Technical Report 2603-TDK8M

Test Development for easyCBM® in Grades K–8: Mathematics

Gerald Tindal, PhD – University of Oregon

Sara McCaslin, PhD – mccaslinwordsmithing.wordpress.com

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. & McCaslin, S. (2026). *Test Development for easyCBM® in Grades K-8 (Technical Report 2603-TDK8M)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Table of Contents for 2603-TDK8M_ItemDevMath

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 the form difficulty level.

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 the attached 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.

---

[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

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.

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 that 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. For early grades, items often include strong visual supports and simplified vocabulary mathematics-specific terms related to the construct. Writers also attend to formatting details that can 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 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. Rather, equivalence is supported by converging indicators: matched mean difficulties across forms, overlapping test characteristic curves or information functions, strong inter-form relationships where reported, 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 statistics 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)
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 focused on establishing the 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. Both general education students and students eligible for the 2% alternate assessment population were represented. Inclusion criteria ensured appropriate grade-level placement, while demographic data were collected to support 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 the majority of 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 from operational forms.

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)*

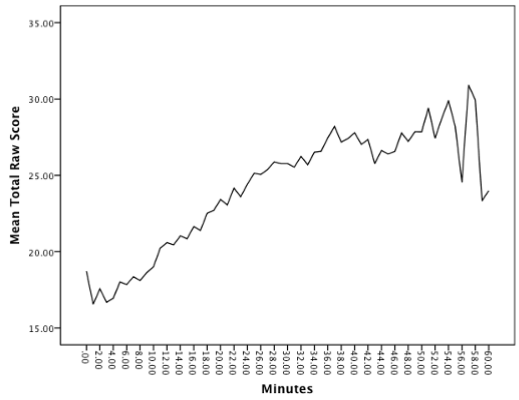

Figure 2. Example Test Information from Technical Report 1207

Figure 2

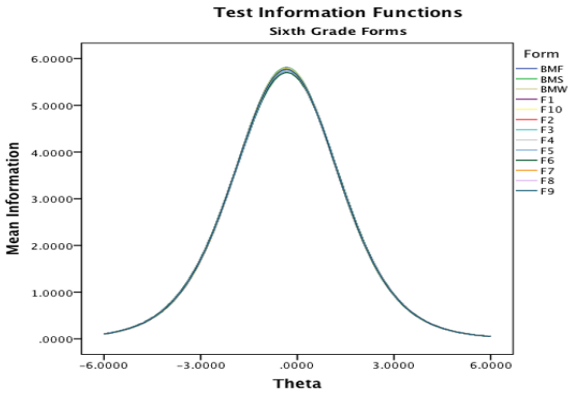

Figure 3. Example Test Characteristic Curves from Technical Report 1207

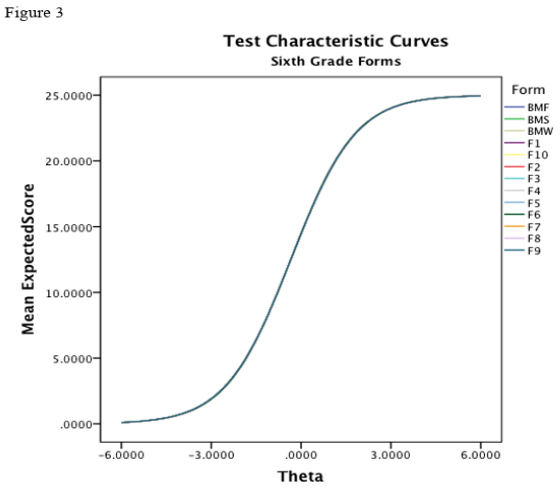

Table 53. Key Findings Summary from Technical Report 1207

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

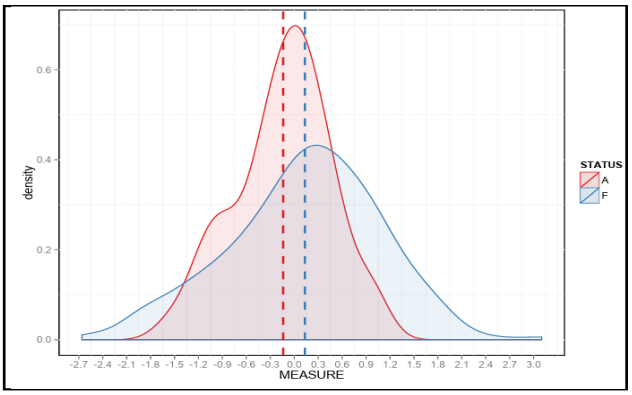

Figure 2. Distribution of Grade 6 item difficulties for anchored and freely estimated items.

Figure 6. Illustrative Box Plots of Item Difficulty from Technical Report 1313

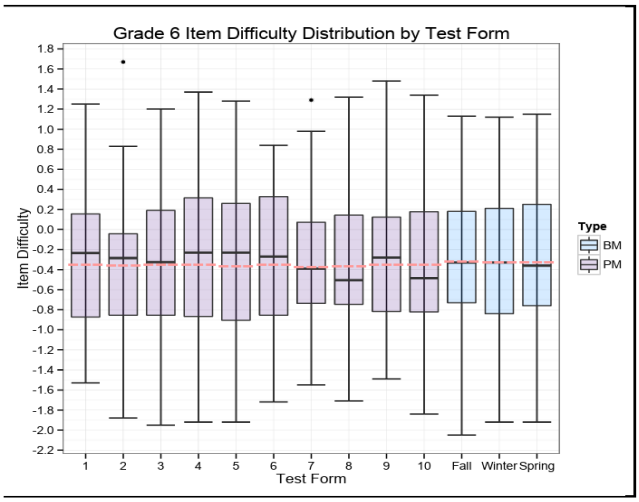

Figure 5. Distribution of item difficulties within Grade 6 test forms. Note that the solid black line within each boxplot represents the median item difficulty for the respective test form, while the hatched red line represents the mean.

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); subsequent 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 1408

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.