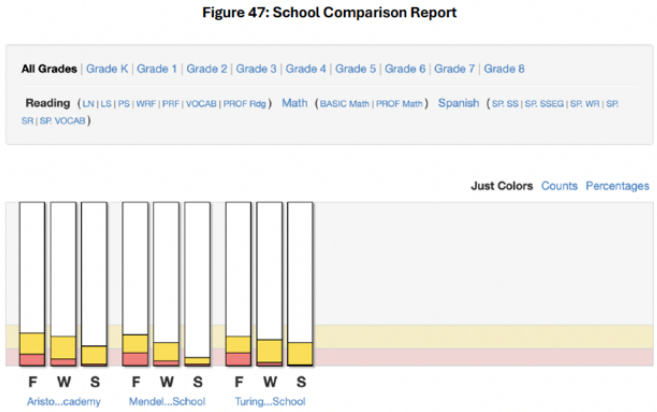
Figure 8 extends risk analysis for individual measures across benchmarks and grades, supporting a broader longitudinal reflection. The decision supported is whether instructional programs are producing sustained improvement. This figure aligns with Testing Standards 1.4 and 7.4 by supporting evaluation of long-term consequences.

Figure 8. Risk within Measures Across Benchmarks and Grade across (User's Manual, page 105)



Finally, **Figure 9** 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 9. Comparisons of Schools at Risk across Benchmarks (User's Manual, page 106)



Once students' risk can be 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 Titles and Figures

Table 1. Validation Supporting Decision Making

Figure 1. Math Sub Scores Critical to the Overall Measure of Math Problem Solving

Figure 2. Adjusting Benchmark (BM) Test Windows and Risk Values (User's Manual, pages 76-77)

Figure 3. Display of Risk for Different Benchmark easyCBMs (User's Manual, page 93)

Figure 4. Displaying Risk for Individuals across Measures and Benchmarks (User's Manual, page 101)

Figure 5. Classroom Analysis of Risk and Change in Risk across Benchmarks (User's Manual, page 103)

Figure 6. Displaying Risk across Measures, Benchmarks, and Grades (User's Manual, page 100)

Figure 7. Displaying Risk within Measures and Grade Benchmarks (User's Manual, page 104)

Figure 8. Risk within Measures Across Benchmarks and Grade across (User's Manual, page 105)

Figure 9. Comparisons of Schools at Risk across Benchmarks (User's Manual, page 106)

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.

Criterion 3.4: Student Progress

This evaluation applies Criterion 3.4 (Indicators 3.4.a–3.4.c) to all sections of the MGCI_3.4 documentation and finds strong evidence that easyCBM mathematics progress-monitoring reports and supporting resources appropriately facilitate valid interpretations and uses of student growth data.

Indicator 3.4.a: The design of mathematics progress reports is clearly aligned with intended interpretations and users. Reports provide multiple grain sizes—from individual student time-series displays to classroom, school, and district summaries—supporting decisions about instructional effectiveness, tier movement, and intervention adjustment. Visual elements such as percentile bands, trend lines, goal and aim lines, and phase-change markers directly link observed growth to instructional decisions. Documentation demonstrates attention to varied audiences, including teachers, administrators, and parents, and includes explicit cautions against common misuses, such as over-interpreting sparse data or monitoring students on measures outside an appropriate instructional range.

Indicator 3.4.b: The system provides defensible representations of uncertainty in growth interpretations. While numeric confidence intervals are not always displayed, variability is communicated through percentile bands, slope stability, overlap, and comparisons of observed trends to expected progress. Supporting guidance emphasizes the importance of sufficient measurement density and cautions users against reacting to short-term fluctuations, helping ensure that instructional decisions are based on stable patterns rather than noise.

Indicator 3.4.c: Extensive, research-based guidance supports appropriate use of mathematics progress data across the full performance continuum. Guidance is grounded in CBM theory, national testing standards, and long-standing empirical research, and reflects consultation with experienced educators. Clear decision rules are provided for selecting appropriate progress-monitoring measures, adjusting grade levels, setting realistic goals, and evaluating response to intervention. The integration of intervention documentation tools, professional development modules, blogs, and “Did You Know” resources further strengthens interpretive support and promotes consistent, defensible use.

Overall, MGCI_3.4 demonstrates strong alignment among report design, treatment of error, and interpretive guidance, supporting valid, transparent, and instructionally meaningful use of mathematics 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 **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.

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.

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. The 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.

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?

Most students should receive benchmark testing on Proficient Math, but for students performing well below grade level (i.e., recently scored at or below the 10th percentile on other Proficient Math forms), Basic Math might provide a more accurate assessment. It is important to remember that these students should still be considered at high risk even if their performance on Basic Math places them at a higher percentile rank.

For progress monitoring, the Basic Math measures are split into individual NCTM focal point standards, allowing you to focus instruction and monitoring on a specific skill area, depending on student need, or to rotate between skill areas. The Proficient Math progress monitoring measures include test items covering a range of skills appropriate for the grade in question, but unlike the benchmark measures, do not include test items from earlier or later grades.

For students who scored at or below the 25th percentile on the Proficient Math benchmark assessment, we recommend conducting progress monitoring with the Basic Math measures. For students who scored between the 26th and 49th percentile, we recommend monitoring with Proficient Math. Students who performed at or above grade level (50th percentile or above on the Proficient Math benchmark assessment) do not require monitoring.

Optimally, the **Math Measures** should be used no more than once every three weeks for monitoring progress. While weekly progress monitoring in mathematics is not recommended, in situations where such frequent monitoring is required, either (a) focus on one Basic Math measure type at a time, transitioning to the next measure type after all ten progress monitoring forms have been used for a given type, or (b) rotate through the different Basic Math measures so each gets tested every three weeks or four weeks if Proficient Math is also being monitored.

- If a student requires progress monitoring in multiple math skill areas, you can either rotate through the different Basic Math measures or—for those students who perform above the 25th percentile rank at that time of the year—monitor with Proficient Math.
- If a student requires progress monitoring in multiple math skill areas, you can either rotate through the different Basic Math measures or—for those students who perform above the 25th percentile rank at that time of the year—monitor with Proficient Math.
- If a student requires progress monitoring in multiple math skill areas, either rotate through the different Basic Math measures or—for those students who perform above the 25th percentile rank at that time of the year—you can monitor with Proficient Math.

Concluding Recommendation: Frequency of Monitoring Progress should be every 3-4 weeks for Basic and Proficient Math

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 1. Content of Progress Monitoring in Math using Basic Measures

easyCBM Math Measures by Content Area

Basic Math The content areas assessed by Basic Math were based on the National Council of Teachers of Mathematics (NCTM) Curriculum Focal Point Standards in Mathematics. For each grade, the content areas below are grouped into three focal point standards. For example, in Grade 8, the focal point standards are 'Algebra,' 'Geometry and Measurement,' and 'Data Analysis, Numbers and Operations, and Algebra.'							**Proficient Math** The content areas for Proficient Math were based on Common Core State Standards (CCSS).
Grade	**Numbers and Operations**	**Geometry**	**Measurement**	**Algebra**	**Data Analysis**	**Ratios**	**Common Core**
K	✓	✓	✓	*	*		✓
1	✓	✓	*	✓	*		✓
2	✓	*	✓	✓	*		✓
3	✓	✓	*	✓	*		✓
4	✓		✓	✓	*		✓
5	✓	✓	✓	✓	*		✓
6	✓	*		✓	*	✓	✓
7	✓	✓	✓	✓	*	*	✓
8	✓	✓	✓	✓	✓	*	✓

** Note: Asterisks indicate Connections to Focal Points as identified by NCTM. Within the constructs of mathematics, elements are woven in to build the foundation and progress a student to the next level or next topic. For example, as a Kindergarten student identifies, duplicates, and extends simple number patterns and sequential growing patterns, they are receiving foundational preparation for creating rules that describe relationships in algebra (adapted from NCTM Focal Points).*

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 easyCBMs (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 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 all grades of Math, PMs rare recommended in the measure that reflects low proficiency levels.

Figure 1. Math Progress Monitoring for Students in All Grades (User's Manual, page 121)

low	below the 30th percentile
medium	30th - 49th percentile
high	50th percentile and above

Math: All Grades, When Administering Both Basic and Proficient Math

Basic Math	Proficient Math	Monitoring Suggestion
low	low	Basic Math
medium	low	Proficient Math
high	low	Proficient Math
low	medium	-
medium	medium	-
high	medium	-
low	high	-
medium	high	-
high	high	-

Math: All Grades, When Administering Only Basic Math

Basic Math	Monitoring Suggestion
low	Basic Math
medium	-
high	-

Math: All Grades, When Administering Only Proficient Math

Proficient Math	Monitoring Suggestion
low	Proficient Math
medium	-
high	-

[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.

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 2 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 2. Monitoring Student Progress Relative to Percentile Bands (User's Manual, page 97)

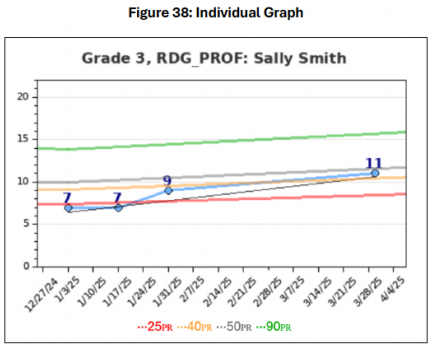
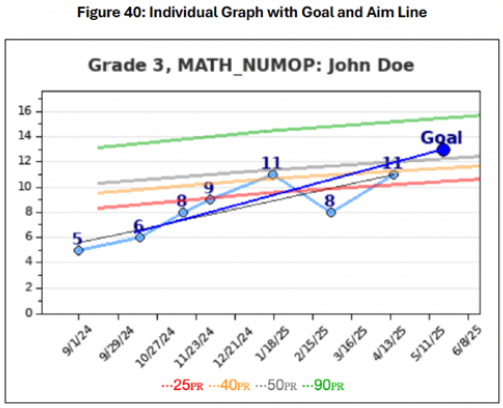


Figure 3 displays change over time relative to instructional goals and the subsequent aimlines displayed to achieve these goals. Notice how decision making can be advanced by ensuring changes are timely (noting the date and level for final progress to be made) and made relative to the variation and slope 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.

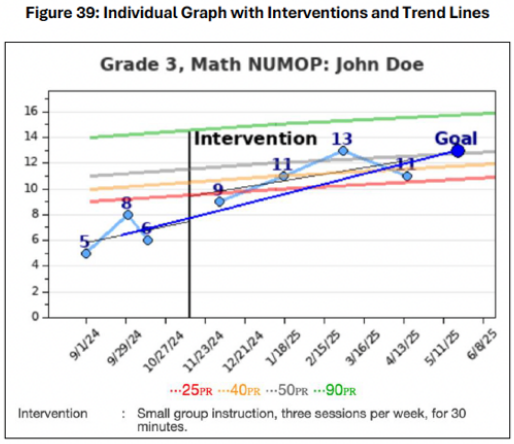
Figure 3. Monitoring Student Progress Relative to Goals and with Aim Lines (User's Manual, page 98)



Finally, the essence of decision-making is impact with the introduction of interventions as displayed in **Figure 4**. 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). Figure

12 emphasizes instructional impact by 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 4. 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 5** 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 5. 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

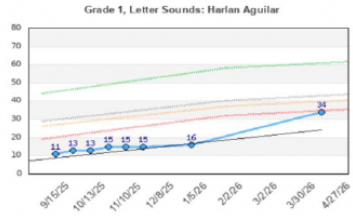
Percentile Rank
The percentage of scores that are equal to or less than a given score. Percentile ranks, like percentages, fall on a continuum from 0 to 100. For example, a test score that is greater than or equal to 75% of the scores of people taking the test is at the 75th percentile rank.

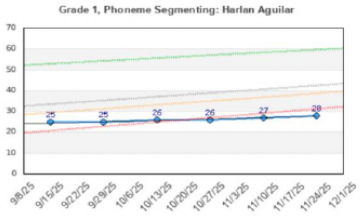
Legend/Key: District Risk Settings

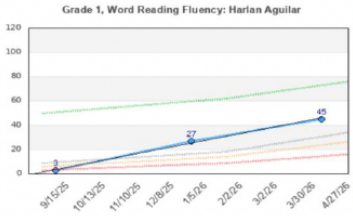
Low Risk
Some Risk
High Risk

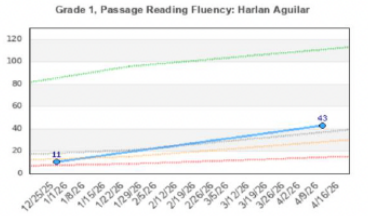
Reading Assessments

	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	



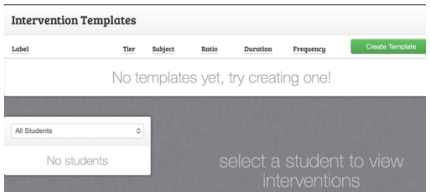


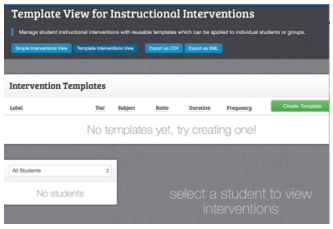




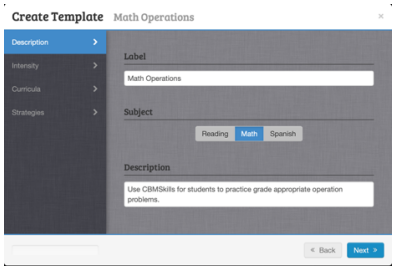
Interventions

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.

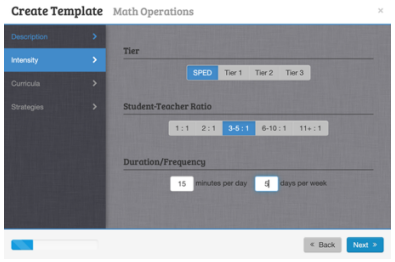




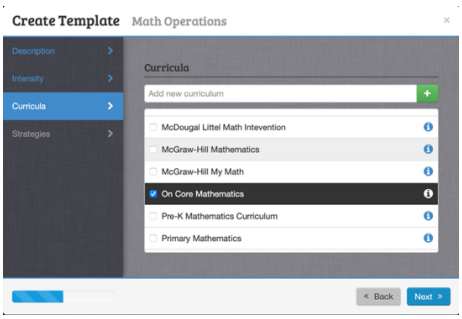
Step 1. Both a label and brief description of the intervention are entered as the initial information to input. Note that the brief description is displayed on the progress (time series) graph.

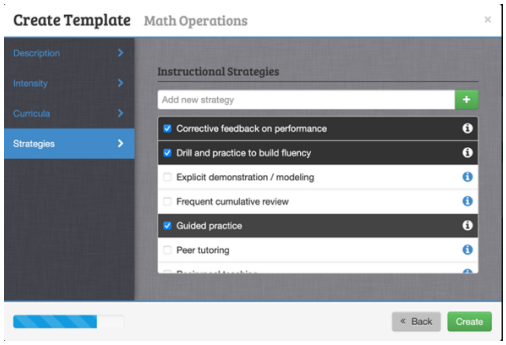
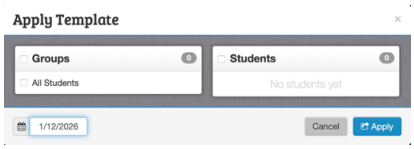


Step 2 in creating the template provides information on the **intensity** of the intervention, which is defined as the Tier (Special Education or Tiers 1 to 3) as well as the ratio of students to teachers and the number of minutes and number of days.



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.



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 Titles and Figures

Table 1. Content of Progress Monitoring in Math using Basic Measures

Table 2. Content of Reading Progress Monitoring Measurement (User's Manual, page 14)

Figure 1. Math Progress Monitoring for Students in All Grades (User's Manual, page 121)

Figure 2. Monitoring Student Progress Relative to Percentile Bands (User's Manual, page 97)

Figure 3. Monitoring Student Progress Relative to Goals and with Aim Lines (User's Manual, page 98)

Figure 4. Evaluating Instruction Based on Student Progress and Goal Attainment (User's Manual, page 98)

Figure 5. Parent Reports Showing Several Different Outcomes (User's Manual, page 102)

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, & EdReports.org. (2023b). *Review Criteria: Interim Assessment Mathematics 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.