aimsweb PLUS

Development Manual

ALWAYS LEARNING

PEARSON

Introduction

aimswebPlus® is an assessment, data management, and reporting system that provides national and local performance and growth norms for the screening and progress monitoring of math and reading skills for all students in Kindergarten through Grade 8. aimswebPlus uses two types of measures: *curriculum-based measures* (CBMs)—brief, timed measures of fluency on essential basic skills—and *standards-based assessments* (SBAs), which are comprehensive measures aligned to current learning standards.

By combining these two types of measures, aimswebPlus provides the data that schools need for program planning and evaluation and for tiered assessment (multi-tiered system of supports [MTSS], also known as response to intervention [RTI]). Furthermore, aimswebPlus data provides teachers with the information needed to differentiate instruction and determine who will benefit from intensive intervention.

In the following sections, an overview of aimswebPlus revision principles and goals is first presented, followed by a discussion of each stage of test development and a summary of each measure's content.

Revision Principles and Goals

The assessment approach of aimswebPlus is based on two principles. The first principle is to provide highly reliable and valid measurement of the automaticity of critical basic skills and short-term skill growth using CBM (i.e., fluency measures). Demonstrating automaticity of the skills measured in brief CBM tests is often a prerequisite for mastering more complex and higher-order skills. The second principle is the practical incorporation of content representing the breadth and depth of current grade-level expectations into assessments that can be completed within a single class period. Standards-based tests for Kindergarten and Grade 1 students enable the measurement of additional foundational skills shown to predict future performance; for Grades 2 through 8, these standards-based tests facilitate measurement of higher-order thinking skills and concepts.

With these principles in mind, development of aimswebPlus began with a review of published CBM research and consultations with CBM experts. Through this effort, published empirical studies of curriculum-based measures that provide predictive validity evidence as well as sensitivity to growth were identified. CBM expert consultants aided in the review and identification of the math and reading skills with the greatest measurement potential that were also highly valued by teachers. Additionally, current **aims**web measures were evaluated based on their psychometric properties (e.g., adequacy of floor/ceiling, reliability and validity data) and ease of administration and scoring.

Based on this research, revision goals were identified that sought to enhance:

- measurement of essential skills across the full range of abilities at each grade level,
- instructional planning data for students and classrooms,
- predictive capability, and
- alignment to current learning standards.

In addition to these goals, the following guiding ideals were established: keep what's working, measure what's important, and keep testing brief and developmentally appropriate. Adhering to these revision goals and guiding ideals, the final aimswebPlus measures were identified, revisions made to those measures carried over, and content written for new measures.

Test Development Stages

The implementation of the aforementioned principles, revision goals, and guiding ideals can be effective only if sound data defend them and guidance is provided for interpreting student and classroom scores in a way that directly informs instruction. In support, each aimswebPlus measure, revised or new, was put through multiple rounds of field testing, with refinements made as needed based on the results of this testing.

aimswebPlus field testing comprised the following research studies, with each study type spanning the Kindergarten through Grade 8 range:

- Pilot studies: multiple studies, 1,000+ students tested
- National tryout study: 14,000+ students tested
- National norms study: 16,000+ students tested
- Progress monitoring form equivalency studies: multiple studies, 15,000+ students tested

These new normative, reliability, and validity data were collected based on a representative sample of U.S. students. Additionally, the psychometric properties of all the aimswebPlus measures were evaluated to meet Pearson's and industry standards during the field testing process.

Analyses confirmed that using a multi-test battery approach provides stronger predictive data for student performance and risk status, as well as additional information about specific skills or knowledge areas that can be useful when interpreting student test scores. The combined information about automaticity of foundational skills and standards-based assessment of skills required for classroom success allow aimswebPlus to provide a more complete picture of what each student knows and can do.

The following sections provide further details of each development phase.

Content Development and Pilot Testing

Early Numeracy Curriculum-Based Measures

At the outset of this revision, the aimswebPlus research team chose to take a fresh look at the existing **aims**web math CBMs, evaluating them against the growing body of research on these types of math measures. Publications that provided empirical results on reliability, concurrent and predictive validity, and growth sensitivity were investigated to find evidence of brief math measures that met the following criteria:

- alternate form reliability of 0.80 or higher;
- concurrent and predictive validity with standardized math measures of 0.50 or higher; and
- seasonal or annual performance gains that exceed 0.50 points per week.

Our search identified about a dozen studies, conducted between 2000 and 2012, that provided results on all three criteria. Some of these studies included **aims**web math CBMs, and several introduced new CBMs. Most math CBMs in the identified studies met the reliability criteria; however, few met the validity criteria (Feldmann, 2012; Gersten et al., 2012; Lembke & Foegen, 2009; Methe, Begeny, & Leary, 2011) and even fewer met both the validity and growth criteria (Clarke, Baker, Smolkowski, & Chard, 2008; Floyd, Hojnoski, & Key, 2006; Gersten et al., 2012; Jordan, Kaplan, Oláh, & Locuniak, 2006; Lembke & Foegen, 2009; Lembke, Foegen, Whittaker, & Hampton, 2008; Methe, Begeny, & Leary, 2011; Seethaler & Fuchs, 2011).

Results of these studies showed that concurrent and predictive validity of single math CBM indicators were at best modest, with coefficients typically from the 0.30s to low 0.50s (Jordan, Kaplan, Oláh, & Locuniak, 2006; Lembke, Foegen, Whittaker, & Hampton, 2008; Methe, Begeny, & Leary, 2011). However, other studies show that the predictive validity can be boosted when scores from several CBMs are combined in a multiple regression model approach (Baglici, Codding, & Tyron; 2010; Martinez, Missall, Graney, Aricak, & Clarke, 2009).

The research also suggests that broad indicators that are ***not*** speeded—such as the Number Knowledge Test (NKT; McGraw-Hill Education, 2008)—tend to achieve higher validity with standardized math assessments. The NKT contains 45 items (testing time is approximately 25 minutes) and assesses a range of early numeracy skills, including counting, number recognition and comparison, story problems, and simple addition and subtraction. For example, Jordan, Kaplan, Oláh, and Locuniak (2006), following 378 Kindergarten students for 4 years, reported predictive validity coefficients with the Woodcock–Johnson® III (Woodcock, Shrank, McGrew, & Mather, 2005) math subtests in the low 0.70s 1 year later and correlations in the low to mid-0.60s 3 years later.

While several typical early numeracy measures exceed the growth sensitivity target (e.g., oral counting, number identification, counting), most such measures do not. Moreover, there appears to be a tradeoff between predictive validity and growth sensitivity (Lembke, Foegen, Whittaker, & Hampton, 2008). Sensitive measures such as oral counting tend to be less predictive.

In the absence of a set of early numeracy CBMs that are simultaneously highly predictive and sensitive to growth, aimswebPlus researchers chose to take a hybrid approach in which a brief, untimed, multi-item assessment is combined with high-quality fluency CBMs. By combining scores from both types of measures, the R&D team surmised that such a composite score would achieve high predictive validity and growth sensitivity while keeping administration under 10 minutes.

The aimswebPlus team also reviewed scientific research and position papers in math education to help identify a subset of math skills that are most essential for success in mathematics. Important findings of this literature review are described in the following paragraphs.

Several researchers note the importance of number sense development (Berch, 2005; Jordan, Glutting, Ramineni, & Watkins, 2010; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Markovitz & Sowder, 1988; McIntosh, Reys, & Reys, 1992). Number sense can be defined as *the ability to understand the meaning of numbers, define different relationships among numbers, recognize the relative size of numbers, and think flexibly with numbers* (National Council of Teachers of Mathematics, 1989). Robert Reys, a noted number sense researcher, also includes mental computation and estimation in his definition of number sense (McIntosh, Reys, & Reys, 1992).

The National Mathematics Advisory Panel (NMAP) Final Report (2008) defined three clusters of concepts and skills called *critical foundations of Algebra,* which included fluency with whole numbers, fluency with fractions, and certain aspects of geometry and measurement, especially analysis of the properties of two- and three-dimensional shapes using formulas to determine perimeter, area, volume, and surface area. Fluency included understanding various number systems, the relationship among number systems, and computation skills.

The Partnership for Assessment of Readiness for College and Careers (PARCC) created content frameworks for mathematics that organized standards into three clusters of emphasis: major, additional, and supporting (PARCC, 2014). The major clusters represent math topics that should be emphasized in each grade and form a continuum of knowledge and skills that puts students on track for success in college math. The framework for Kindergarten through Grade 2 draws heavily from empirical research, findings from international assessment systems, and recommendations from the National Association for the Education of Young Children.

The widely regarded publication *Adding it Up: Helping Children Learn Mathematics* (National Research Council, 2001) discusses the state of math achievement and education in the United States and how to improve achievement. This report cites evidence from the Third International Mathematics and Science Study, the National Assessment of Education Progress, and a range of scientific publications, with the authors concluding that students must have a strong foundation in number (whole numbers and fractions), number systems, and procedural fluency to be successful.

Finally, the National Governors Association, in partnership with the Council of Chief State School Officers, published the Common Core State Standards (CCSS; 2010). Drawing from decades of research and results from international math assessments, the CCSS for Mathematics adopted the approach of fewer standards with deeper understanding. These standards, describing the knowledge and skills that students should attain in each grade, have been adopted by nearly every state in the U.S., either in their entirety or with minor modifications. As such, it was important to reflect the curricular expectations defined by the CCSS for Kindergarten and Grade 1 in the new and revised content for aimswebPlus.

For example, the CCSS recommends emphasizing an understanding of addition and subtraction through 20 and an understanding of whole number and place value with two-digit numbers in Grade 1. By the end of first grade, the CCSS states that students should be fluent with addition and subtraction facts through 10. PARCC expanded on the CCSS by applying a framework, as previously mentioned, for focus areas and coherence across grades (PARCC, 2014).

Based on this research and instructional priorities, the skills listed below represent key areas through which successful students advance from Kindergarten through Grade 1. Note that, according to PARCC, the last three items listed are examples of skills that should receive in-depth focus in instruction and practice.

- Oral counting to 100, beginning from any number
- 1-to-1 counting (associating objects with number words)
- Cardinality (associating the last number in a sequence with the total number of objects in a set)
- Comparing the number of objects in two sets (more, less, and how many more)
- Comparing two numbers
- Fluently add and subtract within 10
- Counting to answer *How many?*
- How many more to make 10
- Quickly find 10 more or 10 less than a two-digit number

Oral counting is often included among early numeracy indicators and has been shown to correlate with broad measures of math computation and problem solving. In addition, this is one of the curriculum-based measures included in the **aims**web Tests of Early Numeracy. However, despite its popularity, some consider oral counting to be a rote memorization task that, much like reciting the ABCs, may provide little evidence of a student's conceptual understanding of numbers, quantities, and the number line (Griffin & Case, 1997). In a recent study, Martin, Cirino, Sharp, and Barnes (2014) evaluated the mediating effect of early literacy skills on the predictive validity of early numeracy CBMs. Nearly 200 students from a large school district completed oral counting and quantity discrimination CBMs in the spring of Kindergarten and the math fluency, applied problem solving, and story problem measures in the spring of Grade 1. Oral counting showed moderate predictive validity, with coefficients ranging from 0.33 to 0.53. Quantity discrimination showed somewhat stronger predictive validity, with coefficients ranging from 0.52 to 0.60. However, after controlling for verbal working memory and phonological awareness, oral counting became a non-significant predictor of fluency, computation, and story problems; conversely, quantity discrimination remained a significant predictor of each outcome measure.

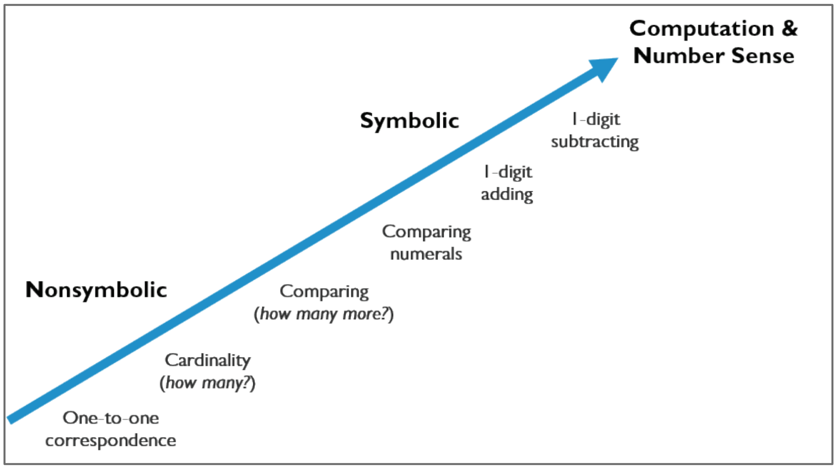
In another study with the same sample, Cirino evaluated the concurrent validity of 12 brief, Kindergarten early numeracy indicators, including oral counting, quantity discrimination, and missing number. The outcome measure was a 55-item, single-digit addition test. The 12 math measures were combined into five clusters: nonsymbolic comparison (e.g., comparing number of dots in two groups), symbolic comparison (e.g., quantity discrimination), symbolic labeling (e.g. number naming and missing number), rote counting (e.g., oral counting), and counting knowledge (e.g., one-to-one correspondence). Participants also completed spatial working memory tasks, as well as the rapid naming and phonological awareness tasks from the Comprehensive Test of Phonological Awareness (CTOPP; Wagner, Torgesen, & Rashotte, 1999). Simple correlations of the math clusters with single-digit addition were in the 0.60s and 0.70s for all but the nonsymbolic comparison ($r = 0.22$). After controlling for non-symbolic comparison, working memory, phonological awareness, and rapid naming, the relationship between rote counting and single-digit addition dropped to 0.04 and was nonsignificant. Symbolic comparison also became nonsignificant, whereas symbolic labeling and counting knowledge remained significant.

Counting and cardinality, numeral recognition, and an understanding of the relationship between number and quantity are the building blocks for math computation and problem solving. Research shows that fall Kindergarten performance on measures that assess a broad range of early numeracy is highly predictive of math achievement at the end of Grade 1 and moderately predictive of math achievement as late as the end of Grade 3 (Burland, 2011; Jordan, Kaplan, Ramineni, Locuniak, 2009; Mazzocco & Thompson, 2005; Purpura, Reid, Eiland, & Baroody; 2015). Math difficulties in late elementary grades have been linked to a failure to develop basic number sense and poor performance on one-to-one correspondence, number identification, and number line estimation (Locuniak & Jordan, 2008; Mazzocco & Thompson, 2005). Other studies show that math difficulties in elementary school can be traced to weaknesses in basic number competency (Gersten, Jordan, & Flojo, 2005; National Mathematics Advisory Panel, 2008).

Drawing from all of these sources, the aimswebPlus research team constructed a continuum (see Figure 1) that formed the basis for the selection and development of the new math curriculum-based measures. In this figure, the top right corner represents the part of the continuum assessed in Grades 2 through 8. It includes computational fluency with whole and rational numbers and number sense, both of which have been described as essential for success in Algebra.

The portion of the figure starting from the bottom left corner and moving toward the top right corner signifies essential prerequisite math skills. Before learners develop the ability to add and subtract using numerals, they must first develop an understanding of the correspondence between quantity and number and be able to compare quantities by answering questions about *how many* and *how many more*. This stage typically precedes the formal use of Arabic numerals, which is labeled as the nonsymbolic portion of this figure. Numeral recognition typically develops at this time as well. Once learners have a good grasp of counting and cardinality and can associate numerals with quantities, they are ready to begin formal instruction in addition with numerals, with instruction in subtraction following soon after.

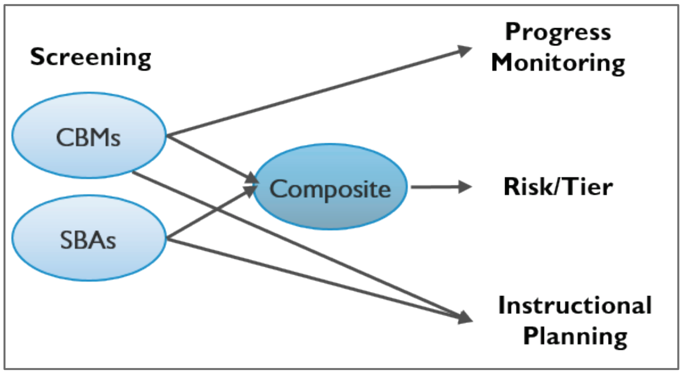
Figure 1 aimswebPlus Math Curriculum-Based Measure Continuum



As previously stated, for this revision the research team decided to include a brief, untimed, multi-item math indicator. This brief measure, called Concepts & Applications (CA), taps essential math concepts and skills and includes content from the domains of number, measurement, geometry, and data. Additionally, this measure includes items that assess reasoning and problem solving. CA was designed to assess the essential math standards that are not amenable to speeded measures and the higher-order thinking skills for which fluency is not an important goal.

Figure 2 illustrates the aimswebPlus hybrid model. For universal screening—also known as benchmarking—the system uses both CBMs and standards-based assessments (SBAs). Scores from the aimswebPlus math CBMs are combined with CA, an SBA, to form a composite. This composite score represents the broadest indicator of math achievement and is the best predictor of success; as such, this score is used to form instructional tiers. Scores from each measure can be used to isolate strengths areas and learning gaps, so all screening measures are used to facilitate instructional planning. The CBMs, because of their sensitivity to short-term learning, are also used for frequent progress monitoring.

Figure 2 aimswebPlus Hybrid Model



The following early numeracy curriculum-based measures were included in the aimswebPlus pilot studies, by grade.

Kindergarten:

- Quantity Comparison: The student indicates which of two boxes has more dots.
- Number Naming: The student names numbers (1–20).
- Quantity Total: The student indicates the total number of dots in one box or two boxes.
- Quantity Match: The student indicates how many more blue dots in a box are needed to match the number of red dots in a separate box.

Grade 1:

- Number Comparison: The student indicates which of two numbers is greater (1–99).
- Math Facts: The student mentally computes addition and subtraction facts through 10.
- Mental Computation: The student adds and subtracts 10 from two-digit numbers.

The early numeracy pilot studies, which began in the fall of 2012 and were completed in the winter of 2013, were conducted by Pearson's research and development team in conjunction with local schools. All math CBMs were individually administered to students by trained Pearson staff, with all students in the participating grades being tested at each site. These studies are further discussed in the following section.

Pilot Study 1

The purpose of this study was to evaluate the validity of the new and revised curriculum-based measures. Oral Counting and Missing Number were also included to provide additional context for interpreting the relative validity of the new measures. Note that Oral Counting rules were modified slightly for this study: Students counted until two sequence errors were made or the student skipped more than two numbers in sequence. Two criterion measures were used in this study: The KeyMath™–3 Diagnostic Assessment (KeyMath–3; Connolly, 2007) and NWEA's Measures of Academic Progress® (MAP®; 2009). The problem-solving subtest of the KeyMath–3 was individually administered to all participating students, while the MAP scores were obtained from the school. A 1-minute time limit was used for all CBMs except Math Facts and Mental Computation, each of which used a 2-minute limit. Testing occurred in October of 2012.

At the Kindergarten level, 37 students were administered the following measures:

- Oral Counting: The student counts aloud, beginning at 1 and up to a maximum of 100.
- Quantity Comparison: The student indicates which of two boxes has more dots. The number of dots in a set ranged from 5 to 50. Students were instructed to quickly respond to each item and were discouraged from counting individual dots to determine their answers.
- Number Naming: The student names numbers (1–20).

At the Grade 1 level, 35 students were administered:

- Number Comparison: The student indicates which of two numbers is greater (1–99).
- Missing Number: The student names the missing number in a three-number sequence (1–10).
- Math Facts: The student mentally computes addition and subtraction facts through 10.
- Mental Computation: The student adds and subtracts 10 from two-digit numbers.

Study results showed that participants in each grade performed at a level typical of for each grade. KeyMath–3 problem solving standard score means were 103 in Kindergarten and 101 in Grade 1, with a national average of 100 for this test. For the new and revised **aims**web measures, mean scores indicated that these CBMs were of appropriate difficulty for fall administration in each grade, with the exception of Mental Computation being too difficult. The correlations with the KeyMath–3 problem solving measure were moderate to strong across the entire set of measures. Correlations with MAP were even stronger. Study 1 results are found in Table 1, including means and *SD*s for each CBM and the concurrent validity coefficients with KeyMath–3 and MAP.

Table 1 Early Numeracy Pilot Study 1 Concurrent Validity Coefficients

Grade	Measure	Mean	*SD*	KeyMath–3 Applied Problem Solving	MAP
K	Oral Counting	38.0	17.3	0.60	0.64
K	Number Identification	28.3	12.1	0.63	0.74
K	Quantity Comparison	19.9	9.1	0.65	0.67
1	Quantity Discrimination	23.9	8.7	0.59	0.69
1	Missing Number	15.1	7.2	0.55	0.63
1	Math Facts	13.5	7.0	0.68	0.78
1	Mental Computation	2.9	3.7	0.71	0.69

Pilot Study 2

This study was conducted in November of 2012 with 30 Kindergarten students. The purpose of this study was to evaluate two new counting and cardinality CBMs (Quantity Total and Quantity Match), along with the Quantity Comparison measure that was tested in the first pilot study. Two forms of each CBM were administered to each student, along with the Numeration subtest of KeyMath–3. The administered CBMs are as follows:

- Quantity Total (Forms A and B): The student indicates the total number of dots in one box or two boxes, with totals ranging from 2 to 10.
- Quantity Comparison (Forms A and B): The student indicates which of two boxes has more dots (same form used in Study 1).
- Quantity Match (Forms A and B): The student indicates how many more blue dots in a box are needed to match the number of red dots in a separate box.

All students in this study were able to perform adequately on Quantity Comparison and Quantity Total; however, they were unable to perform the quantity-matching task. Poor performance on Quantity Match appeared to be a result of not understanding the instructions, and may have been further exacerbated by the generally low ability of this sample. To illustrate, these students' performance on the KeyMath–3 Numeration subtest was below the national average and Quantity Comparison scores were about 6 points lower than in Study 1. As a result of this evidence, the decision was made to revise the Quantity Match instructions and assess additional students on this measure later in the school year. Table 2 shows the means, *SD*s, concurrent validity coefficients, and alternate-form reliabilities for each CBM tested in Study 2. Note that validity coefficients and alternate-form reliabilities were at acceptable levels for both Quantity Comparison and Quantity Total.

Table 2 Early Numeracy Pilot Study 2 Concurrent Validity and Alternate-Form Reliability Coefficients

Grade	Measure	Mean	*SD*	KeyMath–3 Numeration	Reliability
K	Quantity Comparison	13.8	3.9	0.64	0.80
K	Quantity Total	11.0	3.1	0.54	0.83
K	Quantity Match	0.2	1.9	0.22	0.66

Pilot Study 3

This study was conducted with 41 Kindergarten students and 39 Grade 1 students in early February of 2013. The purpose of this study was to accumulate additional validity evidence for the new CBMs and to reevaluate the Quantity Match and Mental Computation measures, determining their appropriateness for use during winter and spring benchmarking.

In addition to the KeyMath–3 Numeration subtest, 41 Kindergarten students were administered the following measures:

- Number Naming: The student names numbers (1–20).
- Quantity Total: The student indicates the total number of dots in one box or two boxes.
- Quantity Match (Forms A and B): The student indicates how many more blue dots in a box are needed to match the number of red dots in a separate box.

Similarly, 39 Grade 1 students were administered the KeyMath–3 Numeration subtest and the following curriculum-based measures:

- Number Comparison: The student indicates which of two numbers is greater (1–99).
- Missing Number: The student names the missing number in a three-number sequence (1–10).
- Math Facts: The student mentally computes addition and subtraction facts through 10.
- Mental Computation (Forms A and B): The student adds and subtracts 10 from two-digit numbers.

Study results are shown in Table 3, including means, *SD*s, validity coefficients, and alternate-form reliabilities. Performance on KeyMath–3's Numeration subtest was slightly above average in Kindergarten (mean scale score = 10.5 using spring norms) and above average in Grade 1 (mean scale score = 12.0 using spring norms). The national average for the Numeration subtest is 10, with a standard deviation of 3.

Performance on the revised Quantity Match measure improved dramatically. With this new format, concurrent validity was high (*r* = 0.74) as was alternate form reliability (0.82). The new Mental Computation measure showed moderate concurrent validity and good reliability. The relatively low concurrent validity demonstrated for Number Naming and Number Comparison may reflect a decrease in their relationship with general math ability for students who have mastered these skills and have become fluent.

Table 3 Early Numeracy Pilot Study 3 Concurrent Validity and Alternate-Form Reliability Coefficients

Grade	Measure	Mean	*SD*	KeyMath–3 Numeration	Reliability
K	Number Naming	44.0	9.2	0.35	--
K	Quantity Total	15.4	3.3	0.48	--
K	Quantity Match	9.0	3.0	0.74	0.82
1	Number Comparison	29.8	4.4	0.44	--
1	Math Facts	17.8	4.8	0.60	--
1	Mental Computation	8.5	4.7	0.60	0.83
K	Number Naming	44.0	9.2	0.35	--

Math Curriculum-Based Measures

The original **aims**web includes two 10-minute, group-administered, paper/pencil math CBMs for Grades 2 through 8: Math Computation (M–COMP) and Math Concepts & Applications (M–CAP). All M–COMP items use a constructed response format in which students calculate answers to math operations items by hand, while M–CAP uses a combination of constructed response and selected response formats. Note that students are allowed to skip items in both of these measures. Scoring is done by hand and results are entered then into the **aims**web online system. For aimswebPlus, this system was replaced with computer-delivered administrations and the content was modified to improve alignment to current grade-level learning standards and to accommodate computer administration.

The replacement measure for M–COMP, called Mental Computation Fluency (MCF) requires students to solve one- and two-step math operations involving friendly (e.g., round) numbers. The problems do not require carrying or borrowing, and they can readily be solved through mental computation. Students select each answer from three response options; this selected-response format is efficient because it removes the need to type answers and can be designed to produce valid results. Moreover, the difference between the two formats—written (M–COMP) and mental computation (MCF)—was made less significant because many M–COMP are one- or two-step problems using friendly numbers that can be mentally solved without paper/pencil computation.

Computational fluency with whole and rational numbers is a critical foundational skill, but it may not provide broad enough representation of mathematics to be used on its own for progress monitoring. As a result, a second math CBM of number sense was developed to provide broader coverage and improve sensitivity to growth. Number sense can be defined as *the ability to understand the meaning of numbers, define different relationships among numbers, recognize the relative size of numbers, and think flexibly with numbers* (National Council of Teachers of Mathematics, 1989). Furthermore, mental computation and estimation are also specifically mentioned as components of number sense in research by McIntosh, Reys, and Reys (1992). In addition, the value of number sense in math education has been discussed extensively (Berch, 2005; Jordan, Glutting, Ramineni, & Watkins, 2010; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Markovitz & Sowder, 1988; McIntosh, Reys, & Reys, 1992; National Math Advisory Panel, 2008).

This new curriculum-based measure, Number Comparison Fluency–Triads (NCF–T), requires students to simultaneously compare three numbers represented as a triad and to determine whether the top number is closer in value to the bottom left number, the bottom right number, or is numerically exactly between the two numbers. This format reflects the concept of a number line, which was intentional and based on research showing the importance of forming and using a mental number line to achieve success in math. Like MCF, NCF–T uses friendly numbers and includes whole numbers, fractions, decimals, and exponents, as appropriate for each grade. Note that NCF–T represents an extension of the early numeracy CBMs Number Comparison Fluency–Pairs and Quantity Match Fluency, allowing grade-appropriate coverage of these concepts to span from Kindergarten to Grade 8.

Instead of another timed CBM, M–CAP was replaced by Concepts & Applications—an untimed, standards-based assessment. This decision was based on best assessment practices for measuring conceptual knowledge and problem solving skills. With such skills, accuracy is more relevant than speed. Allowing students the time they need to reason through problems and to give every student the time needed to attempt *every* item is consistent with educational research and policy; moreover, it is a fairer approach to assessing complex, higher-order thinking skills.

This shift to an untimed standards-based assessment also meant that aimswebPlus would not include frequent progress monitoring on conceptual knowledge and problem solving skills. To be effective for weekly progress monitoring, measures must be brief, reliable, and sensitive to growth over relatively short intervals of time. To be sensitive to growth, the skills assessed must develop fairly rapidly and lend themselves to brief administration.

Basic skills, like counting and computation, can develop rapidly with appropriate instruction and the tasks used to assess these skills can be brief. Complex skills—those that involve reasoning and analysis—develop more slowly and generally require tasks that take additional time to complete. Math word problems, for example, require reading, analyzing the knowns and unknowns, identifying a strategy, and solving the problem, which often involves computation. Compared to one- or two-step computation problems that typically can be solved in less than 15 seconds, a word problem may take 1 minute or more to solve. As a consequence, more testing time is required. This is exacerbated when the skill domains are broad because a brief measure can only sample a few items from each domain or skill area; thus, its ability to detect the growth that may occur in just one or two areas is diminished.

To further support the decision not to use CA for frequent progress monitoring, a simulation study was conducted to evaluate the level of growth sensitivity and length of time that is needed to detect real growth. This simulation used properties of M–CAP and M–COMP to generate the data. For growth to be considered real with a high degree of confidence, it should exceed the amount of error typically observed in the test score. The standard error of measurement (*SEM*) is used to quantify the amount of random error in a score. It is defined as:

$$SD \times \sqrt{1 - r_{xx}}$$

Where *SD* is the standard deviation and *rxx* is the reliability of the measure. The simulation used the *SEM* and average growth rates of M–CAP and M–COMP to generate typical growth patterns across 8, 12, 16, and 20 weeks. Table 4 shows the percentage of simulated cases in which the ratio of gain score (last score—first score) to *SEM* (gain/*SEM*) exceeded 1.28. This ratio provides a high level of confidence that the gain score is real. At this level, fewer than 10 percent of gains could be attributed to chance.

The first column in Table 4 shows results that could be expected on M–CAP for a low-performing student receiving intensive intervention. It uses an ROI growth rate that exceeds 85% of the students in the national norm sample in Grade 3. After 8 weeks of moderately high growth, true growth could be detected in just 28% of the sample. At 16 weeks, true growth was detected in 58% of the sample. In other words, score gains were within the margin of error for nearly half of the sample after 16 weeks.

The last column in Table 4 shows what could be expected using M–COMP with a struggling student. Growth was simulated using the ROI growth rate that exceeded 85% of the norm group in Grade 3. As shown in the table, M–COMP is much more sensitive to growth than M–CAP. By week 12, 95% of the scores exceeded the threshold for true growth. In light of these results, the research team determined that it would not be possible to revise M–CAP in a way that would make it suitable for frequent progress monitoring.

Table 4 Percentage of Instances That Growth Was Significantly Greater Than Zero

	Percentage of time growth differed from zero			
Weeks	ROI = 0.10 x *SEM*	ROI = 0.15 x *SEM*	ROI = 0.20 x *SEM*	ROI = 0.25 x *SEM*
8	28%	43%	59%	72%
12	42%	66%	85%	95%
16	58%	85%	97%	100%
20	74%	95%	100%	100%

Several pilot studies were conducted to evaluate the reliability and validity of the new mental computation and number sense CBMs developed for aimswebPlus. These pilot studies, which began in the fall of 2012 and were completed in the spring of 2013, were conducted by Pearson's research and development team in conjunction with local schools. For all studies, NWEA's MAP (2009) was used as the criterion measure to assess concurrent validity.

All measures were group administered to entire classrooms at each participating school. For each testing session, the proctor distributed paper test booklets, read general instructions, and walked student students through each measure's two practice items. Test booklets included a cover page to record the student's name, a page with practice items, and up to three pages of test items. After completing the practice items and answering any remaining questions, students were instructed to complete as many items as possible, noting that correct responses were more important than speed. Students were then told to turn to the first page of test items, and at this point timing began. When the time limit was reached, students were instructed to turn their text booklets over and wait for the proctor to collect them. If a student finished early, he/she raised his/her hand and the proctor collected the completed test booklet and recorded the elapsed time.

Pilot Study 1

In the first stage of development, the Number Sense task emphasized estimation. Students compared three numbers arranged in a triangular format and indicated which of the two lower numbers was closest in value to the top number, with test items presented in rows across the test page (see Figures 3 and 4). The Mental Computation task also emphasized estimation. For this task, students solved one-step operations mentally and then selected from two options the one closest in value to the answer.

For this study, Number Sense test forms contained 40 items and Mental Computation forms had 30 items, with unique forms constructed for each grade. For each grade level, the type and magnitude of numbers represented content that would have introduced in the prior grade. Students were given 3 minutes to complete the Number Sense form and 5 minutes to complete the Mental Computation form. Approximately 75 students per grade participated in this pilot study.

Figure 3 Math Pilot Study 1 Sample Number Sense Items, Grade 2

30	**300**	**80**	**90**	**600**
80 40	200 700	50 100	60 50	700 800

Figure 4 Math Pilot Study 1 Sample Mental Computation Items, Grade 5

25,000 + 20,000	**30,000 + 95,000**	**10,000 + 60,000**	**35,000 + 85,000**	**75,000 + 75,000**
30,000 50,000	200,000 140,000	50,000 100,000	130,000 100,000	120,000 140,000

Based on the results of this study, the following conclusions were drawn:

- Both measures were moderately correlated with MAP scores, with coefficients in the 50s and 60s.
- Combining scores from both measures raised validity coefficients by about 0.1.
- Guessing rates were moderate and corrections for guessing improved validity.
- The Mental Computation task of computing and then estimating to arrive at an answer was a bit difficult and somewhat confusing for students.
- Item content was a bit too difficult for a timed fluency measure.

These results lead to the changes to each measure and two additional pilot studies, which are discussed in the following sections.

Pilot Study 2

For this study, the Mental Computation item format was modified such that the student calculated and then selected the correct answer from three response options, thus eliminating the need to estimate (see Figure 5). The Number Sense format remained the same as in the first study. Study 2 included approximately 150 students in Grades 6 through 8 attending a suburban middle school. Each test form had four sections (two NS sections and two MC sections, alternating), with 25 items and a 2-minute time limit per section. Two test forms were developed, with one form presenting sections in the reverse order of the other form (i.e., NS1, MC1, NS2, MC2 and NS2, MC2, NS1, MC1).

Figure 5 Math Pilot Study 2 Sample Mental Computation Items, Grade 6

$\frac{3}{7} + \frac{6}{7}$	$\frac{9}{12} - \frac{4}{12}$	2×0.4	$15 + 5 \times 10$
$\frac{15}{7}$ $\frac{9}{7}$ $\frac{9}{14}$	$\frac{5}{12}$ $\frac{13}{12}$ $\frac{2}{12}$	1.8 0.6 0.8	65 200 30

Table 5 shows the sample sizes, means, *SD*s, alternate-form reliabilities, and validity coefficients for each measure, by grade. Reliability and validity were strong in Grades 6 and 8 and moderate to low in Grade 7. Note that the coefficients were based on scores corrected for guessing.

Based on the results of this study, the following conclusions were reached:

- Students completed the Number Sense items in about two-thirds the time it took to complete the Mental Computation items.
- The new Mental Computation format was clearer to students and seemed to improve validity.
- Correcting for guessing improved reliability and validity.
- Mental Computation items were a bit too difficult at each grade level.

Table 5 Math Pilot Study 2 Concurrent Validity and Alternate-Form Reliability Coefficients

Grade	Measure	*N*	Mean	*SD*	MAP	Reliability
6	Number Sense	49	16.7	13.4	0.65	0.89
	Mental Computation		9.8	7.2	0.71	0.77
	Composite		26.5	17.8	0.75	0.91
	MAP		216	16.9		
7	Number Sense	43	14.4	7.7	0.46	0.74
	Mental Computation		9.5	4.5	0.48	0.55
	Composite		23.9	10.6	0.52	0.79
	MAP		221	9.7		
8	Number Sense	48	26.9	8.0	0.61	0.73
	Mental Computation		17.8	9.0	0.72	0.79
	Composite		44.7	14.8	0.76	0.84
	MAP		236	15.8		

Pilot Study 3

The purpose of this study was to evaluate the validity of the new Mental Computation item format in early elementary grades and to evaluate a new format for the Number Sense items. In this new format, a third response option was added, with students now required to determine whether the top number was closer in value to the bottom left number, the bottom right number, or numerically exactly between (see Figure 6). Each response option was the correct answer for approximately one-third of the items in each form. For this new format, this measure's name was changed from Number Sense to Number Triads.

Figure 6 Math Pilot Study 3 Sample Number Triads Items, Grade 3

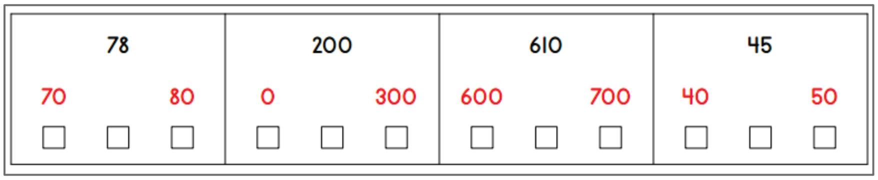


Table 6 shows the sample sizes, means, *SD*s, and concurrent validity coefficients, by grade and based on scores corrected for guessing. Results indicated that the new Number Triads (NT) format was appropriate for Grades 3 and above. While students in Grade 2 were able to accurately determine the answer when either the left or the right number was closest in value, these students had difficulty when the value was numerically in the middle. The overall difficulty was about right for both MC and NT content at each grade level, and validity was sufficiently high enough to recommend moving forward with these two measures and formats.

Table 6 Math Pilot Study 3 Concurrent Validity Coefficients

Grade	Measure	*N*	Mean	*SD*	MAP
2	Number Triads	52	9.8	4.7	0.52
	Mental Computation		14.9	5.3	0.59
	Composite		24.7	9.0	0.61
3	Number Triads	66	14.2	5.6	0.56
	Mental Computation		17.9	7.1	0.69
	Composite		32.1	11.6	0.69
4	Number Triads	50	13.9	6.1	0.66
	Mental Computation		20.0	8.3	0.62
	Composite		33.9	13.4	0.68
6	Number Triads	57	12.8	4.8	0.72
	Mental Computation		18.6	6.1	0.77
	Composite		31.4	9.9	0.83

The cumulative results of these studies allowed the research team to refine the test blueprints for the Math CBMs at each grade level, which were subsequently used to create each grade's benchmark and progress monitoring forms for these measures. Tables 7 and 8 provide item counts by topic area for each grade's Number Comparison Fluency–Triads and Mental Computation Fluency forms, respectively.

Table 7 Number Comparison Fluency–Triads Item Counts by Topic Area, by Grade

	Grade						
	2	3	4	5	6	7	8
2-digit comparisons	17	5	--	--	--	--	
3-digit comparisons	23	23	10	--	--	--	--
4-digit comparisons	--	12	20	15	--	--	--
5-digit comparisons	--	--	--	5	--	--	--
Fractions	--	--	10	15	24	16	14
Fractions & decimals	--	--	--	--	8	8	4
Decimals	--	--	--	5	8	8	4
Negatives	--	--	--	--	--	8	7
Scientific notation	--	--	--	--	--	--	7
Squared numbers	--	--	--	--	--	--	4
Item total	40	40	40	40	40	40	40

Table 8 Mental Computation Fluency Item Counts by Topic Area, by Grade

	Grade						
	2	3	4	5	6	7	8
Add and subtract multiples of 10 and 100	36	--	--	--	--	--	--
Add and subtract 2- and 3-digit numbers	6	--	--	--	--	--	--
Add and subtract 3- and 4-digit numbers	--	26	--	--	--	--	--
Multiply 1-digit with 2- or 3-digit multiples of 10	--	16	--	--	--	--	--
Multiply 1-digit with 2- or 3-digit multiples of 10 or 100	--	--	9	--	--	--	--
Divide 3-digit multiples of 10 by 1-digit numbers	--	--	6	--	--	--	--
Add and subtract 4- and 5-digit numbers	--	--	27	16	--	--	--
Multiply and divide multiples of 10	--	--	--	8	9	--	--
Add and subtract fractions with like denominators	--	--	--	6	3	--	--
Order of operations	--	--	--	6	12	12	--
Add and subtract fractions with unlike denominators	--	--	--	6	10	10	10
Multiply decimals (tenths) by whole numbers or decimals (tenths)	--	--	--	--	8	5	4
Divide whole numbers by fractions	--	--	--	--	--	4	6
Solve for *y* in 1- or 2-step equations	--	--	--	--	--	11	16
Add and subtract negative numbers	--	--	--	--	--	--	6
Item total	42	42	42	42	42	42	42

Early Literacy Curriculum-Based Measures

For Early Literacy (Kindergarten and Grade 1), the goal of content development was to measure each important skill area across the range of grades and seasons where that skill is most important. These skill areas reflected the current consensus of experts, as expressed in the Common Core State Standards for English Language Arts (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), and the recommendations of the National Reading Panel Report (National Institute of Child Health and Human Development, NIH, & DHHS,E 2000).

The skill areas of focus were: foundational skills (concepts of print, knowledge of letter names), vocabulary, phonological awareness, phonics (at the level of letters and groups of letters), and word reading (both isolated words and connected text). On the following page, Table 9 illustrates how the retained **aims**web measures map onto these skill areas and identifies the gaps to be filled with new measures.

Table 9 Target Blueprint of Early Literacy Skills, by Grade and Season

		Kindergarten			**Grade 1**		
Skill	**Measure**	**F**	**W**	**S**	**F**	**W**	**S**
Print Concepts	*Print Concepts* *(features of books and text)*	×					
	Letter Naming Fluency	×	×	×			
Phonological Awareness	*Initial Sounds* *(easier precursor to phoneme segmentation)*	×	×				
	Phoneme Segmentation		×	×	×		
Phonics	*Letter Word Sounds Fluency* *(sounds of letters and letter combinations)*		×	×	×		
Word Reading	*Word Reading Fluency* *(reading highly-decodable and high-frequency words)*			×	×	×	×
Text Fluency	*Oral Reading Fluency* *(oral reading of highly-decodable and standard text)*				×	×	×
Vocabulary	*Auditory Vocabulary* *(picture vocabulary)*	×	×	×	×	×	×

Note. Measure names in *italics* are new (brief summary of content focus in parentheses).

Of the original **aims**web measures, only Letter Naming Fluency remains unchanged. A significant design revision was made to Grade 1 Oral Reading Fluency so that this measure could be used at the beginning of first grade. The content and administration of Phoneme Segmentation were also revised, and new measures were developed to assess foundational reading skills (Print Concepts), vocabulary (Auditory Vocabulary), phonological awareness (Initial Sounds), phonics (Letter Word Sounds Fluency), and word reading (Word Reading Fluency).

The following sections detail the content development process for each aimswebPlus Early Literacy measure, including discussion of relevant pilot testing undertaken to guide decision making.

Print Concepts

For this measure, questions were developed that allow students entering Kindergarten to demonstrate—using a real storybook—the most fundamental understanding of print concepts: how a book is used, the basic elements of text (letters, words), and sequence (left-to-right, top-to-bottom, page-to-page) (Lonigan, Burgess, & Anthony, 2000). Not included were questions asking students to name letters or recognize words, skills that are assessed in other measures in the Early Literacy battery. In addition, words or concepts (e.g., *sentence*) typically not yet introduced in school were avoided. During the development of Print Concepts, emphasis was placed on brevity and ease of scoring, as well as an opportunity to establish rapport with young students likely unfamiliar with testing.

An initial pool of potential items was assembled based on a review of previous research and existing measures (Clay, 1993; Owocki, 2010). A 9-item pilot version was administered to 53 Kindergarten children in October 2012. The test was fairly easy (one-fourth of the children got 8 or 9 items correct), but about 10% missed more than half of the items, indicating that the measure was functioning properly as a screener. Pilot data led to selection of the final set of items according to a combination of content relevance and statistical properties (difficulty, and correlation with the total score).

Letter Naming Fluency

The content for Letter Naming Fluency is essentially the same as in the previous editions of **aims**web, with all letters included. However, for aimswebPlus, student pages are available only in a child-friendly, serif font to avoid any confusion between upper-case letter *i* and lower-case letter *L*.

Initial Sounds

The Initial Sounds measure was developed to assess phonological awareness with a task that was easier than Phoneme Segmentation and therefore appropriate for children entering Kindergarten. According to Adams (1990), rhyming is the earliest phonological awareness skill to be mastered, followed by sound matching (e.g., identifying words with a given initial or final sound), blending, segmenting, and manipulating (i.e., deleting or replacing sounds). Measures of sound matching, blending, segmenting, and manipulation all assess a common dimension, which is somewhat different from what is measured by rhyming tasks (Runge & Watkins, 2006).

Initially, three potential tasks were considered: initial sound matching, final sound matching, and blending. All three measures were evaluated in a pilot study. The pilot forms for Initial Sound and Final Sound had the same format: two stimulus pages, with nine objects pictured on each page. Each page included six receptive items (e.g., "Show me two pictures that start with *[sound]*") and three expressive items (e.g., examiner points to a picture and asks, "What sound does *[word]* start with?"). The reason for including both receptive and expressive items was to enable comparison with respect to psychometric quality and administrative ease. Each item type has its own obvious advantage: receptive items can be objectively scored, while expressive items do not require the student to know the name of a picture. Before the administration of each test page, the child was asked to name all the objects shown, and was told the name of any object for which they did not produce the standard name.

For this pilot study, the Initial Sound and Final Sound measures were given ***without*** a time limit, but the response time for each receptive item was recorded in order to gather information about whether timing would be useful in administration and scoring. The Blending measure required the student to say a single word by combining two syllables, or two to five phonemes, spoken by the examiner. A 90-second time limit was used with the 18 Blending items on the study test form.

All three measures were administered to 53 Kindergarten children in October 2012. Most of the children struggled with both Final Sound and Blending, not understanding what to do and often just repeating the sound spoken by the examiner. Therefore, these two tasks were dropped from further consideration.

Pilot results showed a very wide range of raw scores for the Initial Sound measure, with 26 students scoring very high (15–18 correct out of 18 items) and the remaining 27 students evenly spread across the range from 1 to 14 items correct. Reliability was promising, with a correlation of 0.82 between number-correct scores on the two test pages. Receptive and expressive items demonstrated similarly strong item-total correlations (medians of 0.72 and 0.73, respectively); additionally, the receptive and expressive item scores correlated 0.72 with each other. Based on these findings, the decision was made to retain both receptive and expressive items for the final measure. To reduce administration time, the number of pictures per page was reduced from nine to four, and the receptive task revised to have the student point to one picture (rather than two) starting with a given sound.

Performance on phonological awareness tests is affected by the students' familiarity with the words used for test items (Metsala, 1999); as such, tests and other educational materials that involve picture-naming words were consulted when selecting Initial Sound items. To ensure that nearly all Kindergarten students would recognize and name the included pictures in the same way, a pilot study was conducted in December 2012 in which 12 preschool children were asked to name a set of images. Of those included in this pilot study, 34 objects that ***all*** the participants named in the same way and 13 objects that all but one participant named in the same way were retained. These selected images/words started with 18 different consonant sounds. Art images for these items were then created specifically for use in aimswebPlus. Finally, the decision was made for Initial Sound to remain an untimed measure because (a) children just entering Kindergarten may not be accustomed to working quickly and accurately and (b) an untimed administration is easier to accurately administer when working with young children.

Phoneme Segmentation

The ability to perceive and say the separate phonemes of a spoken word is a skill that is highly predictive of future reading ability (Nation & Hulme, 1997; Torgesen, Wagner, & Rashotte, 1994; Vellutino & Scanlon, 1987; Yopp, 1988). This skill is part of the Common Core State Standards for English Language Arts in Kindergarten and Grade 1 and, as previously noted, it is one of the tasks included in the domain of Phonological Awareness.

The design of the aimswebPlus Phoneme Segmentation measure incorporates several simplifications to administration and scoring that addressed the feedback from **aims**web users and trainers, as well as advice from experts in the field of early literacy assessment. The primary goal of these revisions was to simplify administration, recording, and scoring to help examiners give the test consistently and accurately.

The most obvious change was to make this a nonspeeded (i.e., untimed) measure. Because rate-based measurement has distinct benefits, the decision to base scoring on accuracy alone rather than rate was made very carefully. Rate-based measures reflect the degree of automaticity of the skill, and automaticity can be valuable in the learning process because it frees cognitive resources to focus on developing new skills. In addition, rate-based measures usually have wider score ranges, which can be more sensitive to growth and therefore advantageous for monitoring progress.

Nevertheless, rate-based measurement is not always the most appropriate approach, for both practical and theoretical reasons. On the practical side, experience over the years with **aims**web Phoneme Segmentation Fluency indicated that it was a challenging measure to administer, record, and score accurately. The aimswebPlus development team heard concerns from customers and trainers about the difficulty of recording the details of correct sounds and blends when the student was responding quickly, as they tend to do with timed tests. This challenge was magnified when the examiner recorded responses on a computer or digital device because of the complexity of differentiating between errors/omissions and blends. Furthermore, the pressure on students to work quickly can interfere with saying the phonemes distinctly. The combination of these challenges posed by rate-based scoring creates the potential not only for scoring errors but also for significant inter-examiner variation. Users reported concerns that examiners who were more adept (or, perhaps, less conscientious) would present stimulus words at a more rapid pace than other examiners, thereby raising the students' scores.

From a theoretical perspective, one basic consideration was that phonological awareness is not an aspect of reading; rather, it is a *supporting* skill for learning to read. In other words, developing students' proficiency in breaking down words into sounds is not a primary goal of instruction. Curriculum-based measurement focuses on tracking growth in the targeted outcomes of instruction; as such, there is less imperative to design measures of supporting skills (e.g., phonological awareness) to be optimally sensitive for progress monitoring.

Phonological awareness is the understanding that spoken words are made up of a series of sounds (i.e., phonemes) and the ability to perceive those phonemes (Snow, Burns, & Griffin, 1998). An accuracy-based assessment sufficiently captures these skills. While a rate-based score captures additional information about automaticity, the incremental value of this information is not known with regards this task.

As a result, primarily for the practical reasons just described and supported by theoretical considerations, the aimswebPlus measure of phoneme segmentation ability was developed as an unspeeded assessment, scored on accuracy rather than rate. The benefits of a user-friendly and consistent administration approach were judged to be more important for this skill than information about automaticity.

Another change to this measure is the scoring of each item (word) as correct or incorrect according to whether the student produced a complete segmentation, rather than giving partial credit based on the number of correct sound segments (phonemes or sound blends). The reasons for this change were a combination of the same practical and theoretical considerations that applied to the question of speeded versus unspeeded administration. Users and trainers reported that recording and scoring blends accurately was challenging and this is especially true for computer-based administration, during which it is difficult to use the digital interface to accurately record both blends and correct individual sounds.

The content of Phoneme Segmentation is similar to that of its predecessor Phoneme Segmentation Fluency, with many of the earlier stimulus words retained. Consistent with the Common Core State Standards, all words are one syllable and about half are consonant–vowel–consonant (CVC) words. Furthermore, all included words have either three (73% of test items) or four (27% of test items) phonemes. Approximately one-third include digraphs and one-fourth contain consonant blends. Lastly, only highly familiar words were used, as research shows that performance on phonological awareness tasks is affected by familiarity with the words used (Metsala, 1999).

Letter Word Sounds Fluency

LWSF is the aimswebPlus measure of automaticity in phonics; in other words, the ability to produce the sounds of letters and combine those sounds into the sounds of letter strings (i.e., syllables and words). The original **aims**web Tests of Early Literacy include two phonics measures: Letter Sound Fluency (LSF) and Nonsense Word Fluency (NWF). Researchers determined there to be two reasons for updating how to assess phonics ability. First, to broaden the range of measurement by combining the two original measures; that is, to have a measure that transitions from letter sounds to the sounds of letter groups. Spanning a longer period of development would provide the measure with greater utility for progress monitoring because it would include content that is easy enough for early learners (early-to-mid Kindergarten) ***and*** hard enough for more advanced learners who can combine letter sounds. Second, to move away from nonsense words because customers have demonstrated that a significant proportion of educators believe it is undesirable to ask students to say nonwords. Nevertheless, the task of saying these nonsense words is valuable because it provides a purer measure of decoding (phonics) ability by eliminating the influence of prior word knowledge.

The first version of this new measure started with individual letters, transitioned to consonant blends, and concluded with CVC syllables (nonwords). The letters included were ones that could be pronounced correctly, as an initial sound, by more than 85% of children aged 5 years 0 months, according to the Goldman-Fristoe Test of Articulation developmental norms (Goldman & Fristoe, 2000). The blends, which did not include digraphs, were selected from the Direct Instruction text (Carnine, Silbert, Kame'enui, & Tarver 2010) and The Book of Lists (Fry & Kress, 2006). The syllables, selected from **aims**web Nonsense Word Fluency forms, were ones that could be found in real words. Letters and blends that are difficult for English language learners (e.g., *j* and *z*) were excluded, as were *c* and *g*. Pilot testing in November 2012 with a small group of Grade 1 students quickly showed that blend items—such as *pl*, *br*, or *mp*— were not effective for measurement. Students tended to insert vowels, and the blend items were more difficult than the CVC items. In hindsight, this made sense because blends tend to be taught after CVC words.

The second version of LWSF as designed to be more aligned with the process of learning to decode CVC words. Again, the initial items were individual letters, thus giving students who are not yet able to say the sounds of letter groups introductory content that they can handle. The remaining items consisted of CVC words, both in pieces and as whole words: first the initial consonant, then the rime (the VC syllable, usually not a word), and finally the full CVC word. Credit was given for each of these three units.

This version was piloted in December 2012 with 26 Grade 1 students. Each pilot form included ten letters and ten word-part items. Two orientations of the word-part items were evaluated: horizontal (with C, VC, and CVC on the same row) and vertical (with the three parts stacked from C at the top to CVC at the bottom). About half of the participating students took each orientation. Although students performed correctly on both orientations of the word-part items, they responded much faster using the horizontal orientation, without clear separation between the parts. Because this defeats the purpose of assessing the ability to say the sounds of various letter combinations, the vertical orientation was chosen for further development.

Because LWSF was designed to serve the assessment functions of both Letter Sound Fluency and Nonsense Word Fluency, it needed to include enough single-letter items to provide a reliable score for students who are not yet able to combine letter sounds, as well as enough word-part items to challenge the more advanced students. The optimal number of single-letter items is the number that can be completed in 1 minute by students who are just beginning to be able to say the sounds of letter combinations. Students who are at this level or below will encounter only single letters within the 1-minute time limit. Thus, for those students, LWSF functions like Letter Sound Fluency. As students develop above this level, their letter-sound automaticity increases and they reach an increasing number of word-part items in the same 1-minute time limit. This allows the measure to also assess the higher levels of phonics ability.

The number of single-letter items included per form was empirically determined via analysis of the rate of letter-sound production shown by students who are beginning to be able to say the sounds of two-letter and three-letter groups. Students who say letter sounds at that rate or faster are able to handle the letter-group items, although their performance may be slow.

Scoring of the letter-group items is easier than the scoring method used on Nonsense Word Fluency. In LWSF, each two-letter syllable and three-letter word is scored as correct or incorrect, because the purpose of this part of the measure is to see whether the student can synthesize the letter sounds into a whole. There are ample single-letter items in the measure, so credit does not need to be given for saying the sounds of individual letters within a group of letters, as is done on NWF.

In its final format, each LWSF form includes all five vowel sounds in each word box row and samples across a range of initial consonant sounds, excluding: /c/, /j/, /L/, q, x, and /y/.

Word Reading Fluency

WRF was developed as a potential alternative to the modified Oral Reading Fluency measure at Grade 1 (described in the following section) and as a measure that could be used at the end of Kindergarten to assess actual reading, which the original **aims**web does not do until Grade 1. Research suggests that, among early readers, fluency in naming sight words is as good as text-reading fluency (ORF) at predicting reading competence (Clemens, Shapiro, & Thoemmes, 2011; Fuchs, Fuchs, & Compton 2004). As such, WRF can be thought of as an intermediate step between phonetic decoding (LWSF) and reading connected text (ORF).

WRF consists of high-frequency (i.e., sight) words that are some of the first words students learn to read, with many of them being highly decodable. These words were selected primarily according to the Zeno word list (Zeno, Ivens, Millard, & Duvvuri, 1995), with each WRF word found within the first 250 words of this commonly used resource. In addition, about 70% of the WRF words are included in the Dolch 200 and Dolch 95 (nouns) lists, and about 90% are included on the Fry 300.

All WRF forms were built according to the same structure of decreasing word frequency. Of the 100 words on each form in this study, the first 10 were drawn from the 20 most frequent words according to the Zeno list; the next 10 were selected from the next 30 most frequent word; and each succeeding set of 20 was drawn from the next most frequent 50 words. Word order was controlled to avoid adjacent words that formed a meaningful phrase (e.g., *that man did*) or had some other pattern (e.g., three words in a row with the same initial sound).

Because WRF was constructed according to well-established information about word frequency and is administered in the same way as other word reading measures, no pilot testing was conducted. The length of each form is comparable to that of similar timed measures, and is a length that prevents most students from reading all of the form's words in less than the 1-minute time limit.

Oral Reading Fluency

A longstanding problem with the **aims**web R–CBM measure for Grade 1 has been that the stories are too difficult for many students during the fall, though they do work well in the winter and spring. For this reason, R–CBM is not part of the recommended **aims**web battery until winter of Grade 1. One of the goals of aimswebPlus was to develop an Oral Reading Fluency (ORF) measure that would be effective from fall through spring of Grade 1, in order to accurately assess the growth in text reading fluency across the entire school year.

Several years ago, Mark Shinn, author of the original **aims**web measures, had developed a set of Highly Decodable (HD) passages consisting of words that were phonetically regular, easy to decode or were high-frequency sight words, and incorporated repetition and rhythm to make these words even more accessible to beginning readers. Research indicated that the HD stories were accessible to almost all entering first graders and resulted in scores that were reliable and sensitive to growth (Shinn, 2012). However, these passages appeared to be inappropriately easy for students at the end of the school year.

The solution created for aimswebPlus was "progressive" ORF stories that combine the benefits of HD and standard text. Each Grade 1 R–CBM story was modified by rewriting the initial section (approximately 60 words) to use only highly decodable words (CVC, CCVC, or CVCC) or high-frequency sight words and creating a smooth transition into the previously existing text. Almost all of the HD words are one-syllable words.

The size of the initial HD section was determined empirically by finding the reading rate for HD words at which students became able to read the standard text reasonably accurately. In a pilot study in December of 2012 using Grade 1 stories with an initial 40-word HD section, 26 students had an average score of about 50 words correct per minute, which is higher than in the national norm sample for R–CBM. Approximately two-thirds of the students reached the standard-text section within the 1-minute time limit. Of those students, half attempted fewer than 15 standard-text words, and these students had a very high error rate on those words. The other half (i.e., those who attempted more than 15 standard-text words) had accuracy above 90% on those words. This result is similar to that found by Shinn (2008) in an earlier study of 291 Kindergarten students who took both an HD probe and a standard R–CBM probe. At HD rates from 0 to about 50 to 60 words correct per minute (WPM), the R–CBM rate was relatively flat at under 30 WPM; however, once students exceeded an HD rate of 60 WPM, their R–CBM rate began to accelerate in tandem with their HD rate.

Based on these findings, the HD section was lengthened to approximately 60 words per story so that students who are unable to read the standard text accurately are unlikely to reach that section. In the progressive ORF forms using this design, as the level of students' reading proficiency increases, they spend an increasing portion of testing time on standard text. This design enables the measure to differentiate among beginning readers in the fall ***and*** among relatively proficient readers in the spring, as well as showing fall–spring growth for all students.

Reading Curriculum-Based Measures

For Reading (Grades 2–8), the skill coverage again reflects recommendations from the CCSS and the National Reading Panel: reading comprehension, language (vocabulary), and fluency (both silent and oral). For this level, minor revisions were made to improve the content and interest level of the Oral Reading Fluency stories (formerly called R–CBM). Moreover, new measures were developed to assess vocabulary and reading comprehension skills, as well as an additional measure designed to assess silent reading rate with comprehension: Vocabulary, Reading Comprehension, and Silent Reading Fluency.

As previously mentioned, one of the goals when developing the aimswebPlus measures for Early Literacy and Reading was to assess each of the important reading-related skills at the grades and seasons where it is most important, according to the National Reading Panel Report (2000) and the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Table 10 shows how the retained **aims**web measure and the new measures map onto those skills, by grade and season.

Table 10 Target Blueprint of Reading Skills, by Grade and Season

		Grades 2–3	Grades 4–8
Skill	Measure	Fall/Winter/Spring	Fall/Winter/Spring
Text Fluency	Oral Reading Fluency (previously called R–CBM)	x	x
	Silent Reading Fluency *(comprehension-based silent reading rate)*		x
Reading Comprehension	*Reading Comprehension* *(passage comprehension)*	x	x
Vocabulary	*Vocabulary* *(reading vocabulary)*	x	x

Note. Measure names in *italics* are new (brief summary of content focus in parentheses).

Oral Reading Fluency

At Grades 2 through 8, the content for aimswebPlus ORF is essentially the same as the R–CBM probes in the previous **aims**web edition. However, stories have been selected for use in aimswebPlus benchmarking or progress monitoring based on their content appropriateness and equivalency data (e.g., evidence showing to what degree performance is similar across stories). Only a few stories are entirely new, while a few others have minor updates (e.g., to replace references to obsolete technology). As described by Howe and Shinn (2002), the original stories were written by educators with a close knowledge of the kinds of writing typically encountered by students at different grade levels. Teachers and paraprofessionals were trained to write passages that were appropriate in length (approximately 250 words for Grades 1 and 2, 300 words for Grade 3, and 350 words for Grades 4–8) and that contained the grade-appropriate number of syllables and sentences per 100 words, based on the Fry readability formula.

Silent Reading Fluency

One of the development goals of aimswebPlus was to create a reading fluency measure based on silent reading with comprehension. Students in the upper elementary grades and above must engage in comprehension-based silent reading in their academic subjects, making this is an important skill. Furthermore, oral reading is an increasingly artificial task as students get older.

Measuring comprehension-based silent reading rate in a way that is not only reliable and valid but also authentic and sensitive to growth is a considerable challenge, and many task types have been utilized (Hiebert, Samuels, & Rasinski, 2012; Meisinger, Dickens, & Tarar, 2015), including **aims**web Maze. For aimswebPlus, the development team worked to develop a task that would be more sensitive to growth (to better support progress monitoring), have a more explicit connection to comprehension, and be a more realistic task.

The basic concept was to present text to the student, allow the student as much time as they needed to read and comprehend the text, and then present a question that the student had to answer without referring back to the text. The score would be a reading rate (words per minute) based on the amount of time the student took to read the text, conditional on a reasonably high level of accuracy on the comprehension questions. The questions were designed to be very easy for students to answer ***if*** they understood the gist of the story segment. The purpose of the questions was to force the student to read with comprehension because the student could not predict what they would be asked. This task concept is similar to the one proposed by Nese et al. (2011).

This basic assessment model was refined through five pilot studies conducted in 2012 and 2013. These studies used one-on-one administration in which the student flipped through printed pages and the examiner timed the page turns, recorded the student's responses, and obtained the student's opinions about the tasks. These studies are briefly described in the following sections.

Pilot 1: Evaluation of potential item types. This study included 16 items of three types: sentence verification (true–false), a complex sentence followed by a three-option question, and story segments followed by three-option questions. Ten students in Grade 7 participated in this study. The WPM score had high odd-even reliability (0.89) and correlated 0.67 with the participants' MAP Reading scores. Students preferred the story-segment format and found the sentence verification items relatively confusing. In addition, teachers commented that the story-segment format was more like real reading. As such, the remaining pilot studies used the story-segment format.

Pilot 2: Evaluation of earliest grade. Eighteen students in January of Grade 2 took a measure that included three stories, each broken into four segments of about 20 to 40 words. The questions that followed each segment were very easy multiple-choice questions with three response options. At the end of each story, the examiner gave the student a token for each question they had answered correctly. Although the rate score was highly reliable (0.97), the task did not appear to be appropriate for this grade level because accuracy was quite low: p values ranged from 0.24 to 0.94, with a median of 0.65. Students tended to read aloud even though they were not asked to, and they focused on decoding. This finding implied that the task, as administered, would not be suitable for all entering third graders, and so further development assumed that Grade 4 would be the starting point.

Pilot 3: Effect of a time limit. This study addressed the question of whether students would read faster if they were working under a time limit. Each of the 26 Grade 4 students were administered the same two stories in the same sequence. For one story, students knew that they had an overall time limit of 90 seconds. The other story administered had no time limit. The sequence of administration conditions was counterbalanced. Results showed there was no difference in reading rate or question accuracy between the two conditions. Split-half reliability of WPM was 0.93. As a result, later versions of the task did not emphasize speed.

Pilot 4: Effect of showing a hint about the upcoming question. At this stage of the development cycle, it was becoming clear that increasing students' comprehension accuracy was important. This study evaluated the effect of providing a hint on the text-segment screen of what the following question would be about, in the hope that students would be sure that they understood that aspect of the text before moving on to the question screen and thus answer more accurately. Each of the 20 Grade 5 students were administered the same three stories in the same sequence, with one of the stories (chosen at random) containing hints. There was no effect of the hints on either accuracy or rate. Split-half reliability of WPM was 0.96.

Pilot 5: Effect of accuracy feedback. This study evaluated two other methods designed to increase comprehension accuracy. With the "reread" method, a student who answered incorrectly was asked to immediately reread the segment and then answer again. These students' rate score (WPM) was based on the total of initial and reread times. The "number-correct" method involved telling all students at the beginning of each story that they should try to answer at least three of the four questions correctly, as well as telling them at the end of each story how many of their answers were correct. This study involved 32 students in Grade 4, 22 students in Grade 5, and 31 students in Grade 7.

At each grade, all of the students read the same two stories, with about half of the sample assigned to the reread condition and the other half to the number-correct condition. Accuracy was shown to be the same with both methods. Rate scores using the two methods had similar reliabilities and similar correlations with Oral Reading Fluency and the Wisconsin state reading test, as shown in Table 11. Because it is easier to administer and emphasizes paying attention on the initial presentation of each text segment, the number-correct method was selected for use going forward.

Table 11 Reliability and Validity of WPM Using Reread and Number-Correct Methods, by Grade

	N		**Reliability**		**Correlation w/ORF**		**Correlation w/WKCE**	
Grade	Reread	# Correct	Reread	# Correct	Reread	# Correct	Reread	# Correct
4	17	15	0.81	0.91	0.84	0.73	0.84	0.54
5	14	8	0.89	0.91	0.69	0.86	0.70	0.87
7	16	15	0.78	0.84	0.58	0.69	0.57	0.69

Note. WKCE = Wisconsin Knowledge and Concepts Examination.

Standards-Based Measures

The section describes the development of the aimswebPlus standards-based assessments: Concepts & Applications (Early Numeracy and Math), Auditory Vocabulary (Early Literacy), Vocabulary (Reading), and Reading Comprehension (Reading). These assessments were designed to provide greater coverage of current grade-level math content standards and English Language Arts in Kindergarten through Grade 8.

Items and passages were developed by content experts with training in the development of high-quality standardized test items. Content developers were sensitive to the need to avoid content that is biased on the basis of gender, race/ethnicity, socioeconomic status, or religious group. Moreover, Pearson's item-writing guidelines specify the need to avoid such bias, to choose contexts that will be equally familiar or unfamiliar to all students regardless of ethnicity or gender, and to choose a variety of topics. In addition, care was taken when using names in items to choose names from a wide variety of ethnic backgrounds to facilitate diversity of representation.

Development of these new assessments employed the procedures that Pearson follows when creating or revising tests, including adhering to principles of universal design and the standards established by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, NCME, & APA; 1999); being sensitive to fairness; and collecting data to evaluate the items empirically. Each item and all corresponding test instructions were submitted to multiple rounds of internal review by highly trained Pearson content specialists and editorial staff. Reviewers evaluated content on the following qualities:

- construct and knowledge-and-skill alignment,
- factual accuracy,
- clarity and fairness,
- conformity with professionally accepted item-writing practices,
- balance by gender and race/ethnicity,
- adherence to the principles of universal design,
- appropriateness of visuals and language for the designated age range,
- item difficulty, and
- answer key validation.

All items were submitted to an out-of-house bias/sensitivity review that focused specifically on identifying items that might display bias. Our bias reviewers were members of the target populations and had prior experience conducting sensitivity reviews on standardized assessments. These external experts:

- reviewed each item for potential bias;
- reviewed diversity of background, cultural tradition, and viewpoints;
- evaluated changing roles and attitudes toward various groups;
- reviewed the role of language in setting and changing attitudes toward various groups; and
- edited for inappropriate language usage or stereotyping.

The principles of universal design are intended to enhance accessibility of an assessment's central constructs, reduce the need for additional accommodations, and improve validity, especially for English language learners (ELL) and students with disabilities. The National Center on Educational Outcomes (NCEO, 2002) at the University of Minnesota articulated the following set of principles of universally designed assessments:

- Inclusive assessment population
- Precisely defined constructs
- Accessible, non-biased items
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

To meet the conditions of an inclusive assessment population, development should include, where practicable, students from all of the unique groups for which the test will be used. Many steps were taken to ensure test content was appropriate and accessible to ELL students and students with mild to moderate sensory or physical disabilities. English language learners and students with disabilities were included in the item tryout and standardization stages of development. Furthermore, rigorous analyses, both qualitative and quantitative, were used to minimize bias.

When the NCEO report on universal design was published, the most common accommodations were extra time, individual or small group administration, and oral presentation of items. In the development of aimswebPlus' standards-based assessments, liberal time limits were employed, individual administration was used in Kindergarten and Grade 1, and audio presentation of all instructions as well as the vocabulary and math word problem items were available to ***all*** examinees during tryout and standardization. Additionally, test items and instructions are simple, clear, and intuitive and the verbal complexity was intentionally controlled to reflect below grade level expectations.

Concepts & Applications

aimswebPlus' Concepts & Applications (CA) assesses conceptual knowledge and problem solving skills that reflect current learning standards in mathematics. Content specifications were guided by the Common Core State Standards (CCSS) and supporting documentation. The CCSS define a sequence of topics and performances that describe what students should understand and be able to do by the end of each grade. The standards are organized into clusters, which are grouped into the domains listed in Table 12.

Table 12 CCSS Domains and Corresponding Grade Ranges

Domain	Grade range
Counting & Cardinality	K
Operations & Algebraic Thinking	K–5
Number & Operations in Base 10	K–5
Number & Operations—Fractions	3–5
Measurement & Data	K–5
Geometry	K–8
Ratios & Proportional Relationships	6–7
The Number System	6–8
Expressions & Equations	6–8
Functions	8
Statistic & Probability	6–8

Many CA items use graphs, tables, and real-world situations designed to assess a student's ability to apply conceptual knowledge, computational skills, and reasoning to solve problems. This measure is untimed so that students can work at their own pace to carefully analyze and solve every item in a given form. Many of the multi-step word problem items can take several minutes to solve; this, along with the goal of keeping testing time under 40 minutes in Grades 2 through 8 and under 15 minutes for Kindergarten and Grade 1, it was not possible to assess every standard found in the CCSS. As such, standards that research has shown to be most essential for success in formal Algebra were prioritized.

The National Math Advisory Panel Final Report (NMAP, 2008) defined three clusters of concepts and skills they called "critical foundations of Algebra" that included: fluency with whole numbers; fluency with fractions; and particular aspects of geometry and measurement, especially analysis of the properties of two- and three-dimensional shapes using formulas to determine perimeter, area, volume, and surface area. Fluency included understanding of various number systems, the relationship among number systems, and computation skills. Furthermore, the Partnership for Assessment of Readiness for College and Careers developed a content framework for mathematics that organized the standards into three clusters of emphasis: major, additional, and supporting (PARCC), 2014. The major clusters represent topics that should be emphasized in each grade and form a continuum of knowledge and skills that puts students on track for success in college-level math.

The CA blueprint provided representation of each CCSS domain with at least three items per benchmark form and additional items addressing the major clusters at each grade level. Tables 13 (Kindergarten and Grade 1) and 14 (Grades 2–8) show item counts by domain and grade for each CA benchmark form.

Table 13 Concepts & Applications Item Counts by Content Area, Kindergarten and Grade 1

	Kindergarten			**Grade 1**		
	F	W	S	F	W	S
Counting & Cardinality	7	7	7	--	--	--
Operations & Algebraic Thinking	9	9	8	12	12	12
Number & Operations: Base 10	1	1	1	5	5	5
Measurement & Data	4	4	4	5	5	5
Geometry	4	4	5	3	3	3
Item total	25	25	25	25	25	25

Table 14 Concepts & Applications Item Counts by Content Area, Grades 2 Through 8

	Grade 2			**Grade 3**			**Grade 4**			**Grade 5**			**Grade 6**			**Grade 7**			**Grade 8**		
	F	W	S	F	W	S	F	W	S	F	W	S	F	W	S	F	W	S	F	W	S
Operations & Algebraic Thinking	6	7	4	12	10	8	10	6	6	6	4	4	3	--	1	--	--	--	--	--	--
Expressions & Equations	--	--	--	--	--	--	--	--	--	--	--	--	6	8	7	10	8	6	17	15	13
Functions	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1	3	3
Number & Operations: Base 10	9	11	10	6	4	4	4	4	4	1	12	10	3	2	1	--	--	--	--	--	--
Number & Operations: Fractions	--	--	--	3	5	4	5	11	9	11	6	7	3	1	--	--	--	--	--	--	--
Number System	--	--	--	--	--	--	--	--	--	--	--	--	8	9	10	7	5	5	3	2	2
Ratios & Proportional Relationships	--	--	--	--	--	--	--	--	--	--	--	--	3	3	3	8	6	6	1	--	--
Measurement & Data	11	9	13	8	8	10	6	5	7	7	2	3	--	--	--	--	--	--	--	--	--
Statistics & Probability	--	--	--	--	--	--	--	--	--	--	--	--	--	3	3	1	5	6	2	6	6
Geometry	4	3	3	1	2	4	4	4	4	6	6	6	4	3	5	5	6	7	6	4	5
Item total	30	30	30	30	29	30	29	30	30	31	30	30	30	29	30	31	30	30	30	30	29

Concepts & Applications items were written by Pearson content experts and external math experts familiar with the CCSS. All items were designed to be multiple choice, with each item going through a rigorous review and key-checking process prior to field testing. Items were also reviewed for bias by an external expert panel. Items were then organized into a series of test forms that were administered to a large sample of students across the country, the data from which was then used to obtain statistical information about each item's difficulty, reliability, and validity. See the *National Tryout* section of this manual for additional details regarding the tryout studies and analysis procedures.

Auditory Vocabulary

In Kindergarten and Grade 1, students' understanding of spoken words is a good indicator of their language development (National Institutes of Health, 2010); as such, the Common Core State Standards includes Vocabulary Acquisition and Use as a language expectation for these young students. The Auditory Vocabulary measure was developed to provide this information to teachers, using the well-established picture-vocabulary format in which the student hears a word and selects from four pictures the one picture that illustrates the word.

The words for Auditory Vocabulary were chosen to be within the oral vocabulary of young children, according to research and published tests (Bracken, 2006b; Brownell, 2010; Dunn & Dunn, 2007; Kaufman & Kaufman, 2004; Williams, 2007). These words tend to be highly concrete and are familiar parts of the households, communities, and classrooms of most Kindergarten and Grade 1 students (e.g., lamp, grapes). Each form has words distributed across a range of categories that included food, animals, clothing, actions, descriptors, and other (e.g., vehicles, musical instruments).

The illustrations used in Auditory Vocabulary were pretested with a sample of children aged 4 and 5 years to ensure that they are easy to identify. The distractor images for each item depict things that do ***not*** closely resemble the target word in sound or meaning.

Because the Auditory Vocabulary measure was developed on the basis of prior research on oral vocabulary, it was not pilot tested and was first administered in the tryout phase of test development. Overall, the items were easier than expected. As a result, the most difficult words were chosen for use in both Kindergarten and Grade 1.

Note that this new measure was designed to be screener used to identify students who have vocabulary deficits, sometimes due to limited exposure to books, reading, and spoken English at home. A limited knowledge of basic vocabulary (including basic concepts such as outdoors or household items) can be a barrier to children's learning and overall success in school because it can interfere with their ability to understand instructions and/or to perform reading-related tasks, such as phonological awareness (Biemiller, 2003). Limited vocabulary may not be obvious to a teacher; as such, an objective assessment is useful for identifying children who may need targeted instruction.

Vocabulary

Another essential component of the aimswebPlus Reading assessment solution is a measure of reading vocabulary. In order to assess a student's lexicon, the Vocabulary measure was written such that students cannot infer the meaning of the target word from the item context. (Note that because this inferential skill is an important component of reading comprehension, items that measure understanding words within context are included in the aimswebPlus Reading Comprehension measure; knowing word meaning in isolation and in context addresses the general Common Core State Standards expectations for Vocabulary Acquisition and Use at these grade levels.) Each item consists of a stem, which is a brief phrase or sentence containing the target word, and four response options.

The target words and response options originated from retired editions of the *Stanford Achievement Test Series* (SAT). The aimswebPlus development team reviewed all target words and distractor options to identify and remove any words that had become obsolete (e.g., technology-related) or were inappropriate slang. SAT Vocabulary items were written such that the distractors are easier than the target words, so that errors are not the result of misunderstanding the distractor. Because the items were drawn from a carefully researched test, the aimswebPlus Vocabulary measure was not pilot tested and was first administered in the aimswebPlus Tryout.

Reading Comprehension

aimswebPlus was designed to be a comprehensive reading assessment solution; as such, an untimed reading comprehension measure is an essential component for students in Grades 2 through 8. Educators have long relied on tests that mirror the authentic passage-and-question tasks students are expected to master in the classroom; however, national and state academic standards are increasingly specific about the nature of these stories and questions. The introduction of online, group-administered testing for aimswebPlus was an opportunity to develop a reading comprehension measure that uses this traditional assessment format to assess a variety of comprehension skills in a way that is relatively brief and efficient.

Although the Reading Comprehension measure is not designed for progress monitoring over brief intervals, it follows the principles of curriculum-based measurement in that each benchmark form has comparable content and is equivalent in difficulty. This design supports the measurement of student growth over the school year.

At each grade level, half of a form's passages are literary (i.e., fiction) and half are informational (i.e., nonfiction). At Grades 3 through 8, the literary selections include poetry (either full poems or excerpts). These fiction and nonfiction passages were written by Pearson researchers—or by writers under their guidance—or were adapted from grade-appropriate Pearson publications, such as the *Stanford Achievement Test Series, Tenth Edition* (SAT–10; Pearson, 2007). Topics were selected to reflect the kinds of literature familiar in today's classrooms, with themes appealing to students of varying backgrounds and interests. Passages were carefully written to facilitate a range of questions—from finding answers literally in the words of a passage to drawing conclusions and connecting several broader ideas introduced in the text.

During the development of this measure, the appropriateness of passage length and text complexity were also considered. At each grade level, both somewhat shorter and longer texts were included. However, compared to other assessments, even the longer texts may be considered brief (i.e., none at Grade 2 are longer than about 200 words, while none at Grade 8 are longer than 425 words). This allows students to demonstrate the ability to read and understand texts of varying lengths and topics without requiring extraordinarily long test sessions. The complexity of text was estimated using Pearson's Reading Maturity Metric (Landauer, 2011), which generates a grade-level complexity score that incorporates an evaluation of the text structure, syntax, and vocabulary usage, including "word maturity" (i.e., how individual words gradually become known to have unique meanings depending on context [Landauer, 2011; Landauer, Kireyev, & Panaccione, 2011]).

The development of Reading Comprehension items was based in part on Common Core State Standards expectations at each grade level, adapted for the multiple-choice item format. Items were written to fall into the three main categories of Key Ideas, Craft & Structure, and Integration of Ideas (see Table 15). The items span a range of difficulty within each story, with students having the option to revisit previous stories and change responses as desired. The stories are ***not*** accompanied by illustrations to prevent any unintended clues that could influence how a student answers questions.

Table 15 Reading Comprehension Items, by Level and Category

	Grade 2			Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	F	W	S	F	W	S	F	W	S	F	W	S	F	W	S	F	W	S	F	W	S
R.L or I Stds 1–3: Key Ideas & Details	9	12	12	17	15	17	16	20	17	15	20	16	15	15	14	12	13	14	14	12	13
R.L or I Stds 4–6: Craft & Structure	6	5	2	4	5	5	6	4	5	5	2	5	7	6	7	9	8	7	6	11	8
R.L or I Stds 7–9: Integration of Knowledge & Ideas	0	0	0	0	0	0	1	0	1	3	1	1	2	3	3	3	3	3	3	0	1
R.L or I Std 10: Range of Reading	9	7	10	3	4	2	1	0	1	1	1	2	0	0	0	0	0	0	1	1	2
Item total	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24

National Tryout Studies

A series of studies were undertaken in the spring of 2013 to evaluate the new and revised CBMs and the new standards-based assessments (SBAs). The two overarching goals were as follows:

- Select CBM forms to be used as Fall, Winter, and Spring benchmark forms in standardization.
- Select items to build the benchmark SBA forms.

Equivalency studies were conducted to determine the relative difficulty of each alternate CBM form. Forms were organized into sets with counter-balancing of the order of administration across sets. Students completed all the forms in the set and the mean score of each form was used to determine its relative difficulty.

For the SBAs, the design of the forms and composition of the sample were carefully defined for the purpose of obtaining accurate data for:

- item calibration (i.e., estimating item difficulty, item fit and validity),
- differential item functioning, and
- vertical scaling.

A more complete description of these efforts are described in the following sections.

Table 16 shows the configuration of the test booklets by grade. Most booklet were comprised of two or more measures, and multiple versions of each measure were assessed (i.e., different forms). For example, the Kindergarten Early Literacy booklets consisted of either PS, LWSF, and AV ***or*** ORF, WRF, and LWSF. Forms A through D were used to evaluate item difficulty, while forms E through H were used to evaluate equivalency. The forms were designed for either item calibration or equivalency. Forms used in calibration were untimed; that is, students took as much time as needed to complete every item. Equivalency study forms used the same time limits as those defined in the final product.

Table 16 Test Booklet Configuration, by Content Area and Grade

Content area	Grade	Study	Forms	Measures included						
Early Numeracy	K	C, E	A–F	NNF (1)	QTF (1)	QMF (1)	CA	NNF (2)	QTF (2)	QMF (2)
	1	C, E	A–F	MF	CA	MC	--	--	--	--
Early Literacy	preK	C	A & B	IS	AV	--	--	--	--	--
	K	C	A–D	PS	LWSF	AV	--	--	--	--
	K	E	E–H	ORF	WRF	LWSF	--	--	--	--
	1	E	A–D	AV	ORF	WRF	--	--	--	--
Math	2–9	C	A–F	CA	--	--	--	--	--	--
	2–8	C	A	NCF–T	MCF	--	--	--	--	--
	2–8	E	A–D	NCF–T (1)	MCF (1)	NCF–T (2)	MCF (2)	--	--	--
Reading	2–3	C	A–F	VO	RC	--	--	--	--	--
	4–8	C & E	A –F	SRF	VO	RC	--	--	--	--

Note. In the Study column, C = calibration study and E = equivalency study.

Item Calibration Studies

The primary objective of the item calibration studies was to select the highest quality items for building the standardization forms. Item selection was completed on the basis of content, item difficulty, fit, and absence of bias. To ensure the best possible results, each test form included clear and brief instructions, followed by one or two practice items. Students were instructed to do their best to answer each item and, if necessary, to take their best guess. Students were given as much time as needed to complete each test.

In Kindergarten and Grade 1, all test items were individually administered to each student by a trained examiner. Items were presented on printed pages, with students responding orally or by pointing and examiners recording student responses on paper record forms. Results were hand entered into specially designed data entry forms, with each case entered by two people to ensure accuracy. Any differences in data entry were adjudicated by referring back to the record form. Although testing in Kindergarten and Grade 1 was straightforward and the test materials were carefully reviewed and tested, examiners were strongly encouraged to conduct two practice administrations before testing study participants.

In Grades 2 and above, tests were administered via computers ***except*** for Oral Reading Fluency, which was individually administered. Each test included one or two practice items and instructions. All instructions were presented using on-screen text and audio. In addition, Concepts & Applications and Vocabulary test items included audio as an option for every student. Very liberal administration time limits were used.

The pool of test items was too large to be administered in a single test form. To keep the testing burden reasonable, items were distributed across multiple test forms. Generally, test forms were designed to be similar in content coverage and difficulty. Some items appeared on more than one form for the purpose of statistically linking items. Table 17 summarizes the total pool of items in each calibration study and the number of items per form, with each measure's content and form design briefly described below.

Initial Sounds. Each form consisted 40 target words, 12 of which were used on both forms as linking items.

Phoneme Segmentation. Each form consisted of 25 words with three or four phonemes. Twelve words were used to link the forms.

Letter Word Sounds Fluency. Each form consisted of 10 letters and 25 CVC words. Seven words were used for linking across forms. The purpose of this measure's calibration study was to evaluate the difficulty of the CVC words. Letters were included as a warm-up activity and to determine whether the examinee had sufficient skills with letter sounds to attempt the more difficult word-sounds task.

Auditory Vocabulary. PreK forms consisted of 15 items, Kindergarten forms consisted of 50 items, and Grade 1 forms consisted of 45 items. Linking items were distributed across forms using a spiral method, wherein different linking items were assigned to each pair of forms and with each pair sharing 10 items.

Vocabulary. The pool of items was distributed across six forms such that each form was similar in difficulty. Forms were linked within grade and across adjacent grades. Linking items were distributed across forms using a spiral method, wherein different linking items were assigned to each pair of forms. Within grade, each pair of forms shared eight items, and about 12 items on each form were assigned to a form in the next higher grade. Thus, in Grades 3 through 8, each form included a set of items from the next lower grade. This design made it possible to concurrently analyze all of the items in the pool to obtain item difficulty estimates on a single scale and permitting direct comparisons across forms and grades.

Reading Comprehension. Forms design was similar to Vocabulary, except that the construction of forms was determined at the passage level. Each form had two unique passages, two passages linking forms within grade, and one passage linking forms across adjacent grades. Six test questions were developed for each passage.

Math Concepts & Applications (K and 1). The pool of items was distributed across six forms such that each form had an approximately equal number of items from each math domain being assessed. Twelve linking items appeared on each form. Forms were linked within, but not across, grades.

Math Concepts & Applications (2–8). The item pool was distributed across six forms such that each form had a similar number of items from each math domain being assessed. Within each grade, forms contained from 33 to 36 items, with a set of 20 common items assigned to all six forms. In each of Grades 3 through 8, students were randomly assigned to either an on-grade test form or a test form from the next lower grade. This randomly equivalent sample design was used to link items across grades.

Number Comparison Fluency–Triads and Mental Computation Fluency. One form of each math CBM was evaluated per grade. Although these measures are typically speeded, this administration was untimed so that all students would complete every item, with the intent of obtaining estimates of item difficulty. The forms were constructed using the same blueprint used in the equivalency studies.

Table 17 Total Item Pools and Items per Form, by Measure

Measure	Grade	Total item pool	Items per form
Initial Sounds	K	76	40
Phoneme Segmentation	K	98	25
Letter Word Sounds	K	72	25
Auditory Vocabulary	K & 1	160	30
Vocabulary	2–8	503	23–27
Reading Comprehension	2–8	738	20–30
Concepts & Applications	K & 1	188	24
Concepts & Applications	2–9	840	33–36
Number Comparison Fluency–Triads	2–8	40	40
Mental Computation Fluency	2–8	42	42

Sample

Over 50 schools representing a range of SES levels and urbanity (i.e., urban, suburban, rural) from 20 states participated. Participating schools chose which content area and grades to test, but were required to assess ***all*** students in the chosen grades except students with moderate to severe intellectual disabilities, students who were blind or deaf, or students who were English language learners with an English language proficiency score of 1 or 2. Many schools participated in both math and reading studies. Note that when race/ethnicity or English language learner status was not as critical, it was not collected.

Tables 18 through 20 describe the sample for each study. The sample size and demographic composition was chosen to obtain accurate estimates of item difficulty, item fit, and item validity, as well as to evaluate item fairness using a statistical technique called differential item functioning (DIF). DIF was evaluated by gender and race (i.e., White and African American; White and Hispanic/Latinx). The goal was to have 100 students from each of the three race/ethnicity categories complete each test item.

The purpose of the Math Grades 2 through 8 CBM calibration study was to estimate the difficulty of each item and task type. Item difficulty was based on the percentage of students getting the item correct (*p* value). Items with *p* values less than 0.60 were revised or replaced with easier items. Using relatively easy items is consistent with the purpose of CBMs to assess fluency of basic skills.

Table 18 Early Literacy Item Calibration Study Sample, by Grade

		Gender		Race/Ethnicity			
Grade	*N*	% F	% M	% B	% H	% O	% W
preK	491	51	49	33	31	3	33
K	413	50	50	33	32	2	33

Table 19 Concepts & Applications Item Calibration Study Sample, by Grade

		Gender		Race/Ethnicity				
Grade	*N*	% F	% M	% B	% H	% O	% W	% ELL
2	973	49	51	29	31	5	36	13
3	1836	51	49	28	31	6	35	9
4	1627	49	51	26	33	6	36	9
5	1511	47	53	29	30	5	36	6
6	1345	48	52	26	35	7	32	8
7	1250	47	53	37	30	6	27	15
8	822	50	50	23	30	7	40	6
9	895	49	51	38	31	3	28	8

Table 20 NCF–T and MCF Item Calibration Study Sample, by Grade

		Gender		Race/Ethnicity				
Grade	*N*	% F	% M	% B	% H	% O	% W	% ELL
2	235	46	54	20	14	3	63	1
3	247	53	47	25	10	2	63	3
4	306	50	50	28	21	3	47	4
5	285	46	54	25	21	4	51	4
6	198	47	53	16	53	3	29	11
7	271	48	52	40	32	8	20	20
8	190	56	44	43	21	12	25	8

Calibration

Calibration is statistical process that results in estimates of item difficulty and other item characteristics. The Rasch model was used to calibrate aimswebPlus items. With Rasch calibration, examinee ability and item difficulty are estimated on the same scale, such that when the examinee's ability is equal to an item's difficulty, the examinee has a 50% chance of answering the item correctly. An important consequence of using a common scale for item difficulty and person ability is that all test forms can be calibrated in one step. The shared items provide a statistical "anchor" that places the items on the same continuum. This means that items from different grades can also be analyzed concurrently.

The math test forms were not linked across grades; as such, the procedure for putting items onto a single scale spanning forms and grades required a second step. In the first step, all of the items within each grade were calibrated together. Each calibration was centered on item difficulty; that is, the mean difficulty at each grade was fixed to 0.0 on the Rasch scale. In the second step, the mean ability of the two groups in each grade (i.e., the group assigned to an on-grade test form and the group assigned to a test form in the next lower grade) was compared. Because the groups were randomly equivalent and item calibration was centered on item difficulty, a difference in the mean ability (lower grade–higher grade) can be attributed to the average difficulty difference between the test forms in adjacent grades. Subtracting this difference from the difficulty estimates in the lower grade places the items on a common scale. This procedure was repeated for each pair of adjacent grades starting with the Grade 5/4 pair and moving outward. Thus, all items were centered on the average difficulty in the spring of Grade 5.

Item Selection

Items were selected and placed on forms to be used for Fall, Winter, or Spring benchmark testing in standardization. Selection was an iterative process of identifying quality items, placing the items on forms, and estimating the internal consistency reliability of the new forms. To be considered for a final form, an item had to fit the Rasch model reasonably well and be largely free of bias. The final benchmark forms had to be similar in difficulty and content coverage. Thus, the process continued until reliability was optimized without compromising equivalence in difficulty and content coverage.

Fairness and Sensitivity

Fairness—that is, freedom from bias—is a critical characteristic of high quality assessment because it is an aspect of the test's validity. A test score is fair only to the extent that it has the same significance, or the same interpretation, regardless of the demographic characteristics of the examinee who obtained it. Thus, the evaluation of tests and test items for possible bias is a search for content that may cause the score interpretation to be inaccurate for examinees with particular demographic characteristics.

Both qualitative and quantitative methods to maximize the fairness were used. Qualitative evaluations of potential items occur first in the test development process. For this purpose, internal and external sensitivity reviews were conducted. For reading and math, a total of 12 experts participated in the bias review panel: four African American individuals, four Hispanic/Latinx individuals, and four Asian American individuals, with women and men represented in each demographic group.

Sensitivity reviewers performed the following analysis for all content (i.e., both text and art, as appropriate):

- Review diversity of background, cultural tradition, and viewpoints.
- Evaluate changing roles and attitudes toward various groups.
- Review the role of language in setting and changing attitudes toward various groups.
- Edit for inappropriate language usage or stereotyping.

The quantitative evaluation of fairness used the Mantel-Haenszel procedures, which examine differential item functioning (DIF) between reference (majority) and focal (minority) groups, after matching the groups on test scores. Statistical DIF is a flag for items that are in need of careful inspection to determine whether anything in their content could cause them to be less valid for some groups than for others. Ultimately, the decision of whether an item is biased relies on judgment.

Comparisons were made between men and women, White and African American individuals, White and Hispanic/Latinx individuals, and White and Asian American individuals. An item was considered potentially biased if its Chi-square was greater than what would normally be expected by chance. Items showing differences greater than chance were flagged for review and possible exclusion from the final test forms.

Item fit was evaluated using a Rasch-based mean square fit statistic called *infit*. The expected value of this statistic is 1.0. Values greater than 1.0 indicate unpredictability from the model, the cause of which can be due to multidimensionality, vagueness in the question the item is asking, or some other unmodeled property of the item. Items with infit greater than 1.2 were considered for removal. For Vocabulary and Reading Comprehension, poor fitting items were dropped. Because math is a multidimensional construct and infit can be influenced by multidimensionality, poor fitting items were dropped only if the reason for the poor fit could not be addressed through edits, the item also showed significant DIF, or the item's content or difficulty was redundant with another item.

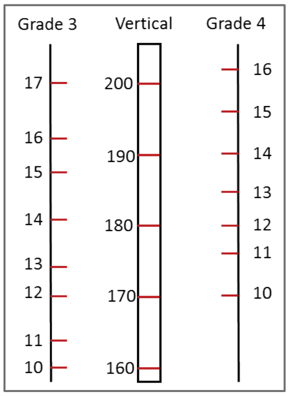
Vertical Score Scale

In Grades 2 through 8, Reading Comprehension, Vocabulary, and Concepts & Applications scores are reported on a vertical (or, developmental) scale. A vertical scale is a score scale that spans multiple grade levels by accounting for differences in item and form difficulty within grade and across grades. A student's vertical scale score is determined solely by his or her total number correct score. Thus, students with the same total score on the same test form will have the same vertical scale score. Students with the same vertical scale score who took different forms have the same overall ability on the construct (e.g., vocabulary).

The vertical scales range from about 80 to 300. The scale is centered on the performance of students in the spring of Grade 5, with the median (50th percentile) approximately equal to a vertical scale score of 200. In aimswebPlus, the vertical scale score is referred to as the *Growth Scale Value*. This name is used to reflect its primary purpose, which is to evaluate student growth across grade levels.

Figure 7 illustrates the relationship between the total correct scale and the vertical scale using Grades 3 and 4 Fall Vocabulary test forms. The two outside scales represent total scores for each grade. The rectangle in the middle depicts the vertical scale. The shift of the Grade 4 total scores upward relative to Grade 3 indicates that the Grade 4 form is a bit harder. Thus, a total correct score of 10 on the Grade 4 form is equivalent to a total correct score of 12 on the Grade 3 form. Likewise, a total correct score of 11 on the Grade 3 form corresponds to a lower vertical scale score (164) than a total correct score of 11 on the Grade 4 form (176). This figure also depicts the equal interval nature of the vertical scale. The spacing between each 10 point segment is always the same, indicating 10 points represents the same amount of ability growth at every point along the scale. The nature of this scale is similar to the popular Lexile™ score scale.

Figure 7 Vertical Scale Illustration



Equivalency Studies

The equivalency studies were designed to evaluate the difficulty of each aimswebPlus benchmark CBM form, as estimated by the mean score of the group completing the forms. The collected data were used to select a subset of the most equivalent forms for use in the Fall, Winter, and Spring benchmark testing in standardization.

To obtain an accurate estimate of the population mean, a minimum of 50 students completed each form. Each student completed multiple forms, which were grouped into sets. The number of forms in each set was carefully chosen to limit the testing burden to a level appropriate for the type of measurement. Each test form was assigned to at least two sets, with the order of administration counter-balanced across sets. Students were then randomly assigned to sets using a spiraling method such that, if there were four sets of measures, every fourth student would take Set 1. By counter-balancing the sets, it was possible to estimate the difficulty of the forms without the need to precisely balance the sample taking each set by all demographic variables. Table 21 shows the grade range and number of alternate forms for each study.

Table 21 Equivalency Study Measures, by Grade

		Grade		
Content Area	Measure(s)	Sample	Content	Forms per grade
Early Literacy	LWSF	K	K	6
Early Literacy	WRF	K & 1	1	6
Early Literacy	ORF	K & 1	1	6
Reading	SRF	4–8	4–8	6
Early Numeracy	NNF, QTF, QMF	K	K	6
Early Numeracy	NCF, MFF, MCF	1	1	6
Math	NCF–T, MCF	2–8	2–8	4

Note. Four of the Early Numeracy measures' names changed slightly between the equivalency studies and standardization, as follows: Quantity Match Fluency (QMF) became Quantity Difference Fluency (QDF), Number Comparison Fluency (NCF) became Number Comparison Fluency–Pairs (NCF–P), Math Facts Fluency (MFF) became Math Facts Fluency–1 Digit (MFF–1D), and Mental Computation Fluency (MCF) became Math Facts Fluency–Tens (MFF–T).

Letter-Word Sounds Fluency. Students say the sounds of letters and components of CVC words (C, CV, and CVC) for 1 minute. Four forms—each consisting of 30 letters and 15 CVC words—were assigned to four sets, with each form appearing on two sets and each set containing two forms. Each form appeared in the first positon in one set and the second position in the other set.

Word Reading Fluency. Students read individual words aloud for 1 minute. Each form consisted of 99 words found in reputable lists of high-frequency words. The examiner recorded errors and the last word attempted. Six forms were evaluated, with each form assigned to two of the three sets. Each set contained three forms, with forms appearing in different orders in each set.

Oral Reading Fluency. Students read each story aloud for 1 minute while an examiner recorded errors and the last word read. Six revised **aims**web Grade 1 stories were organized into four sets, each comprised of three stories. Each story appeared in two sets, in different positions. Results were used to pair stories for use as benchmarks forms in standardization. Note that stories were combined into pairs so that the average reading rate across the three pairs would be nearly equal.

Silent Reading Fluency. Students read four stories silently. Each story was separated into four, 40- to 45-word segments. Story segments were presented to students via online administration. After silently reading each segment, the student advanced to a screen with a simple multiple-choice question. Students were ***not*** allowed to return to the story segment after advancing to the corresponding question screen. The amount of time spent on each story segment was recorded and used to estimate a silent reading rate. To obtain a silent reading rate score, students had to correctly answer at least three of four questions. Six sets of four stories were included, with each story appearing in two or three sets.

Number Naming Fluency (NNF), Quantity Total Fluency (QTF), Quantity Match Fluency (QMF). These kindergarten Early Numeracy CBMs appeared in sets that also included Concepts & Applications (CA) items. Each of the six sets contained the three math CBMs, followed by 15 CA items, then followed by a second set of the three math CBMs (e.g. NNF1, QTF1, QMF1, CA, NNF2, QTF2, QMF2). Each math CBM had six forms, with each assigned to two sets and appearing before the CA items in one of the sets. Each math CBM had a 1-minute time limit; conversely, students were given as much time as needed to complete the 15 CA items. Individual administration was used, with each set taking about 15 minutes to complete.

Number Comparison Fluency (NCF), Math Facts Fluency (MFF), Mental Computation Fluency (MCF). These Grade 1 Early Numeracy CBMs appeared in six sets that also included CA items. The sets were designed and administered using the same format used in Kindergarten, as described above.

Number Comparison Fluency–Triads (NCF–T) & Mental Computation Fluency (MCF). In this study, four forms of each aimswebPlus Math CBM were evaluated. Each form was assigned to two of the four sets, with each set consisting of two NCF–T and two MCF forms. Each form appeared in the first position in one of the two sets (e.g., NCF–T1, MCF1, NCF–T2, MCF2). Each NCF–T form had a 3-minute time limit, while each MCF form had a 4-minute time limit. Students completed the multiple-choice items via computer administration and were instructed to do their best; however, if needed, students could skip items they did not understand. On each form, NCF–T items were presented four per screen and MCF items were presented two per screen. When time expired, students were automatically advanced to the instructions screen for the next form.

Tables 22 through 25 show the sample characteristics for the Early Literacy, Early Numeracy, Silent Reading Fluency, and Number Comparison Fluency–Triads and Mental Computation Fluency benchmark form equivalency studies, respectively.

Table 22 Early Literacy Study Sample, by Grade

		Gender	
Grade	*N*	% F	% M
K	282	53	47
1	356	49	51

Table 23 Early Numeracy Item Calibration and Equivalency Study Sample, by Grade

		Gender		Race/Ethnicity				
Grade	*N*	% F	% M	% B	% H	% O	% W	% ELL
K	757	48	52%	31	33	4	32	8
1	757	51	49	33	30	5	35	8

Table 24 Silent Reading Fluency Equivalency Study Sample, by Grade

		Gender		Race/Ethnicity				
Grade	*N*	% F	% M	% B	% H	% O	% W	% ELL
4	1145	50	50	33	22	5	41	6
5	1418	47	53	31	24	10	40	4
6	1063	49	51	29	26	7	38	7
7	1001	46	54	45	14	4	37	13
8	648	52	48	32	20	6	41	4

Table 25 NCF–T and MCF Equivalency Study Sample, by Grade

		Gender		Race/Ethnicity				
Grade	*N*	% F	% M	% B	% H	% O	% W	% ELL
2	134	53	47	16	34	13	37	19
3	205	53	47	8	22	14	56	11
4	209	52	48	7	22	11	60	11
5	143	50	50	24	13	15	48	3
6	132	47	53	0	6	17	77	11
7	158	46	54	48	13	1	38	13
8	154	44	56	44	27	2	27	6

Upon completion of the equivalency and item calibration studies, the aimswebPlus research team used the data and other information gathered to make final decisions about items and forms for the standardization and finalization phases of test development. The final aimswebPlus measures and corresponding normative data, system reports, and supporting materials represent the culmination of years of research, content development and refinement, studies, data collection, and data analysis. Please see the *aimswebPlus Technical Manual* (Pearson, 2016) for details regarding the standardization and finalization phases of aimswebPlus test development.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: The MIT Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999.) *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baglici, S. P., Codding, R., & Tyron, G. (2010). Extending the research on the tests of early numeracy: Longitudinal analyses over two school years. *Assessment for Effective Intervention, 35*(2), 89–102.

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38*(4), 333–339.

Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology, 24*(3–4), 323–335.

Bracken, B. A. (2006a). *Bracken Basic Concept Scale—Receptive* (3rd ed.). San Antonio, TX: Harcourt Assessment.

Bracken, B. A. (2006b). *Bracken Basic Concept Scale—Expressive*. San Antonio, TX: Harcourt Assessment.

Brownell, J. (2010). *Listening: Attitudes, principles, and skills* (4th Ed.), Boston, MA: Allyn & Bacon.

Burland, A. (2011). *Statistical relationship among number sense, computational fluency and Montana comprehensive assessment system* (Doctoral dissertation). University of Montana, Missoula, MT.

Carnine, D. W., Silbert, J., Kame'enui, E. J., & Tarver, S. G. (2010). Direct Instruction Reading, (5th Ed.), Boston, MA: Merrill.

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education, 29*(1), 46–57. doi:10.1177/0741932507309694

Clay, M. M. (1993). *An observation survey of early literacy achievement* (2nd ed.). Portsmouth, NH: Heinemann.

Clemens, N. H., Shapiro, E. S., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly, 26*, 231–244.

Connolly, A. J. (2007). *KeyMath 3: Diagnostic Assessment*. Bloomington, MN: Pearson.

Cunningham, A.E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6) 934–945.

Daniel, M. H., & Hiebert, F. (2016). Assessing silent reading fluency: The problem of comprehension. Manuscript in preparation.

Dickinson, D.K., & Snow, C. E. (1987). Interrelationships among prereading and oral language skills in kindergartners from two social classes. *Early Childhood Research Quarterly, 2,* 1–25.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Bloomington, MN: NCS Pearson, Inc.

Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, & DHHS. (2000). Report of the National Reading Panel: Teaching Children to Read: Reports of the Subgroups (00-4754). Washington, DC: U.S. Government Printing Office.

Feldmann, G. (2012). Early numeracy: technical adequacy of select kindergarten and first grade screening measures. (Doctoral Dissertation, University of Iowa, 2012). http://ir.uiowa.edu/etd/2869

Floyd, R. G., Hojnoski, R., & Key, J. (2006). Preliminary evidence of the technical adequacy of the preschool numeracy indicators. *School Psychology Review, 35*(4), 627–644.

Foegen, A., Lembke, E., Klein, K., Lind, L., & Jiban, C. L. (2008). Technical adequacy of early numeracy indicators: Exploring growth at three points in time. *Research Institute on Progress Monitoring,* 1–47.

Fry, E. B, & Kress, J. E. (2006). *The reading teacher's book of lists* (5th ed.). San Francisco, CA: Jossey-Bass.

Fuchs, L. S., Fuchs, D. & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*, 7–21.

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Council for Exceptional Children, 78*(4), 423–445.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38*(4), 293–304.

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation 2*. Bloomington, MN: Pearson.

Griffin, S., & Case, R. (1997). Rethinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education, 3*(1), 1–49.

Hiebert, E. H., Samuels, S. J., & Rasinski, T. V. (2012). Comprehension-based silent reading rates. What do we know? What do we need to know? *Literary Research and Instruction, 51*(2), 110–124. http://dx.doi.org/10.1080/19388071.2010.531887

Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features.* Bloomington, MN: Pearson.

Johnson, C. K., Daniel, M. H., & Bielinski, J. S. (2015). Better stories, better scores: Progressive passages for first grade readers. Poster presented at the annual convention of the National Association of School Psychologists, Orlando, FL.

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in Kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review, 39*(2), 181–195.

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice, 22*(1), 36–46.

Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77*(1), 153–175.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*(3), 850–867. doi:10.1037/a0014939.

Kaufman, A. S., & Kaufman, N.L. (2004). *Kaufman Assessment Battery for Children* (2nd ed.). Bloomington, MN: Pearson.

Landauer, T. K. (2011). Pearson's text complexity measure. Iowa City, IA: Pearson White Paper. Retrieved from http://www.pearsonassessments.com/textcomplexity

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1), 92–108.

Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research & Practice, 24*(1), 12–20.

Lembke, E., Foegen, A., Whittaker, T. A., & Hampton, D. (2008). Establishing technically adequate measures of progress in early numeracy. *Assessment for Effective Intervention, 33*(4), 206–214.

Lembke, E. S., & Stecker, P. M. (2007). Curriculum-based measurement in mathematics: An evidence-based formative assessment procedure. Portsmouth, NH: RMC Research Corporation, Center on Instruction, 1–28.

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities, 41*(5;5), 451–459.

Lonigan, C. J. (2003). Development and promotion of emergent literacy skills in preschool children at-risk of reading difficulties. In B. Foorman (Ed.), *Preventing and remediating reading difficulties: Bringing science to scale*, 23–50. Timonium, MD: York Press.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36*(5), 596–613. http://dx.doi.org/10.1037/0012–1649.36.5.596

Markovitz, Z., & Sowder, J. (1988). Mental computation and number sense. In M. J. Behr, C. B. Lacampagne, & M. M. Wheeler (Eds.), *Proceedings of the tenth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 58–64). DeKalb, IL: Northern Illinois University. (ERIC Document Reproduction Service No. ED 411 126).

Martin, N. A., & Brownell, R. (2011). *Expressive One-Word Picture Vocabulary Test* (4th ed.). Novato, CA: Academic Therapy Publications.

Martin, R. B., Cirino, P. T., Sharp, C., & Barnes, M. (2014). Number and counting skills in kindergarten as predictors of grade 1 mathematical skills. *Learning and Individual Differences, 34*. http://dx.doi.org/10.1016/j.lindif.2014.5.006

Martinez, R. S., Missall, K. N., Graney, S. B., Aricak, O. T., & Clarke, B. (2009). Technical adequacy of early numeracy curriculum-based measurement in Kindergarten. *Assessment for Effective Intervention*, *34*(2), 116–125. doi:10.1177/1534508408326204 http://aei.sagepub.com/cgi/content/abstract/34/2/116

Mazzocco, M. M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research and Practice, 20*(3), 142–155. doi:10.1111/j.1540–5826.2005.00129.x

McGraw-Hill Education. (2008). *Number knowledge test*. New York, NY: Author.

McIntosh, A., Reys, B. J., & Reys, R. E. (1992). A proposed framework for examining basic number sense. *For the Learning of Mathematics, 12*(3), 1–7.

Meisinger, E. (2012). Silent reading fluency using underlining: Evidence for an alternative method of assessment. *Psychology in the Schools 49*(6). doi:10.1002/pits.21613

Meisinger E., Dickens, R., & Tarar, J. (2015). Oral and silent reading fluency: Assessment to intervention. Paper presented at the annual meeting of the National Association of School Psychologists, Orlando, FL.

Methe, S. A., Begeny, J. C., & Leary, L. L. (2011). Development of conceptually focused early numeracy skill indicators. *Assessment for Effective Intervention, 36*(4), 230–242. doi: 10.1177/1534508411414150. http://ael.sagepub.com

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology, 91*, 3–19.

Nation, K., & Hulme, C. (1997). Phonemic segmentation, not onset-rime segmentation, predicts early reading and spelling skills. *Reading Research Quarterly, (32)*2, 154–167. doi:10.1598/RRQ.32.2.2

National Center on Educational Outcomes. (2002). *Universal design of assessments*. Retrieved from https://nceo.info/Assessments/universal_design

National Council of Teachers of Mathematics. (1989). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards*. Washington, DC: Authors.

National Institute of Child Health and Human Development *See* Eunice Kennedy Shriver National Institute of Child Health and Human Development

National Institutes of Health. (2010). *Speech and Language Developmental Milestones* (NIH Publication No. 00–4781).

National Mathematics Advisory Panel. (2008). *Final report.* Washington, DC: Author.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

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

Northwest Evaluation Association. (2009) *Measures of academic progress*. Portland, OR: Author.

Owocki, G. (2010). *The RTI daily planning book, K–6*. Portsmouth, HN: Heinemann.

Partnership for Assessment of Readiness for College and Careers. (2014). *Mathematics model content frameworks: Kindergarten through grade 2*. Washington, DC: Author.

Pearson. (2007). *Stanford achievement test series* (10th ed.). San Antonio, TX: Author.

Pearson. (2012). ***aims**web ROI Growth Norms Guide*. Bloomington, MN: Author.

Pearson. (2012). ***aims**web Technical Manual*. Bloomington, MN: Author.

Pearson. (2016). *aimswebPlus Technical Manual*. Bloomington, MN: Author.

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review, 44*, 41–59.

Runge, T. J., & Watkins, M. W. (2006). The structure of phonological awareness among kindergarten students. *School Psychology Review, 35*, 370–386.

Seethaler, P. M., & Fuchs, L. S. (2011). Using curriculum-based measurement to monitor kindergarteners' mathematics development. *Assessment for Effective Intervention, 36*. doi:10.1177/1534508411413566. http://aei.sagepub.com/content/36/4/219

Shinn, M. R. (2008). [Highly decodable reading passages and R–CBM results for Kindergartners]. Unpublished raw data.

Shinn, M. R. (2012). *Progress on early literacy universal screening and progress monitoring: Highly decodable reading passages*. Unpublished manuscript.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washngton, D.C.:National Academy Press.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology 38*(6), 934–947. doi:10.1037//0012–1649.38.6.934

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1994) (PS) Longitudinal studies of phonological processing and reading. *Journal of Learning Disabilities, 27*, 276–286.

Vellutino, F. R., & Scanlon, D. M. (1987) Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study. *Merrill-Palmer Quarterly (Wayne State University. Press), 33*(3), 321–363.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). Comprehensive Test of Phonological Processing (CTOPP). Boston, MA: Houghton Mifflin Harcourt.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development, 69*, 335–357.

Williams, K. T. (2007). *Expressive Vocabulary Test* (2nd ed.). Bloomington, MN: Pearson.

Woodcock, R. W., Shrank, F. A., McGrew, K. S., & Mather, N. (2005). Woodcock–Johnson III. Boston, MA: Houghton Mifflin Harcourt.

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly, 23*, 159–178.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). The educator's word frequency guide. Brewster, NY: Touchstone Applied Science Associates.

aimsweb PLUS

Pearson Executive Office 5601 Green Valley Drive Bloomington, MN 55437

800.627.7271 www.aimsweb.com

Produced in the United States of America.

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16