Technical Report 2603-RK8M

Reliability of 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). *Reliability of easyCBM® in Grades K–8: Mathematics (Technical Report # 2603-RK8M)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Table of Contents for 2603-RK8M_ReliabilityMath

[1] For individual technical reports, see https://brtprojects.org: BRT Tech Reports / Search here…

Abstract

This technical document summarizes primary studies conducted at Behavioral Research and Teaching (BRT) in the development and validation, specifically reliability, of easyCBM® Math measures for Grades K – 8. All studies present summaries of results and then illustrative findings with screen shots of exemplary tables. Note that all primary studies can be obtained at https://brtprojects.org. **Note**: All tables and figures in this summary are examples of those presented in full within the individual Technical Reports but are not exhaustive, just illustrative.

Definition and Types of Reliability

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

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

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

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

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

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

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

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

Summary of Technical Report Findings

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

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

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

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

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

Technical Report 0908 focused on a single grade (7) progress monitoring and benchmark form development and complements classical reliability evidence with item-bank calibration and form comparability evidence. The study piloted a large item pool (912 items) with a national sample of approximately 2,800 students and analyzed item performance using a one-parameter logistic Rasch model (Winsteps). Technical Report 0804 extended this analysis to Grades 1-8 to address alternate form reliability and reported strong inter-form correlations. The calibrated item bank supported the construction of multiple progress-monitoring forms and benchmark screeners with closely matched mean difficulties within focal-point groupings. The report also presented classical reliability evidence (including **Cronbach's alpha** for forms and **alternate-form** correlations), aligning the item-bank work with the practical need for equivalent forms in repeated measurement.

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

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

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

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

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

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

Methods

Participants included large, grade-specific samples drawn from a mid-sized school district in Oregon. Sample sizes ranged from approximately 1,100 to over 1,350 students per grade, with representation across general education, special education, economically disadvantaged, and historically low-achieving populations. Demographic data were collected during the spring of 2009 and used as an approximation of student characteristics during fall 2009 testing.

The mathematics screeners were computer-based assessments administered in group settings during the fall of the 2009 school year. Each grade-level test consisted of 48 dichotomously scored items aligned with National Council of Teachers of Mathematics (NCTM) Curriculum Focal Point Standards. Items were distributed evenly across three focal point domains per grade. Standardized administration procedures and automated scoring were used to reduce procedural error and ensure consistency across schools. Data were prepared by coding each item for response selection, correctness, and focal point domain. Internal consistency reliability was evaluated using Cronbach's alpha for total scores and for subtests corresponding to individual focal point domains. Descriptive statistics, including item means, variances, inter-item correlations, standard deviations, and standard errors of measurement, were calculated for each grade level.

Results

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

Table 1. Example of Overall Statistics from Technical Report 0915

Table 4
Overall Statistics.

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

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

Table 3

Inter-Item Correlations.

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

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

Table 3. Summary of Key Findings from Technical Report 0915

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

Reference:

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

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

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

Methods

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

Results

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

Table 4. Illustrative Reliability Indices from Technical Report 1006

Table 2
Case Processing Summary

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

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

Table 3
Reliability Statistics

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

Table 4
Summary Item Statistics

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

Table 5
Scale Statistics

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

Table 5. Additional Sample Reliability Indices from Technical Report 1006

Split-half Reliability: Fall

Table 6
Case Processing Summary

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

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

Table 7
Reliability Statistics

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

Table 6. Sample Item Statistics from Technical Report 1006

Table 8
Summary Item Statistics

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

Table 7. Sample Scale Statistics from Technical Report 1006

Table 9
Scale Statistics

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

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

Figure 3

Item fit – Grade K Fall Focal Point 1

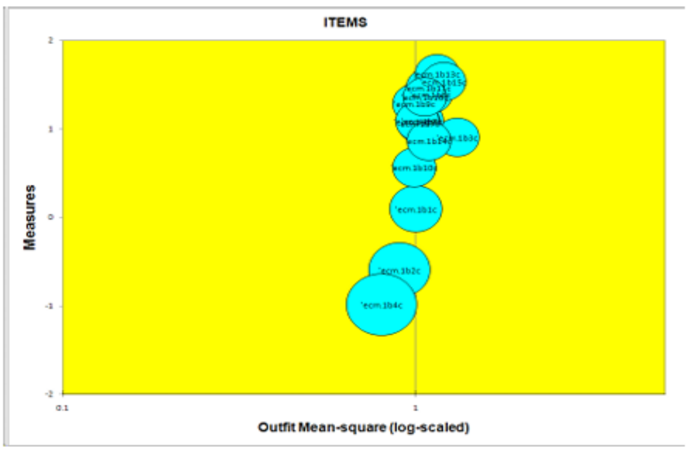

Table 8. Key Findings Summary for Technical Report 1006

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

Reference

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

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

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

Methods

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

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

Results

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

Table 9. Example Internal Reliability from Technical Report 1405

Table 5

Internal Reliability: CCSS Math

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

Table 10. Summary of Key Findings from Technical Report 1405

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

Reference

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

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

Methods

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

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

Results: Proficient Math

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

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

Reference

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

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

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

Methods

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

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

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

Results

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

Table 11. Illustrative Results for Technical Report 0908

Table 4

Grade 8 Test Form Point-Biserial Correlations

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

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

* $p < .05$

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

Table 25

Grade 6: Alternate Form Reliability Coefficients

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

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

Table 13. Summary of Results for Technical Report 0908

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

Reference

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

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

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

Methods

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

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

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

Results

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

Table 14. Illustrative Results for Technical Report 0804

Table 26.
Three Types of Problematic Items.

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

Table 15. Key Findings Summary for Technical Report 0804

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

Reference

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

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

Methods

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

Results

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

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

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

Table 16. Summary of Main Findings for Technical Report 1312

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

Reference

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

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

Methods

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

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

Results

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

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

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

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

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

Table 17. Summary of Main Findings for Technical Report 1804

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

Reference

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

Appendix A: Technical Report Table and Figure Titles

Appendix B: Guide to Spreadsheet Technical Report Value Displays

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

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

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

Technical Report References

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

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

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

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

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

Nese, J. F. T., Anderson, D., Irvin, P. S., & Alonzo, J. (2018). *In-Brief: Reliability of the slope of the easyCBM® math measures (Technical Report # 1804)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

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

Wray, K., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM CCSS math measures Grades K-8 (Technical Report # 1405)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.