Pearson

aimswebPlus

Efficacy Research Report

April 2018

Contents

Introduction

In 2013, Pearson made a commitment to efficacy: to identify the outcomes that matter most to students and educators, and to have a greater impact on improving those outcomes. Our aspiration was to put the learner at the heart of the Pearson strategy; our goal was to help more learners, learn more.

A critical part of Pearson's portfolio is its Assessment business, which is really a services business supporting our customer requests by designing, building, administering, scoring, and reporting on test-taker performance in many different contexts (ranging from K-12 classrooms to the workplace) and for different purposes (ranging from supporting classroom instruction through ongoing progress monitoring to certifying fitness for employment in a given occupation). The people who take these tests are learners on a journey, similar to students who use our courseware products in the classroom to fulfill course requirements. In this case, however, the test is serving a slightly different function along this journey than would one of our digital courseware products. Taking a test is not a learning experience in and of itself, but rather, the scores and diagnostic information from these assessments may be used by instructors and others to make decisions about a learner's progress along their journey. Therefore, a measure of efficacy for assessments is not whether taking the test leads directly to higher achievement or passing the course, but whether the scores and other diagnostic information provide an accurate snapshot of what the learner knows and can do. In other words, the efficacy of an assessment is its fitness for a given purpose.

The fitness of an assessment for a given purpose, in turn, is defined by three primary qualities or attributes of test scores and their use: validity, reliability, and fairness. The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) have defined these attributes as follows:

— **Validity** is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Validity requires evidence that test scores can be interpreted as they are intended and can be appropriately used for a specific, defined purpose.
— **Reliability** is "the consistency of scores across replications of a testing procedure" (p. 33). Reliability requires evidence of the consistency of scores over time, across multiple forms of the assessment, and/or over multiple scorers.
— **Fairness** suggests that "scores have the same meaning for all individuals in the intended population" (p. 50). Fairness requires evidence that when assessments are administered as intended, items are not systematically biased against any particular group of test-takers and students are not hindered in demonstrating their skills by irrelevant barriers in the test administration procedures.

Given the longstanding role of the Joint Committee Standards as a source of guidance on best practices in the development and evaluation of tests and the role these standards play in the legal defensibility of assessment, Pearson has adopted these three attributes as the Assessment Quality Indicators on which we publicly report evidence underlying our assessment products. Each attribute is associated with a range of evidence types that are more or less relevant in a given context depending on the test's particular purpose and intended uses. For example, there are five commonly-accepted types of validity evidence that can be woven together to formulate an argument that a particular test can be interpreted as intended and used in a particular way, including evidence about how the assessment content was developed and how scores on the assessment relate to scores on other measures of the same kinds of knowledge and skills (AERA, APA, NCME, 2014). Similarly, there are different indices of reliability that can be provided, depending on the purpose and implementation of the test - when and how often it is administered, how it is scored, and how scores are reported. Such indices might include the average inter-item correlation or correlations between scores from different forms of the assessment, or across different times when the assessment is administered. Finally, fairness can also be supported by different types of evidence, including the results of analyses that specifically attempt to isolate items that appear to function differently for people in different subgroups (e.g., males versus females) and results from analyses of item content by specially formulated expert committees whose purpose is to identify potentially biasing content.

Pearson's assessment products are designed, built, and maintained over time by teams of subject matter experts and Ph.D. level research scientists trained in the science of assessment. These teams regularly (in some cases, annually) carry out studies to collect the kinds of validity, reliability, and fairness evidence described above, in accordance with the Joint Committee Standards. This evidence is typically consolidated and published in a technical manual or technical report that is updated with each new revision of the test. For that reason, much of the research we summarize on our assessment products has been completed internally and in many cases, we refer the interested reader to the technical manuals for full details of the research studies and associated evidence.

Special thanks

We want to thank all the customers, test takers, research institutions and organizations we have collaborated with to date. If you are interested in partnering with us on future efficacy research, have feedback or suggestions for how we can improve, or want to discuss your approach to using or researching our assessments, we would love to hear from you at efficacy@pearson.com.

Kate Edwards

Kate Edwards
Senior Vice President,
Efficacy and Research, Pearson
April 3 2018

Product summary

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 8th grade. 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.

Launched in the fall of 2016, aimswebPlus currently enjoys approximately 132,000 institutional users, reaching 2 million learners, resulting in over 24 million screening and progress monitoring assessments to date. Its key market is the United States. Intended for students in kindergarten through 8th grade, aimswebPlus provides universal screening for all students, as well as intensive progress monitoring for at-risk students. Key features of aimswebPlus include new standards-based content, new reports, digital and online administration of the assessments, and a new platform that optimizes usability, simplicity, scalability, and performance.

Assessment quality indicators

We define efficacy in assessment by three primary assessment quality criteria — validity, reliability, and fairness, as they apply to the main purpose of the assessment. The purpose of aimswebPlus is to diagnose students' need for intensive intervention in reading or mathematics and track their progress over time toward specific academic goals. The three assessment quality criteria discussed here are the extent to which the assessments allow educators to make sound interpretations of student skills (validity), the consistency and accuracy of scores (reliability) and the fairness of the assessments (AERA, APA, & NCME, 2014).

aimswebPlus research and development was intentionally designed to obtain evidence on item quality; score reliability; content validity; sensitivity to academic growth; and concurrent and predictive validity; as well as to ensure the assessments conform to professional standards and meet customer expectations. Collectively, these indicators, combined with the wealth of evidence on previous versions of aimsweb, support the interpretations and decisions for which aimswebPlus was intended:

— Accurate identification of academic risk in reading and mathematics
— Accurate measurement of progress toward end-of-grade performance targets
— Reliable and accurate measurement of progress toward individual performance goals
    for students needing frequent progress monitoring
— Diagnostic analysis of learning gaps and strengths to support tailored instruction

Assessment quality indicator 1:
Test scores can be interpreted as measures of critical basic skills in early literacy and numeracy, as well as higher order thinking skills and concepts in reading and mathematics, that can be used for identification of academic risk, progress monitoring, and tracking of progress toward individual student goals (validity).
A key aimswebPlus goal is to enable educators to make sound interpretations about student ability that support identification of students at academic risk, progress monitoring, and tailored instructions by providing measures that accurately capture literacy and numeracy skills, as well as profiles of relative strengths and weaknesses across different skill areas.

Assessment quality indicator 2:
Test scores are internally consistent and/or are consistent over multiple forms (reliability).
Another important goal of the aimswebPlus is to minimize errors in judgment and decision making by providing scores that are internally consistent and are consistent across different test forms.

Assessment quality indicator 3:
Test scores can be interpreted the same way for test-takers of different subgroups (fairness).
aimswebPlus also strives to provide scores that can be interpreted in the same way for all test-takers, regardless of gender or race/ethnicity. Fairness implies that when the assessments are administered as intended, items are not systematically biased against any particular group of test-takers and students are not hindered in demonstrating their skills by irrelevant barriers in the test administration procedures.

Foundational research

Overview of foundational research

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 Curriculum-Based Measurement 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, original aimsweb 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, development 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
— 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 development goals and guiding ideals, the final aimswebPlus measures were identified, revisions made to original aimsweb measures carried over, and content written for new measures.

The final aimswebPlus assessments consist of brief, standardized measures of skills in early literacy, reading, early numeracy, and mathematics that are extensively supported by both published research and Pearson-conducted studies. This research is summarized below, by content area (Early Numeracy, Math, Early Literacy, and Reading).

Early numeracy and math

At the outset, the aimswebPlus research team took a fresh look at the existing aimsweb 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
— Seasonal or annual performance gains that exceed 0.50 points per week

This search identified about a dozen studies, conducted between 2000 and 2012, that provided results on all three criteria. Some of these studies included aimsweb 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. For example, 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.
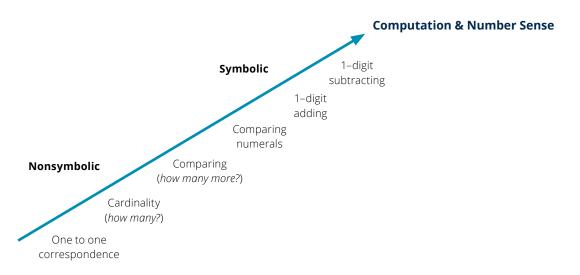
The Partnership for Assessment of Readiness for College and Careers (PARCC) created content frameworks for mathematics that organized standards into three clusters of emphasis: major, additional, and supporting (PARCC, 2014). The major clusters represent math topics that should be emphasized in each 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 sites 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.

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 figure 1, 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.

Figure 1: Continuum for math curriculum-based measures

The final aimswebPlus' Early Numeracy measures represent counting and cardinality, numeral recognition, and an understanding of the relationship between number and quantity, which 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).

For Grades 2 through 8, the original aimsweb 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). For aimswebPlus, content was modified to improve alignment to current grade-level learning standards, and to accommodate computer administration.

The first CBM, 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.

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 second CBM, called 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.

The third and final aimswebPlus Math measure is Concepts & Applications (CA) — an untimed, standards-based assessment (SBA). Including CA as a benchmark-only measure was a decision 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.

Early literacy and reading

For early literacy, 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, 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).

Of the original aimsweb 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).

Research underlying the development of the aimswebPlus Early Literacy measures includes:

— **Print concepts**: The basic elements of text (letters, words), and sequence (left-to-right, top-to-bottom, page-to-page) (Lonigan, Burgess, & Anthony, 2000).

— **Initial sounds**: Assessment of a phonological awareness with a task 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).

— **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). 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.

— **Letter word sounds fluency**: The ability to produce the sounds of letters and combine those sounds into the sounds of letter strings (i.e., syllables and words). Content included letters 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).

— **Word reading fluency**: An intermediate step between phonetic decoding (LWSF) and reading connected text (ORF). 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). 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.

— **Oral reading fluency**: Research indicated that the highly decodable (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. aimswebPlus' solution was to create "progressive" ORF stories that combine the benefits of HD and standard text.

For Grades 2 through 8, multiple studies have demonstrated that Oral Reading Fluency-type measures accurately predict general reading ability (Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Deno, 1982; Fuchs, Fuchs, & Maxwell, 1988; Shinn, Good, Knutson, Tilly, & Collins, 1992) and can be used to reliably determine a student's response to intervention within four to six weeks (Fuchs & Vaughn, 2005). A new reading fluency measure based on silent reading with comprehension — Silent Reading Fluency — was developed for students in the upper elementary grades and above because these older students must engage in comprehension-based silent reading in their academic subjects, making this an important skill. 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 aimsweb Maze. Using a task similar to the one proposed by Nese et al. (2011), the resulting measure is sensitive to growth (to better support progress monitoring), has an explicit connection to comprehension, and is a more realistic task for older students.

In addition, two standards-based measures are included in the aimswebPlus Reading battery: Vocabulary and Reading Comprehension. For Vocabulary, 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.

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. 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.

Items were written to fall into the three main categories of Key Ideas, Craft & Structure, and Integration of Ideas. The items span a range of difficulty within each story, with students having the option to revisit previous stories and change responses as desired. 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. The text complexity of each Reading Comprehension passage 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]).

Intended product implementation

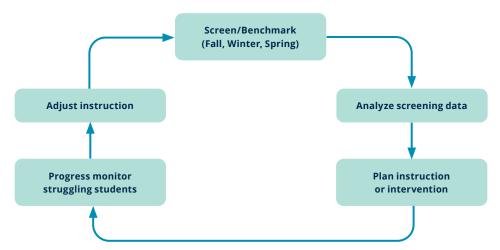## Overview of intended product implementation

aimswebPlus is intended to be implemented across all students in Kindergarten through Grade 8 for the purposes of identifying and monitoring students' foundational skills. As previously mentioned, 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]).

Multi-tiered system of supports are driven by data from three activities:

— Universal screening (i.e., benchmarking): Assessing all students to identify those who need additional instruction to succeed
— Progress monitoring: Tracking the effectiveness of instructional interventions
— Program evaluation: Evaluating the efficacy of core instruction in relation to student progress

As shown in Figure 2 below, all students should be screened three times per year. Based on the results of these screenings, students should be categorized into tiers according to instructional need and the intensity of the intervention needed. Students identified as needing intervention should be placed on a progress monitoring schedule according to the level of instructional intensity needed, with the most frequent monitoring schedule being once per week. Progress monitoring scores are then tracked over the school year and the teacher is able to adjust instruction or interventions as needed.

Figure 2: Screening process

A consistent approach to implementation support has been taken for both aimsweb and aimswebPlus. Customers of aimswebPlus are be provided with a variety of complimentary training options that include quick guides and video tutorials with embedded certification quizzes to ensure teachers understand the fundamental components of aimswebPlus. Additionally, customers will also be provided with an opportunity to purchase additional training and professional development options, which will include onsite, online, and remote workshops.

Onsite, online, remote, and ongoing professional development courses provide essential content, hands-on activities, and next steps in using aimswebPlus, as well as a solid understanding of the principles of assessment and their applications in the classroom and beyond for educators and administrators. Each service will include a customer survey to capture participants' feedback on the quality and value of the workshop. Results will be reviewed quarterly and updates to content will be scheduled to ensure customers are confident in their ability to utilize aimswebPlus for universal screening and progress monitoring in reading and mathematics.

Product research

The aimswebPlus team regularly carries out studies to collect the kinds of validity, reliability, and fairness evidence described above, in accordance with the Joint Committee Standards (AERA, APA, NCME, 2014). This evidence has been consolidated and published in a set of technical and development manuals, which are updated with each new revision of the test. For that reason, much of the aimswebPlus research we summarize in the following section has been completed internally. We encourage readers who are interested in the full details of the research studies and associated evidence to consult the aimswebPlus Technical Manual (Pearson, 2017).

Research studies

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

This new normative, reliability, and validity data was collected based on a representative sample of US 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.

Normative sample

Tables 1, 2 and 3 (below) present the demographic characteristics of the normative samples for the math and reading measures at each grade level. To be included in the norm sample, students had to complete the set of measures assigned to them (reading, math, or both). The percentage of students completing all assigned measures in all three seasons generally exceeded 90% in Math (Grades 2–8) and Early Literacy (Kindergarten and Grade 1). Approximately 85% of students completed all Early Numeracy measures (Kindergarten and Grade 1) and all Reading measures (Grades 2–8) in all three seasons. The dropout pattern was unrelated to demographic characteristics and was generally consistent across participating schools, with two exceptions. First, one school dropped out after the Winter testing session in the Early Numeracy study. Second, Oral Reading Fluency was administered on two separate platforms during Fall testing, which then had to be combined by matching various student characteristics, including student name. About 15% of the cases could not be matched and were excluded from the remaining data analyses.

Table 1: Demographic characteristics of the normative samples — Early Numeracy

| | | Sex | | | | Race/Ethnicity | | | | | | | | SES | | | |
| | | Female | | Male | | Black | | Hispanic | | Other | | White | | ELL | Low | Mod | High |
Subject	Grade	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	%	%	%	%
Early Numeracy	K	1000	50	1000	50	279	14	504	25	204	10	1013	51	10	32	32	36
Early Numeracy	I	1000	50	1000	50	265	13	506	25	201	10	1028	51	10	32	32	36
Early Literacy	K	1000	50	1000	50	279	14	504	25	204	10	1013	51	10	32	32	36
Early Literacy	I	1000	50	1000	50	265	13	506	25	201	10	1028	51	10	32	32	36

Table 2: Demographic characteristics of the normative samples — Math

| | | | Sex | | | | Race/Ethnicity | | | | | | | | SES | | | |
| | | | Female | | Male | | Black | | Hispanic | | Other | | White | | ELL | Low | Mod | High |
Subject	Grade	Measure	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	%	%	%	%
Math	2	NCF—T, MCF, CA	1500	50	1500	50	420	14	700	23	300	10	1580	53	10	30	40	30
Math	3	NCF—T, MCF, CA	1500	50	1500	50	420	14	700	23	292	10	1580	53	10	30	40	30
Math	4	NCF—T, MCF, CA	1500	50	1500	50	430	14	650	22	300	10	1620	54	10	30	40	30
Math	5	NCF—T, MCF, CA	1500	50	1500	50	414	14	693	23	293	10	1600	53	10	30	40	30
Math	6	NCF—T, MCF, CA	1000	50	1000	50	260	13	487	24	187	9	1066	53	10	30	40	30
Math	7	NCF—T, MCF, CA	1000	50	1000	50	275	14	456	23	100	5	1169	58	10	30	40	30
Math	8	NCF—T, MCF, CA	1000	50	1000	50	234	12	446	22	150	8	1170	58	10	30	40	30

Table 3: Demographic characteristics of the normative samples — Reading

| | | | Sex | | | | Race/Ethnicity | | | | | | | | SES | | | |
| | | | Female | | Male | | Black | | Hispanic | | Other | | White | | ELL | Low | Mod | High |
Subject	Grade	Measure	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*%*	*%*	*%*	*%*
Reading	2	ORF	1158	0.49	1187	0.51	355	0.15	614	0.26	246	0.10	1130	0.48	10	32	32	36
Reading	3	ORF	1180	0.51	1125	0.49	266	0.12	583	0.25	225	0.10	1231	0.53	10	32	32	36
Reading	4	ORF	1251	0.50	1233	0.50	307	0.12	591	0.24	250	0.10	1336	0.54	10	32	32	36
Reading	5	ORF	1138	0.52	1066	0.48	337	0.15	518	0.24	251	0.11	1098	0.50	10	32	32	36
Reading	6	ORF	842	0.50	828	0.50	260	0.16	396	0.24	153	0.09	859	0.51	10	32	32	36
Reading	7	ORF	814	0.51	790	0.49	255	0.16	413	0.26	137	0.09	799	0.50	10	32	32	36
Reading	8	ORF	790	0.53	688	0.47	194	0.13	392	0.27	142	0.10	750	0.51	10	32	32	36
Reading	2	RC, VO	1500	0.50	1500	0.50	413	0.14	740	0.25	300	0.10	1547	0.52	10	32	32	36
Reading	3	RC, VO	1500	0.50	1500	0.50	414	0.14	732	0.24	292	0.10	1562	0.52	10	32	32	36
Reading	4	RC, VO, SRF	1500	0.50	1500	0.50	407	0.14	717	0.24	289	0.10	1587	0.53	10	32	32	36
Reading	5	RC, VO, SRF	1500	0.50	1500	0.50	415	0.14	693	0.23	293	0.10	1599	0.53	10	32	32	36
Reading	6	RC, VO, SRF	1000	0.50	1000	0.50	285	0.14	462	0.23	187	0.09	1066	0.53	10	32	32	36
Reading	7	RC, VO, SRF	1000	0.50	1000	0.50	275	0.14	456	0.23	182	0.09	1087	0.54	10	32	32	36
Reading	8	RC, VO, SRF	1000	0.50	1000	0.50	202	0.10	446	0.22	184	0.09	1168	0.58	10	32	32	36

Reliability

Reliability is an estimate of the consistency or stability of test scores. Consistency is affected by random error (which can be caused by many factors, including variations in student motivation and attentiveness), imperfect and incomplete specification of the achievement domain, and guessing. The choice of reliability method depends on how the test is administered and scored, as well as how the results will be used. For untimed tests that assess student achievement at a single point in time, internal consistency reliability is most appropriate. Among the various internal consistency methods, Cronbach's alpha is the most commonly utilized and it is the one reported for all aimswebPlus untimed measures.

Note that for untimed measures, items that were skipped/unanswered were scored as zero. To be included in the analysis, a minimum of five valid item scores were required for any given measure. This number of items was chosen because the administration guidelines for standardization testing indicating that testing should be discontinued if the student failed each of the first five items of a given measure. This occurred, on average, during about 1% of test administrations.

Cronbach's alpha is not appropriate for aimswebPlus timed measures because this type of reliability requires a score on all items in a given measure. The time limits used for aimswebPlus fluency measures are designed to provide strong reliability and growth sensitivity; however, these time limits also have the effect of ensuring that most students will not complete all of the items in a given measure. As such, alternate form reliability is most

appropriate for aimswebPlus timed measures.

Another important reason for using alternate form reliability for these measures is how scores from the timed measures are used. aimswebPlus timed measures are used for benchmark screening and for frequent (e.g., weekly) monitoring of student progress. The timed measures have either 12 or 23 alternate forms for each grade, depending on benchmark seasons administered. Two (Fall/Winter or Winter/Spring) or three (Fall/Winter/Spring) of the forms are used for universal screening, with the remaining 10 or 20 used for progress monitoring. All alternate forms for each measure were constructed from a common test blueprint and are nearly equivalent in difficulty.

Progress monitoring scores are used to estimate rate of growth and to determine whether that rate is sufficient to meet the performance goal set for a student. Therefore, it is important to know how variations in test content and occasion affect score consistency. Alternate form reliability is designed for that purpose.

Reliability results are presented in Table 4, organized by domain: Early Numeracy, Math, Early Literacy, and Reading. Reliability coefficients are provided for each measure, season, and grade within these domains.

Table 4: Reliability of Early Numeracy measures and composites, Kindergarten and Grade 1

Measure	Grade	Season	Cronbach's Alpha		Alternate Form			Stratified Alpha	SEM
			n	Coefficient	n range	Coefficient mean	Range		
NNF	K	F, W, S	—	—	201—207	0.90	0.88—0.90	—	4.63
QTF	K	F, W, S	—	—	93—206	0.80	0.77—0.81	—	2.01
QDF	K	W, S	—	—	201—203	0.74	0.71—0.76	—	2.48
CA	K	F	1378	0.83	—	—	—	—	2.09
CA	K	W	1217	0.83	—	—	—	—	1.93
CA	K	S	1193	0.83	—	—	—	—	1.76
Composite	K	F	—	—	—	—	—	0.88	4.10
Composite	K	W	—	—	—	—	—	0.91	4.41
Composite	K	S	—	—	—	—	—	0.91	4.41
NCF—P	I	F, W, S	—	—	222—239	0.88	0.86—0.89	—	2.47
MFF—ID	I	F, W, S	—	—	217—234	0.86	0.86—0.89	—	2.33
MFF—T	I	F, W, S	—	—	167—175	0.93	0.93	—	1.25
CA	I	F	1459	0.86	—	—	—	—	2.12
CA	I	W	1259	0.87	—	—	—	—	1.88
CA	I	S	1364	0.88	—	—	—	—	1.75
Composite	I	F	—	—	—	—	—	0.96	3.30
Composite	I	W	—	—	—	—	—	0.97	3.29
Composite	I	S	—	—	—	—	—	0.97	3.38

Table 5: Reliability of Math measures and composites, Grades 2 through 8

| Measure | Grade | Season | Cronbach's Alpha | | | Alternate Form | | | Stratified Alpha | | |
			n range	Coefficient mean	Range	*n* Range	Coefficient mean	Range	Coefficient mean	Range	SEM
NCF—T	2	F,W, S	—	—	—	128	0.84	0.82—0.85	—	—	3.16
NCF—T	3	F,W, S	—	—	—	140	0.91	0.91—0.92	—	—	3.10
NCF—T	4	F,W, S	—	—	—	148	0.89	0.88—0.91	—	—	3.53
NCF—T	5	F,W, S	—	—	—	145	0.86	0.85—0.87	—	—	4.05
NCF—T	6	F,W, S	—	—	—	121	0.78	0.78—0.80	—	—	4.22
NCF—T	7	F,W, S	—	—	—	115	0.78	0.76—0.80	—	—	4.05
NCF—T	8	F,W, S	—	—	—	153	0.80	0.80—0.81	—	—	4.21
MCF	2	F,W, S	—	—	—	122	0.85	0.84—0.87	—	—	3.48
MCF	3	F,W, S	—	—	—	140	0.83	0.82—0.84	—	—	3.38
MCF	4	F,W, S	—	—	—	140	0.87	0.87—0.88	—	—	2.69
MCF	5	F,W, S	—	—	—	136	0.85	0.84—0.87	—	—	2.84
MCF	6	F,W, S	—	—	—	120	0.87	0.86—0.89	—	—	3.65
MCF	7	F,W, S	—	—	—	79	0.87	0.86—0.88	—	—	3.60
MCF	8	F,W, S	—	—	—	124	0.91	0.90—0.92	—	—	3.26
NSF	2	F,W, S	—	—	—	113	0.92	0.90—0.93	—	—	4.32
NSF	3	F,W, S	—	—	—	131	0.93	0.92—0.94	—	—	4.45
NSF	4	F,W, S	—	—	—	137	0.93	0.91—0.94	—	—	4.29
NSF	5	F,W, S	—	—	—	132	0.91	0.91—0.92	—	—	5.00
NSF	6	F,W, S	—	—	—	115	0.86	0.83—0.88	—	—	6.60
NSF	7	F,W, S	—	—	—	77	0.88	0.87—0.89	—	—	6.03
NSF	8	F,W, S	—	—	—	123	0.90	0.89—0.91	—	—	6.04
CA	2	F,W, S	1763—1962	0.85	0.85—0.86	—	—	—	—	—	8.87
CA	3	F,W, S	1803—1875	0.85	0.83—0.86	—	—	—	—	—	8.36
CA	4	F,W, S	1726—1925	0.77	0.74—0.82	—	—	—	—	—	8.89
CA	5	F,W, S	1795—1977	0.80	0.80—0.81	—	—	—	—	—	8.52
CA	6	F,W, S	1514—1736	0.81	0.77—0.84	—	—	—	—	—	8.54
CA	7	F,W, S	1144—1409	0.82	0.81—0.83	—	—	—	—	—	8.23
CA	8	F,W, S	1078—1299	0.79	0.77—0.82	—	—	—	—	—	8.55

Measure	Grade	Season	Cronbach's Alpha			Alternate Form			Stratified Alpha		
			n range	Coefficient mean	Range	*n* Range	Coefficient mean	Range	Coefficient mean	Range	SEM
Composite	2	F,W, S	—	—	—	—	—	—	0.92	0.92—0.93	9.93
Composite	3	F,W, S	—	—	—	—	—	—	0.92	0.92—0.93	9.93
Composite	4	F,W, S	—	—	—	—	—	—	0.90	0.89—0.92	9.97
Composite	5	F,W, S	—	—	—	—	—	—	0.91	0.91	9.83
Composite	6	F,W, S	—	—	—	—	—	—	0.90	0.88—0.91	10.90
Composite	7	F,W, S	—	—	—	—	—	—	0.91	0.91—0.92	10.40
Composite	8	F,W, S	—	—	—	—	—	—	0.91	0.91—0.92	10.50

Table 6: Reliability of Early Literacy measures and composites, Kindergarten and Grade 1

Measure	Grade	Season	Cronbach's Alpha		Alternate Form			Stratified Alpha	SEM
			n	Coefficient	n range	Coefficient mean	Range		
LNF	K	F, W, S	—	—	655—672	0.78	0.73—0.82	—	8.74
LWSF	K	PM	—	—	90—217	0.87	0.84—0.90	—	5.39
IS	K	F	1256	0.88	—	—	—	—	1.35
IS	K	W	1221	0.87	—	—	—	—	0.82
PC	K	F	1256	0.63	—	—	—	—	1.06
PS	K	W	1238	0.93	—	—	—	—	3.46
PS	K	S	1221	0.87	—	—	—	—	3.82
AV	K	F	1256	0.82	—	—	—	—	1.65
AV	K	W	1221	0.81	—	—	—	—	1.27
AV	K	S	1238	0.76	—	—	—	—	1.34
Composite	K	W	—	—	—	—	—	0.93	10.09
Composite	K	S	—	—	—	—	—	0.91	10.41
WRF	I	F, W, S	—	—	173—180	0.94	0.93—0.95	—	5.51
ORF	I	F	—	—	1341	0.97	—	—	5.21
ORF	I	W	—	—	1389	0.96	—	—	6.53
ORF	I	S	—	—	1502	0.96	—	—	7.11
PS	I	F	1329	0.83	—	—	—	—	3.84
AV	I	F	1346	0.85	—	—	—	—	0.96
AV	I	W	1390	0.87	—	—	—	—	0.88
AV	I	S	1503	0.87	—	—	—	—	0.6
Composite	I	F	—	—	—	—	—	0.95	8.13

Table 7: Reliability of Reading measures and composites, Grades 2 through 8

Measure	Grade	Season	Cronbach's Alpha			Alternate Form			Stratified Alpha		
			n range	Coefficient mean	Range	*n* Range	Coefficient mean	Range	Coefficient mean	Range	SEM
ORF	2	F,W, S	—	—	—	1719—1900	0.96	0.95—0.97	—	—	7.78
ORF	3	F,W, S	—	—	—	1580—1902	0.96	0.95—0.96	—	—	7.46
ORF	4	F,W, S	—	—	—	1633—2014	0.95	0.95—0.96	—	—	8.40
ORF	5	F,W, S	—	—	—	1649—2009	0.95	0.95	—	—	9.33
ORF	6	F,W, S	—	—	—	1271—1449	0.95	0.94—0.96	—	—	8.54
ORF	7	F,W, S	—	—	—	959—1097	0.94	0.94—0.95	—	—	9.58
ORF	8	F,W, S	—	—	—	850—1051	0.95	0.94—0.96	—	—	8.13
SRF	4	F,W, S	—	—	—	1857—2022	0.87	—	—	—	17.09
SRF	5	F,W, S	—	—	—	1926—2212	0.87	—	—	—	14.90
SRF	6	F,W, S	—	—	—	1322—1632	0.86	—	—	—	18.31
SRF	7	F,W, S	—	—	—	985—1238	0.87	—	—	—	17.77
SRF	8	F,W, S	—	—	—	939—1207	0.86	—	—	—	17.59
VO	2	F,W, S	1842—2084	0.67	0.63—0.71	—	—	—	—	—	13.88
VO	3	F,W, S	1839—1955	0.73	0.72—0.74	—	—	—	—	—	11.43
VO	4	F,W, S	1874—2083	0.74	0.73—0.74	—	—	—	—	—	10.91
VO	5	F,W, S	1910—2291	0.73	0.70—0.75	—	—	—	—	—	11.03
VO	6	F,W, S	1358—1664	0.73	0.72—0.76	—	—	—	—	—	10.93
VO	7	F,W, S	1003—1293	0.75	0.73—0.77	—	—	—	—	—	10.68
VO	8	F,W, S	950—1241	0.82	0.80—0.83	—	—	—	—	—	10.40
RC	2	F,W, S	1870—2053	0.86	0.85—0.88	—	—	—	—	—	10.10
RC	3	F,W, S	1868—1937	0.87	0.86—0.89	—	—	—	—	—	9.82
RC	4	F,W, S	1853—2002	0.84	0.82—0.86	—	—	—	—	—	9.52
RC	5	F,W, S	1903—2117	0.85	0.84—0.87	—	—	—	—	—	9.55
RC	6	F,W, S	1292—1535	0.84	0.84—0.85	—	—	—	—	—	10.07
RC	7	F,W, S	978—1191	0.85	0.83—0.86	—	—	—	—	—	9.36
RC	8	F,W, S	907—1143	0.84	0.83—0.85	—	—	—	—	—	9.61

Table 7: Reliability of Reading measures and composites, Grades 2 through 8

Measure	Grade	Season	Cronbach's Alpha			Alternate Form			Stratified Alpha		
			n range	Coefficient mean	Range	*n* Range	Coefficient mean	Range	Coefficient mean	Range	SEM
Composite	2	F, W, S	—	—	—	—	—	—	0.91	0.91—0.92	17.98
Composite	3	F, W, S	—	—	—	—	—	—	0.92	0.92—0.93	16.40
Composite	4	F, W, S	—	—	—	—	—	—	0.88	0.87—0.89	18.98
Composite	5	F, W, S	—	—	—	—	—	—	0.88	0.87—0.89	18.76
Composite	6	F, W, S	—	—	—	—	—	—	0.87	0.86—0.88	20.33
Composite	7	F, W, S	—	—	—	—	—	—	0.88	0.88	19.54
Composite	8	F, W, S	—	—	—	—	—	—	0.89	0.89—0.90	19.72

In summary, reliability estimates typically met common benchmarks for adequate consistency for measures used to make decisions about individual students. There were only two cases in which a reliability estimate fell below 0.70 — the internal consistency of Print Concepts scores for students in grades K-1 was 0.63, and average internal consistency of Vocabulary scores for students in grades 2 was 0.67. In particular:

— Internal consistency of untimed early numeracy measures for students in grades K-1 ranged from 0.83 to 0.88. Average alternate forms reliability for timed measures ranged from 0.74 to 0.93. Stratified alpha for composite scores ranged from 0.88 to 0.97.
— Average internal consistency of untimed Math measures for students in grades 2-8 ranged from 0.77 to 0.85. Average alternate forms reliability for timed measures ranged from .78 to .93. Average stratified alpha for composite scores ranged from 0.90 to 0.92.
— Internal consistency of untimed early literacy measures for students in grades K-1 ranged from 0.63 to 0.93. Average alternate forms reliability for timed measures ranged from .78 to .97. Stratified alpha for composite scores ranged from 0.91 to 0.95.
— Average internal consistency of untimed reading measures for students in grades 2-8 ranged from 0.67 to 0.87. Average alternate forms reliability for timed measures ranged from .86 to .96. Average stratified alpha for composite scores ranged from 0.87 to 0.92.

Validity

Validity is the degree to which evidence supports interpretations of test scores for a given purpose. There are several different types of validity evidence that can be provided, depending on the proposed use of the test. Because aimswebPlus is used to identify students at risk of academic failure and track progress toward academic goals in reading and math, one particularly relevant form of validity evidence is the extent to which performance on the tests correlates with performance on other measures, which are called criterion measures. Correlations with criterion measures administered at the same time are called concurrent validity coefficients, and correlations with criterion measures administered at a later time are called predictive validity coefficients. These coefficients can range from -1.0 to 1.0, with positive values closer to 1.0 indicating a stronger positive relationship. To the extent that the coefficients are high, this suggests that the tests are doing a good job measuring targeted reading and math skills and predicting future performance on end-of-year achievement tests.

During the 2013–2014 standardization study, Pearson obtained achievement scores for participating students from other reading and math tests used by each school. As a condition of participation, schools provided spring test scores from interim assessments, state NCLB tests or other formative assessments. A secure file transfer protocol was used to share data, with test scores being provided to Pearson without individually identifiable information. A unique, randomly derived student ID assigned by Pearson was used to match each participant's scores to standardization data.

This section presents the concurrent and predictive validity coefficients obtained from these data from criterion measures and aimswebPlus. Concurrent validity represents the correlation of aimswebPlus composite scores and criterion measure scores, both from the Spring testing season. Predictive validity represents the correlation of Fall aimswebPlus composite scores and Spring scores from the criterion measures.

Predicting student achievement in the Spring from Fall benchmark scores is the basis for determining a student's risk status. The National Center on Intensive Intervention (NCII) requires predictive validity coefficients of 0.70 or higher to obtain the maximum rating (i.e., providing convincing evidence) for screeners. However, there is not a single universally accepted standard for defining success, and many different tests are used across U.S. schools; thus, it is important to evaluate predictive validity with several criterion measures.

When a test shows strong prediction with several different criterion measures, there is greater confidence that results can be generalized to other standardized and validated measures of student achievement. In the sections that follow, concurrent and predictive validity coefficients for aimswebPlus Early Numeracy, Math, Early Literacy, and Reading benchmark composites are provided.

Early Numeracy

Table 8 shows the predictive validity coefficients of the aimswebPlus Early Numeracy composite scores with the Tennessee Comprehensive Achievement Program (TCAP) math scores. TCAP assesses math skills aligned to Tennessee's state learning standards. The characteristics of the sample upon which the coefficient was obtained are also provided.

Table 9 shows the concurrent validity coefficients for the aimswebPlus Early Numeracy composites with TCAP math scores. The aimswebPlus Early Numeracy scores were collected in May 2014, while TCAP scores were obtained in late April 2014.

As can be seen, all coefficients, adjusted for range restriction were at least 0.70. In particular, predictive validity ranges from 0.70 to 0.87 and concurrent validity ranges from 0.73 to 0.79.

Table 8: Predictive validity coefficients of the aimswebPlus Early Numeracy composite scores

| Criterion | Grade | n | Predictive | | Sex | | Race/Ethnicity | | | |
			Unadjusted	Adjusted	% Female	% Male	% Black	% Hispanic	% Other	% White
TCAP	K (Fall)	68	0.62	0.70	41	59	9	12	0	79
TCAP	K (Winter)	68	0.70	0.76	41	59	9	12	0	79
TCAP	I (Fall)	55	0.79	0.86	53	47	2	25	0	73
TCAP	I (Winter)	55	0.80	0.87	53	47	2	25	0	73

Table 9: Concurrent validity coefficients for the aimswebPlus Early Numeracy composites with TCAP math scores

| Criterion | Grade | n | Predictive | | Sex | | Race/Ethnicity | | | |
			Unadjusted	Adjusted	% Female	% Male	% Black	% Hispanic	% Other	% White
TCAP	K (Spring)	68	0.66	0.73	41	59	9	12	0	79
TCAP	I (Spring)	55	0.68	0.79	53	47	2	25	0	73

Math

Five criterion measures were used to calculate criterion validity for aimswebPlus Math:

— Iowa Tests of Basic Skills®–Total Math (ITBS®)
— Illinois Standards Achievement Test (ISAT)
— New Mexico Standards Based Assessment (NMSBA)
— Northwest Evaluation Association Measures of Academic Progress® (NWEA–MAP®)
— State of Texas Academic Assessment of Readiness (STAAR)

The ITBS is a comprehensive, group-administered, paper-based assessment of reading and math achievement. ITBS's Total Math score reflects performance on standards-based math concepts, problem solving, and computation. The ISAT is the end-of-year achievement test assessing Illinois learning standards covering five math strands: Number Sense, Measurement, Algebra, Geometry, and Data Analysis and Probability. The NMSBA is used to measure student proficiency on New Mexico's reading and math learning standards. NWEA–MAP is a computer-adaptive test that assesses achievement in reading and mathematics. Results are reported on an RIT scale, which is then linked to each state's performance standards. Finally, the STAAR assesses student performance on Texas's mathematics and reading learning standards.

Table 10 shows the predictive validity coefficients of the aimswebPlus Math composite with each criterion measure. Weighted mean validity coefficients, by grade, are also shown, which provides an estimate of the overall predictive validity. The characteristics of the sample upon which the coefficient was obtained are also provided.

Table 11 shows the concurrent validity coefficients for the aimswebPlus Math composite with each criterion measure, as well as the mean adjusted coefficients by grade. aimswebPlus Math scores were collected in May 2014, while the criterion measures scores were obtained in March through May 2014.

As can be seen, all average coefficients but one, adjusted for range restriction were at least 0.70. In particular, average predictive validity coefficients range from 0.69 to 0.85, and average concurrent validity coefficients range from 0.77 to 0.85.

Table 10: Predictive validity coefficients of the aimswebPlus Math composite

Criterion	Grade	n	Correlation			Sex		Race/Ethnicity			
			Unadjusted	Adjusted	Mean	% Female	% Male	% Black	% Hispanic	% Other	% White
ITBS	2	179	0.79	0.81	0.69	60	40	19	42	21	17
NWEA—MAP	2	218	0.62	0.56		48	52	5	31	12	53
ISAT	3	69	0.85	0.81		49	51	1	25	13	61
NWEA—MAP	3	101	0.83	0.79	0.79	46	54	1	40	14	44
STAAR	3	146	0.74	0.77		55	45	10	39	37	14
ISAT	4	175	0.80	0.79		51	49	4	28	9	58
NWEA—MAP	4	95	0.76	0.75	0.76	59	41	5	35	10	49
STAAR	4	207	0.75	0.73		51	49	8	46	32	14
ISAT	5	189	0.86	0.84		53	47	2	21	9	68
NWEA—MAP	5	81	0.89	0.86	0.83	47	53	3	43	11	43
STAAR	5	91	0.70	0.79		49	51	2	52	41	6
ISAT	6	273	0.84	0.89		59	41	22	6	8	64
NMSBA	6	210	0.75	0.80	0.85	52	48	3	64	1	32
NWEA—MAP	6	86	0.79	0.83		55	45	22	9	10	59
STAAR	6	61	0.63	0.75		55	45	5	44	48	3
ISAT	7	130	0.84	0.90		45	55	13	2	3	82
NMSBA	7	220	0.78	0.78	0.85	47	53	2	62	0	36
STAAR	7	61	0.80	0.90		40	60	5	43	49	4
ISAT	8	122	0.62	0.74		37	63	5	1	3	91
NMSBA	8	223	0.84	0.87	0.83	44	56	6	67	1	26
STAAR	8	75	0.61	0.79		61	39	15	53	32	0

Table 11: Concurrent validity coefficients for the aimswebPlus Math composite

Criterion	Grade	n	Correlation			Sex		Race/Ethnicity			
			Unadjusted	Adjusted	Mean	% Female	% Male	% Black	% Hispanic	% Other	% White
ITBS	2	218	0.82	0.81		60	40	19	42	21	17
NWEA—MAP	2	179	0.73	0.71	0.77	48	52	5	31	12	53
ISAT	3	46	0.84	0.82		49	51	1	25	13	61
NWEA—MAP	3	101	0.87	0.85	0.83	46	54	1	40	14	44
STAAR	3	211	0.76	0.82		55	45	10	39	37	14
ISAT	4	126	0.85	0.83		51	49	4	28	9	58
NWEA—MAP	4	95	0.82	0.80	0.79	59	41	5	35	10	49
STAAR	4	277	0.77	0.76		51	49	8	46	32	14
ISAT	5	154	0.85	0.84		53	47	2	21	9	68
NWEA—MAP	5	81	0.84	0.84	0.82	47	53	3	43	11	43
STAAR	5	157	0.72	0.80		49	51	2	52	41	6
ISAT	6	231	0.85	0.88		59	41	22	6	8	64
NMSBA	6	210	0.77	0.85	0.85	52	48	3	64	1	32
NWEA—MAP	6	86	0.74	0.76		55	45	22	9	10	59
STAAR	6	61	0.68	0.79		55	45	5	44	48	3
ISAT	7	130	0.78	0.83		45	55	13	2	3	82
NMSBA	7	220	0.76	0.85	0.84	47	53	2	62	0	36
STAAR	7	61	0.74	0.84		40	60	5	43	49	4
ISAT	8	122	0.68	0.73		37	63	5	1	3	91
NMSBA	8	223	0.80	0.87	0.82	44	56	6	67	1	26
STAAR	8	75	0.56	0.77		61	39	15	53	32	0

Early Literacy

An important outcome of Kindergarten early literacy instruction is to move students from elementary phonological awareness, such as letter identification and letter sounds, to word reading and eventually to reading connected text in the form of sentences and short stories. Thus, the aimswebPlus measure Word Reading Fluency is used as the predictive criterion measure of Fall and Winter Kindergarten scores. Word Reading Fluency assesses a student's automaticity with reading high frequency and highly decodable words. Students are given 1 minute to read as many words as possible.

In the Fall testing season of Kindergarten, aimswebPlus requires only Letter Naming Fluency for assessing risk status. This measure was selected because research shows it to be a strong predictor of end-of-year oral reading fluency ability (Clemens et al., 2015), and because it is a very appropriate measure of foundational reading skills in beginning Kindergarten. By midyear, Kindergarten students typically have had formal instruction on letter identification, letters sounds, and parsing simple words into phonemes. As such, the aimwebPlus Early Literacy Winter composite for Kindergarten also includes Letter Word Sounds Fluency and Phoneme Segmentation. The composite of these three measures is used to identify risk and predict end-of-grade performance on Word Reading Fluency.

In Grade 1, early literacy instruction continues with a greater emphasis on word reading, as well as reading and comprehending connected text. For Grade 1 students, Oral Reading Fluency has been shown to provide strong prediction of end-of-grade performance on broad measures of reading. The Iowa Test of Basic Skills Level 6 measures vocabulary, word reading, and reading comprehension at the end of Grade 1, making it an appropriate criterion measure for ORF.

Table 12 shows the unadjusted and adjusted predictive validity coefficients of aimswebPlus LNF (Kindergarten, Fall), the composite comprised of LNF, LWSF, and PSF (Kindergarten, Winter), and ORF (Grade 1, Fall). The characteristics of the sample upon which the coefficient was obtained are also provided. Because WRF was administered to all Kindergarten students in the Spring testing season, data from this measure was used to obtain the validity coefficient.

Table 13 shows the concurrent validity coefficients for the composite comprised of LNF, LWSF, and PSF (Kindergarten, Spring) and ORF (Grade 1, Spring). ITBS scores were obtained in April 2014.

As can be seen, coefficients varied by criterion measure, with higher coefficients seen for ITBS scores. In particular, when coefficients were adjusted for range restriction, predictive validity ranges from 0.58 (for Word Reading Fluency administered in the fall) to 0.72 (for ITBS administered in the fall) and concurrent validity was 0.57 (for Word Reading Fluency) and 0.74 (for ITBS).

Table 12: Unadjusted and adjusted predictive validity coefficients of aimswebPlus

			Predictive		Sex		Race/Ethnicity			
Criterion	Grade	*n*	Unadjusted	Adjusted	% Female	% Male	% Black	% Hispanic	% Other	% White
WRF	K (Fall)	1075	0.58	0.58	50	50	14	25	10	51
WRF	K (Winter)	1075	0.63	0.63	50	50	14	25	10	51
ITBS	I (Fall)	61	0.57	0.72	41	59	25	25	17	33

Table 13: Concurrent validity coefficients

Criterion	Grade	n	Predictive		Sex		Race/Ethnicity			
			Unadjusted	Adjusted	% Female	% Male	% Black	% Hispanic	% Other	% White
WRF	K (Spring)	1075	0.57	0.57	50	50	14	25	10	51
ITBS	I (Spring)	61	0.67	0.74	41	59	25	25	17	33

Reading

Four criterion measures were used to calculate criterion validity for aimswebPlus Reading:

— Illinois Standards Achievement Test (ISAT)
— Missouri Assessment Program Grade Level Assessment (MAP–GLA)
— Northwest Evaluation Association Measures of Academic Progress (NWEA–MAP)
— State of Texas Academic Assessment of Readiness (STAAR)

The ISAT is the end-of-year achievement test assessing Illinois learning standards, including reading comprehension. The MAP–GLA is the end-of-year achievement test that assesses Missouri reading and math standards, including reading comprehension. NWEA–MAP is a computer-adaptive test that assesses achievement in reading and mathematics. Results are reported on an RIT scale, which is then linked to each state's performance standards. Finally, the STAAR assesses student performance on Texas's mathematics and reading learning standards.

Table 14 shows the predictive validity coefficients of the aimswebPlus Reading composite with each criterion measure. Weighted mean validity coefficients, by grade, are also shown, which provides an estimate of the overall predictive validity. The characteristics of the sample, upon which the coefficient was obtained, are also provided.

Table 15 shows the concurrent validity coefficients for the aimswebPlus Reading composite with each criterion measure, as well as the mean adjusted coefficients by grade. aimswebPlus Math scores were collected in May 2014, while the criterion measures scores were obtained in March through May 2014.

As can be seen, all average coefficients but two, adjusted for range restriction were at least 0.70. In particular, mean predictive validity coefficients range from 0.69 to 0.83, and mean concurrent validity coefficients range from 0.68 to 0.80.

Table 14: Predictive validity coefficients of the aimswebPlus Reading composite

| Criterion | Grade | n | Correlation | | | Sex | | 6Race/Ethnicity1 | | | |
			Unadjusted	Adjusted	Mean	% Female	% Male	% Black	% Hispanic	% Other	% White
NWEA—MAP	2	128	0.83	0.83	0.83	52	48	2	23	21	53
ISAT	3	113	0.80	0.84		47	53	2	28	20	49
MAP-GLA	3	317	0.71	0.69	0.77	55	45	24	2	2	72
NWEA—MAP	3	150	0.78	0.79		45	55	2	25	20	52
STAAR	3	208	0.70	0.74		56	44	10	49	14	27
ISAT	4	230	0.77	0.79		56	44	4	39	10	47
MAP-GLA	4	292	0.62	0.58	0.69	49	51	32	1	5	62
NWEA—MAP	4	125	0.76	0.77		53	47	4	28	16	52
STAAR	4	277	0.60	0.61		44	56	8	52	10	29
ISAT	5	250	0.73	0.75		48	52	4	22	13	61
MAP-GLA	5	222	0.65	0.65	0.73	50	50	42	0	7	50
NWEA—MAP	5	141	0.81	0.79		48	52	3	30	18	48
STAAR	5	157	0.66	0.71		53	47	9	57	3	31
ISAT	6	332	0.74	0.77	0.75	58	42	9	14	12	65
NWEA—MAP	6	124	0.67	0.73		52	48	4	21	12	63
ISAT	7	179	0.78	0.81		44	56	12	12	7	68
MAP-GLA	7	101	0.71	0.78	0.73	46	54	41	4	0	55
NWEA—MAP	7	207	0.51	0.61		51	49	9	24	12	55
ISAT	8	202	0.72	0.80	0.78	46	54	10	11	6	74
MAP-GLA	8	218	0.69	0.76		57	43	28	3	1	68

Table 15: Concurrent validity coefficients for the aimswebPlus Reading composite

			Correlation			Sex		6Race/Ethnicity1			
Criterion	Grade	n	Unadjusted	Adjusted	Mean	% Female	% Male	% Black	% Hispanic	% Other	% White
NWEA—MAP	2	128	0.80	0.80	0.80	52	48	2	23	21	53
ISAT	3	113	0.85	0.88		47	53	2	28	20	49
MAP-GLA	3	317	0.69	0.69	0.77	55	45	24	2	2	72
NWEA—MAP	3	150	0.80	0.80		45	55	2	25	20	52
STAAR	3	208	0.70	0.72		56	44	10	49	14	27
ISAT	4	230	0.73	0.76		56	44	4	39	10	47
MAP-GLA	4	292	0.70	0.68	0.70	49	51	32	1	5	62
NWEA—MAP	4	125	0.67	0.71		53	47	4	28	16	52
STAAR	4	277	0.67	0.66		44	56	8	52	10	29
ISAT	5	250	0.79	0.80		48	52	4	22	13	61
MAP-GLA	5	222	0.64	0.67	0.73	50	50	42	0	7	50
NWEA—MAP	5	141	0.77	0.76		48	52	3	30	18	48
STAAR	5	157	0.65	0.69		53	47	9	57	3	31
ISAT	6	332	0.79	0.81	0.78	58	42	9	14	12	65
NWEA—MAP	6	124	0.72	0.74		52	48	4	21	12	63
ISAT	7	179	0.78	0.80		44	56	12	12	7	68
MAP-GLA	7	101	0.64	0.67	0.68	46	54	41	4	0	55
NWEA—MAP	7	207	0.50	0.57		51	49	9	24	12	55
ISAT	8	202	0.72	0.79	0.76	46	54	10	11	6	74
MAP-GLA	8	218	0.69	0.72		57	43	28	3	1	68

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, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Baglici, S. P., Codding, R., & Tryon, G. (2010). Extending the research on the tests of early numeracy: Longitudinal analyses over two school years. *Assessment for Effective Intervention, 35*(2), 89102.

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

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.K., Smolkowski, K., and Chard, D. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education*, 29, 46–57.

Clemens, N.H., Hagan-Burke, S., Luo, W., Cerda, C., Blakely, A., Frosch, J., … Jones, M. (2015). The predictive validity of kindergarten and first-grade reading skills. *School Psychology Review, 44*(1), 76–97.

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.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49(1), 3645.

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.

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.

Fuchs, L. S., Fuchs, D., & Deno, S. L. (1982). Reliability and validity of curriculum-based informal reading inventories. *Reading Research Quarterly, 18*(1), 626.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, 9(2), 2028.

Fuchs, L.S., & Vaughn, S.R. (2005). Response-to-intervention as a framework for the identification of learning disabilities. *Trainer's Forum: Periodical of the Trainers of School Psychologists, 25*(1), 12–19.

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.

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

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.

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.

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

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).

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), 116125.

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., 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

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 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 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.

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

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.

Partnership for Assessment of Readiness for College and Careers (PARCC). (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. (2017). 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

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

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. (1992). Curriculum-Based reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*(3), 458–478.

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

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.

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.