Efficacy Report

# aimswebPlus

March 23, 2016

## Product Summary

aimsweb is a curriculum-based measurement system used for universal screening, progress monitoring and data management. This brief form of assessment measures overall performance of key foundational skills at each grade level and draws upon over 30 years of research that demonstrates both its accurate prediction of reading and math achievement, as well as its sensitivity to growth. Measures include literacy, reading, numeracy and spelling. Based on direct, frequent and continuous student assessment, results are used to identify students in need of intensive intervention and to track their progress in response to intervention.

aimsweb currently enjoys approximately 640,000 users, reaching 3.8 million learners, resulting in over 50 million screening and progress monitoring assessments annually. Its key markets are the US and Canada. Intended for students in grades K-8, with normative information based on 8th grade content for grades 9-12, aimsweb provides universal screening for all students, as well as intensive progress monitoring for at-risk students or those with severe needs.

The original aimsweb screening and monitoring assessments were launched in 2000. A new and updated assessment suite, called aimswebPlus, will be launched in fall 2016. Key features of aimswebPlus include new standards-based content, new reports, digital and online administration of the assessments, and a new platform that optimises usability, simplicity, scalability, and performance. However, because the new assessment suite is not yet available, there is still limited research evidence. Thus, this report summarizes evidence on the previous version of the product--aimsweb.

# Intended Outcomes

## Overview of Intended Outcomes

The efficacy of aimswebPlus can be conceptualized as its quality is a signal to educators of students' need for intensive intervention in reading or mathematics and as an indication of their progress over time toward specific academic goals. Signal quality, in turn, can be characterized as a function of the consistency and accuracy of scores (*reliability*) and the extent to which the assessments allow educators to make sound interpretations of student skills (*validity*) (AERA, APA, & NCME, 2014).

**Intended Outcome 1: Test scores can be interpreted as measures of student achievement in key foundational skills in reading and math, and can be used for universal screening and progress monitoring (Validity).**

One of the key outcomes for aimswebPlus is to enable educators to make sound interpretations about student skills and to support their decisions about students' need for intensive intervention in reading and mathematics by providing measures that accurately predict student performance on end-of-year reading and mathematics achievement tests, identifying students unlikely to pass.

**Intended Outcome 2: Test scores are consistent over passages/probes and different scorers, and individual student growth estimates are consistent over time (Reliability).**

Another important goal of aimswebPlus is to minimize errors in judgment and decision-making by providing scores that are consistent over different forms and raters, and individual student growth estimates that are consistent over time.

# Foundational Research

## Overview of Foundational Research on aimsweb

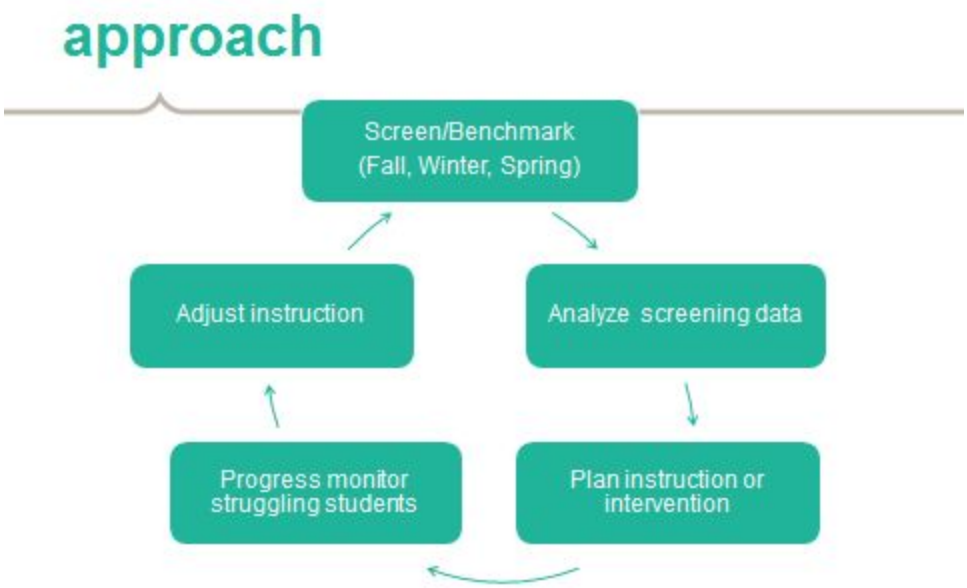
aimsweb assessments consist of brief, standardized measures of skills in early literacy, reading, early numeracy, mathematics, spelling and writing that are consistent with research on 'Curriculum-Based Measurement' (CBM). For example, aimsweb's reading measures include oral reading fluency probes for calculating the number of words correctly read aloud in one minute. Multiple studies have demonstrated that these types of 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). Several studies have found that three minute writing probes, similar to aimsweb's 'Written Expression' measure, are moderately consistent over alternate probes, and moderately related to both a district writing test and teacher ratings of writing ability (Espin et al., 2000; Gansle et al., 2002; Jewell & Malecki, 2005; McMaster & Campbell, 2007). Early numeracy measures, similar to aimsweb's oral counting, number identification, quantity discrimination, and missing number measures, were found to be moderately related to performance on a standardized mathematics achievement test (Clarke et al., 2008). Moreover, early literacy measures similar to aimsweb's letter naming fluency, letter sound fluency, and phonemic segmentation fluency, demonstrate high levels of consistency over forms, over time, and over raters (Elliott et al., 2001).

# Intended Product Implementation

## Overview of Intended Product Implementation

aimswebPlus is intended to be implemented across all students in grades K-8 for the purposes of identifying and monitoring students' foundational skills. aimswebPlus is also intended to provide information for educators to evaluate the performance of groups of students and the impact of instructional programs on the development of these foundational skills.

As shown in the graphic 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.




A consistent approach to implementation support has been taken for both aimsweb and aimswebPlus. Customers of aimswebPlus will 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

## Overview of Product Research

aimswebPlus research and development was designed intentionally to obtain evidence on item quality, test score reliability, validity and sensitivity to academic growth, and to ensure the assessments conform to professional standards and meet customer expectations. In addition, there is a wealth of evidence on previous versions of the aimsweb system supporting the consistency and accuracy of scores and the ability to support sound interpretations and decision-making.

Future research will seek to replicate these results for the new version. Immediate plans for 2016 studies include inter-rater reliability studies to show evidence of scoring objectivity. The team is currently prioritizing other 2016 and longer-term research efforts, which will likely include evidence that scores from frequent monitoring of students are reliable and that decision rules built into the progress monitoring system are accurate and valid.

## Research Studies

| aimsweb R-CBM Passage Field Test | |
|---|---|
| Study Citation | NCS Pearson, Inc. (2012). *aimsweb Technical Manual*. Bloomington, MN: NCS Pearson, Inc. |
| Research Study Contributors | NA |
| Type of Study | Field test of aimsweb passages |
| Sample Size | 24 students per grade level from grades 1-7 in 4 elementary and 3 middle schools in a suburban/rural midwestern school district, plus 183 grade 8 students in 14 different schools and 5 states. |
| Description of Sample | The students represented a range of ability levels, according to their performance on standardized reading tests (⅓ each at the 75th, 50th, and 25th percentiles).<br><br>Other demographic information on the field test sample was not reported, so it is unknown how representative the sample was of the target population in terms of gender, racial/ethnic background, and other characteristics. |
| Outcomes Measured | Average passage difficulty, alternate-forms reliability, and passage readability. |

The technical manual provides a description of the process used to write and field-test the R-CBM passage sets across grades 1-8, as well as analysis and results of the passage field test.

In the study, 24 students per grade from grades one to-seven from four elementary and three middle schools in a suburban/rural midwestern school district read each passage in a set. In addition, passages were field-tested with 183 grade eight students in 14 different schools and five states, with each student reading each passage in a set. There were 33 passages for grade one, and 50 passages per grade level for grades two to eight. Each student read all of the passages at his or her grade level, spread across five sessions of 10 passages each (six to seven passages per session for grade one). The student read each passage aloud for one minute, and the examiner recorded the number of words read correctly (WRC) and the number of errors. Passages were evaluated according to difficulty (average WRC) and alternate-form reliability (average correlation with the other passages at that grade level). In addition, the readability of each passage was assessed by calculating its Lexile measure (Stenner, 1997).

Generally, passages whose average WRC differed from the overall average WRC by more than the standard error of the mean (SEM), whose Lexile measure was outside the grade range, or whose correlation with other passages was below .70 were rejected. The pool of retained passages included 23 at grade one and 33 at each of grades two through seven.

The grade eight passages were not retained in their initial form because the means and standard deviations (SD) of their WRCs were lower than those for the grade seven passages and because most of them had average alternate-form correlations below .70. These passages were revised and subjected to a second field test involving 83 students randomly drawn from four schools in five states. Each participating student read 10 randomly-selected passages; each passage was administered to an average of 36 students (range = nine to 55 students). The data from this second field test indicated that even though average WRCs were still lower than at grade seven, the revised passages had good alternate-form reliability; therefore, 33 passages were retained.

For the final passage sets, the average WRC shows an increasing trend across grades one to seven, before dropping at grade eight. Benchmark passage levels were aligned to common readability indices (Lexile, Fry, Flesch, Powers, Spache and the SMOG), with correlations ranging from 0.83 to 0.97. Similarly, progress monitoring passage levels were aligned to common readability indices (e.g., Lexile, Flesch), with correlations ranging from 0.81 to 0.91. Similar information for aimswebPlus oral reading passages will be available in the aimswebPlus technical manual, scheduled to be available 8/1/2016.

It should be noted that demographic characteristics for the passage field test sample were not reported, so it is unknown how representative the sample was of the target population in terms of gender, racial/ethnic background, and other characteristics. In addition, the sample did not include urban school representation.

| aimsweb Criterion-Related Validity Study | |
|---|---|
| Study Citation | NCS Pearson, Inc. (2012). *aimsweb Technical Manual*. Bloomington, MN: NCS Pearson, Inc. |
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | R-CBM: Participants included roughly 1000 students in each of grades 3-8.<br><br>Reading Maze: Participants included between roughly 2,000 to 10,000 students in each of grades 3-8.<br><br>M-CAP: Participants included roughly between 550-960 students in each of grades 3-8.<br><br>M-COMP: Participants included between 54-98 students in each of grades 1, 3, and 8.<br><br>TEL: Participants included between 28-438 students in each of grades K-1 who were in the aimsweb database. |
| Description of Samples | Demographic information on the samples were not reported, so it is unknown how representative the samples were of the target population in terms of gender, racial/ethnic background, and other characteristics.<br><br>R-CBM: Participants were drawn from the aimsweb user database from the 2009-2010 school year.<br><br>Reading Maze: Participants were drawn from the aimsweb user database from the 2009-2010 school year.<br><br>M-CAP: Participants were drawn from the aimsweb user database from the 2009-20010 school year.<br><br>M-COMP: Participants were drawn from the field test sample.<br><br>TEL: Participants were drawn from the aimsweb user database from the 2007–2008 and 2009–2010 school years. |
| Outcomes Measured | Student performance on the following aimsweb subtests: R-CBM, Reading Maze, M-CAP, M-COMP, and TEL, as well as |

| | |
|---|---|
| | scores on the following external measures of reading and mathematics achievement: North Carolina End of Grade Test, Illinois Standards Achievement Test, G-MADE, Pennsylvania System of School Assessment, and the Minnesota Comprehensive Assessment. |

In each study, students were administered one or more aimsweb subtests during either the fall, winter, or spring administration period, as well as a relevant external achievement measure in reading and/or mathematics. In all but one study, the external criterion measures were administered at the end of the same school year. For the TEL study, the interval between administration of the TEL and the criterion measure was often as long as two to three years.

Scores on the aimsweb and external criterion measures were correlated. Correlations between the R-CBM and R-COMP measures and external achievement were corrected for range restriction, using the variability of the norming sample as the referent.

Estimated correlations for each subtest were as follows:

- R–CBM scores correlated approximately .70 with Illinois and North Carolina state reading test scores in grades three through five and in the mid-to-low .60s in grades six through eight.
- Median correlations between Reading Maze and the various achievement measures ranged from 0.51 to 0.59.
- Correlations between M-CAP and Illinois and North Carolina state mathematics test scores ranged from 0.56-0.65 in grades three through five, and from 0.73-0.80 in grades six through eight.
- Correlations between the M-COMP and the G-MADE ranged from 0.73 to 0.84.
- Correlations between the TEL and the Pennsylvania System of School Assessment, Illinois Standards Achievement Test, and the Minnesota Comprehensive Assessment ranged from 0.33-0.76, with most correlations in the moderate range.

In addition, cut scores on the R-CBM and M-CAP were used to predict whether students would pass the North Carolina End of Grade Test and the Illinois Standards Achievement Test for a subset of the overall samples. Sample sizes for R-CBM ranged from 200 students at grade three to 1,656 students at grade seven. Across the three benchmark periods, between 72-80% of students who ultimately failed the end-of-year tests were correctly flagged by the aimsweb R-CBM measure. Corresponding percentages for the aimsweb M-CAP measure ranged from 75-85%. Similarly, across the three benchmark periods, between 73-91% of students who ultimately passed the end-of-year tests were correctly identified as such by the aimsweb R-CBM measure, with corresponding percentages for the M-CAP measure ranging from 68-82%.

| Curriculum-Based Measures and Performance on State Assessment and Standardized Tests | |
|---|---|
| Study Citation | Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *24*(1), 19-35. |
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | 617 students in grades 3-5 |
| Description of Samples | Students came from one moderate-sized urban/suburban district in Pennsylvania, where approximately 33% of the student population were identified as low-income, 8% were identified as having limited English proficiency, and 11% of students as having Individualized Education Programs (IEPs). |
| Outcomes Measured | aimsweb R-CBM measures, Pennsylvania System of School Assessment, Metropolitan Achievement Test 8. |

Shapiro, Keller, Lutz, Santoro, and Hintze (2006) collected data at grades three through five from an urban/suburban school district in eastern Pennsylvania, reporting correlations of R–CBM scores taken from the fall, winter, and spring administrations with scores on group reading tests administered in the spring of the same grade. Estimated correlations (all significant at the $p<.001$ level) for each external criterion measure were as follows:

- For the PSSA, correlations ranged from 0.669 to 0.693 at grade five and from 0.647 to 0.671 at grade three.
- Correlations with the Total Reading Score on the MAT 8 ranged from 0.701 to 0.724 at grade four and from 0.711 to 0.740 at grade two.

| Using Reading Curriculum-based Measurements as Predictors for the Measures of Academic Progress (MAP) Standardized Test in Nebraska. | |
|---|---|
| Study Citation | Merino, K., & Beckman, T. O. (2010). Using reading curriculum-based measurements as predictors for the measures of academic progress map standardized test in Nebraska. *International Journal of Psychology: A Biopsychosocial Approach*, *6*, 85-98. |
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | Participants included 233 students in grades 2-4. |
| Description of Samples | Participants were drawn from a Nebraska public school, characterized as enrolling “diverse students from all backgrounds,” with 39% of the student population English language learners, 15% students with disabilities, and 80% non-white students. |
| Outcomes Measured | aimsweb R-CBM measures, as well as the Measures of Academic Progress reading composite scores. |

Merino and Beckman (2010), using a sample of 233 students in Grades two through four at a Nebraska elementary school, evaluated the correlation of R–CBM screening scores in the spring with scores on the Measures of Academic Progress (MAP) collected during the following fall. Across grades two to four, the correlation between spring R-CBM scores and fall MAP scores ranged from 0.679-0.720, all significant at the $p<.05$ level.

| An Analysis of the Concurrent and Predictive Validity of Curriculum Based Measures (CBM), the Measures of Academic Progress (MAP), and the New England Common Assessment Program (NECAP) for Reading | |
|---|---|
| Study Citation | Andren, K. J. (2010). *An analysis of the concurrent and predictive validity of curriculum based measures (CBM), the Measures of Academic Progress (MAP), and the New England Common Assessment Program (NECAP) for reading* (Doctoral dissertation, The University of Southern Maine). |

| | |
|---|---|
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | 137 students in grade 3 |
| Description of Samples | Participants were sampled from two schools in a suburban public school district in the northeastern US. The schools were described as having a student population that was evenly split between boys and girls, and was:<br>● 96% white<br>● 19% low-income<br>● 11% special education. |
| Outcomes Measured | R-CBM, Reading Maze, Measures of Academic Progress Reading Test, and the New England Common Assessments Program Reading Test. |

Andren (2010) studied the correlation of R–CBM scores from administrations conducted in the fall and winter of grade three with scores from group reading tests given in the fall, winter, and spring of the same grade. The interval between administration of the aimsweb measures and both external criterion measures ranged from several weeks to several months. Estimated correlations (all significant at the $p<.05$ level) were as follows:

- Correlations between R-CBM and MAP measures ranged from 0.762 to 0.807, depending on the interval.
- Correlations between R-CBM and NECAP measures ranged from 0.678 to 0.714, depending on the interval.
- Correlations between Reading Maze and MAP measures ranged from 0.652 to 0.708, depending on the interval.
- Correlations between Reading Maze and NECAP measures ranged from 0.621 to 0.713, depending on the interval.

| **The Utility of CBM Written Language Indices: An Investigation of Production-dependent, Production-independent, and Accurate-production Scores** | |
|---|---|
| Study Citation | Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, *34*(1), 27-44. |
| Research Study | NA |

| | |
|---|---|
| Contributors | |
| Type of Study | Correlational |
| Sample Size | Participants included 203 students in grades 2, 4, and 6 |
| Description of Samples | Students were recruited from three schools in one rural northern Illinois school district: an early elementary, an elementary, and a middle school. The sample was 56% female, with around 6% enrolled in special education. No other demographic data on the sample was available, although the district was 94% white, 11% low-income, and 3% with limited English proficiency. |
| Outcomes Measured | aimsweb Written Expression subtests for Total Words Written, Words Spelled Correctly and Correct Writing Sequences, as well as the Stanford Achievement Test 9 Language subtest. |

Jewell and Malecki (2005) administered a three-minute narrative-writing probe to students in grades two (N=87), four (N = 59), and six (N = 57) in a rural Illinois school district. The students also took the Stanford Achievement Test, Ninth Edition (SAT9; Harcourt Brace Educational Measurement, 1996) in the same month as the writing probe. Estimated correlations were as follows:

- The correlations between Total Words Written and the SAT-9 Language score ranged from -.14 to 0.24, none of which were significant.
- The correlations between Words Spelled Correctly and the SAT-9 Language score were: R=.05 at grade six, R=.29 at grade four, and R=0.38 at grade two, with only the latter correlation significant at the p<.01 level.
- The correlations between Correct Writing Sequences and the SAT-9 Language score were R=0.23 for grade six, R=0.46 for grade four, and R=0.57, with the latter two correlations significant at the p<.01 level.

| **Moving Beyond Total Words Written: The Reliability, Criterion Validity, and Time Cost of Alternate Measures for Curriculum-based Measurement in Writing.** | |
|---|---|
| Study Citation | Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review*, *31*(4), 477-497. |

| | |
|---|---|
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | 179 students in grades 3 and 4 |
| Description of Samples | Students were drawn from a suburban elementary school in a Southeastern state. The sample was roughly evenly split between males and females, with 82% of the sample white. |
| Outcomes Measured | aimsweb Written Expression subscores for Total Words Written, Words Spelled Correctly, and Correct Writing Sequences, as well as scores from the Iowa Tests of Basic Skills (ITBS) Language test, the Louisiana Educational Assessment Program (LEAP), and teacher ratings of student writing skills. |

Gansle, Noell, VanDerHeyden, Naquin, and Slider (2002) administered two three-minute story-writing probes in February to all 179 students in grades three and four of a suburban school in the Southeast. Probe scores were then correlated with either the Iowa Tests of Basic Skills Language Score (grade three) or the Louisiana Educational Assessment Program Write Competently and Conventions of Language subtests (grade four), both administered in March. Estimated correlations were as follows:

- Correlations between aimsweb Written Expression subtests and the ITBS at grade three were R=0.15 for Total Words Written (ns), R=0.24 (p<.05) for Words Spelled Correctly, and R=0.43 (p<.001) for Correct Word Sequence.
- Correlations between aimsweb Written Expression subtests and LEAP Write Competently subscores at grade four were R=0.28 (p<.01) for Total Words Written, R=0.29 (p<.01) for Words Spelled Correctly, and R=0.28 (p<.01) for Correct Word Sequence.
- Correlations between aimsweb Written Expression subtests and LEAP Use Conventions of Language subscores at grade four were R=0.16 (ns) for Total Words Written, R=0.26 (p<.05) for Words Spelled Correctly, and R=0.41 (p<.001) for Correct Word Sequence

| **A Preliminary Investigation Into the Identification and Development of Early Mathematics Curriculum-Based Measurement** | |
|---|---|
| Study Citation | Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, *33*(2), 234-248. |

| | |
|---|---|
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | 52 students in grade 1 |
| Description of Samples | Students were drawn from two schools in a medium-sized school district in the Pacific Northwest. The sample was 56% female, with approximately 58% of the sample low-income, 10% non-white, and 6% in special education. |
| Outcomes Measured | aimsweb TEN subtests Oral Counting, Number Identification, Quantity Discrimination, and Missing Number as well as the Woodcock-Johnson Applied Problems Subtest and the Number Knowledge Test. |

Clarke and Shinn (2004) administered TEN to 52 students in grade one at two schools in the Pacific Northwest. Two forms of each TEN measure were given in the fall and winter, and one form was administered in the spring. In addition, students took the Applied Problems subtest of the Woodcock-Johnson Psycho-Educational Battery, Revised (Woodcock & Johnson, 1990) in the fall and spring, and the Number Knowledge Test (Case & Okamoto, 1996) in the fall. Correlations were as follows:

- For Oral Counting, correlations with criterion measures ranged from 0.60 to 0.72.
- For Number Identification, correlations with criterion measures ranged from 0.63 to 0.72 and alternate forms reliability ranged from 0.89 to 0.93 across intervals.
- For Quantity Discrimination, correlations with criterion measures ranged from 0.71 to 0.80 and alternate forms reliability ranged from 0.92 to0.93 across intervals.
- For Missing Number, correlations with criterion measures ranged from 0.68 to 0.74 and alternate forms reliability ranged from 0.78 to 0.83 across intervals.

It should be noted that sample sizes for some correlations were extremely small.

| **Technical Adequacy of Early Numeracy Curriculum-Based Measurement in Kindergarten** | |
|---|---|
| Study Citation | 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. |
| Research Study Contributors | NA |

| | |
|---|---|
| Type of Study | Correlational |
| Sample Size | 59 students in kindergarten |
| Description of Samples | Students were drawn from a large K-2 primary school in the rural Midwest, where 30% of the student population were characterized as low income. The sample was 54% male and 94% white. |
| Outcomes Measured | aimsweb TEN subtests Oral Counting, Number Identification, Quantity Discrimination, and Missing Number, as well as the Stanford Achievement Test, 10th edition. |

Martinez, Missall, Graney, Aricak, and Clarke (2009) administered three of the TEN measures (Oral Counting, Quantity Discrimination, and Missing Number) and the Stanford Achievement Test, Tenth Edition (SAT-10; Harcourt Educational Measurement, 2002) in the spring to 59 kindergarten students at a rural Midwestern school. In total 52 of the students had also taken the TEN measures (including Oral Counting) the previous fall. Correlations were as follows:

- For Oral Counting, the correlation with the SAT-10 was 0.45.
- For Number Identification, correlations with the SAT-10 ranged from 0.31 to 0.44 and alternate forms reliability was 0.91.
- For Quantity Discrimination, correlations with the SAT-10 ranged from 0.46 to 0.63 and alternate forms reliability was 0.77.
- For Missing Number, correlations with the SAT-10 ranged from 0.36 to 0.47 and alternate forms reliability was 0.79.
- All correlations were significant at the $p<.05$ level.

| **Extending the Research on the Tests of Early Numeracy: Longitudinal Analyses Over Two School Years** | |
|---|---|
| Study Citation | 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), 89-102. |
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | 61 students in kindergarten |
| Description of Samples | Students were drawn from a suburban public school district near New York City. The sample was 56% male, 56% non-white, 3% English language learners, 13% in |

| | |
|---|---|
| | self-contained special education classrooms, and 10% receiving other special services (such as speech therapy). |
| Outcomes Measured | aimsweb TEN subtests Oral Counting, Number Identification, Quantity Discrimination, and Missing Number as well as teacher ratings of students' mathematics skills on the Academic Competence Evaluation Scales - Mathematics. |

In a longitudinal study, Baglici, Codding, and Tryon (2010) administered the TEN measures in the winter and spring to 61 kindergarten and grade one students in a suburban school near New York City. The authors also administered the aimsweb Mathematics-Curriculum Based Assessment in the winter and spring of grade one , and collected teacher ratings of students' mathematics abilities at the end of grade one using the Academic Competence Evaluation Scales—Mathematics (ACES; DiPerna & Elliott, 2000). Correlations were as follows:

- For Oral Counting, correlations with ACES ratings were R=0.36 (Spring) andR=0.39 (Winter), both nonsignificant.
- For Number Identification, correlations with ACES ratings were R=0.46 (winter) and R=0.57 (spring) (both significant at the p<.01 level) and alternate form reliability was R=0.71 (kindergarten, spring) and R=0.84 (kindergarten, winter).
- For Quantity Discrimination, correlations with ACES ratings were R=0.40 (spring) and 0.51 (winter) (with the latter significant at the p<.01 level) and alternate form reliability was R=0.89 (kindergarten, spring) and R=0.91 (kindergarten winter).
- For Missing Number, correlations with ACES ratings were R=0.44 (winter) and R=0.58 (spring) (both significant at the p<.01 level) and alternate form reliability was R=0.81 (kindergarten, spring) and R=0.86 (kindergarten, winter).

| **Linking the R-CBM and MAZE with The Lexile Framework® for Reading: A Study to Link the Reading Curriculum-Based Measurement and Reading Maze with the Lexile Framework** | |
|---|---|
| Study Citation | MetaMetrics. (2011). *Linking the R-CBM and MAZE with the Lexile Framework for Reading: A study to link the reading curriculum-based measurement and reading maze with the Lexile Framework*. Durham, NC: MetaMetrics. |
| Research Study Contributors | Pearson |
| Type of Study | Linking study |
| Sample Size | 5,444 students in grades 1-8 for R-CBM and 5,316 students in grades 1-8 for MAZE. |

| | |
|---|---|
| Description of Samples | Participants were sampled from 25 schools in 12 districts and 8 states. Both samples were roughly evenly balanced between males and females (for students reporting gender), and were approximately 56% non-white. |
| Outcomes Measured | aimsweb R-CBM and MAZE scores, as well as scores on a Lexile Linking Test. |

In this study, the R–CBM measure was aligned with the Lexile scale through a linking study (MetaMetrics, 2011). A total of 5,444 students from school districts in several states took R–CBM and a specially-constructed Lexile calibration test. The test, a group-administered, 40-minute, 35-item assessment of reading comprehension, asked students to read a brief passage and then select (from four options) the word that best completed a sentence related to the content of the passage. The internal-consistency (alpha) reliability of the Lexile test at each grade ranged from .90 to .92. Scores from the assessments were linked through an equipercentile linking method, with a cubic spline post-smoothing technique applied. Correlations between the R-CBM and the linking test ranged from 0.59 to 0.73, and Lexile-scaled R-CBM scores generally increased monotonically as the grade levels progressed. Similarly, correlations between the MAZE and the linking test ranged from 0.47 to 0.63 and Lexile-scaled MAZE scores increased monotonically as the grade levels progressed, although for both R-CBM and Reading Maze tests, corresponding Lexile linking test scores decreased slightly for grade seven students in comparison to grade six students.

| **Curriculum-Based Measurement of Oral Reading: Standard Errors Associated With Progress Monitoring Outcomes From DIBELS, AIMSweb, and an Experimental Passage Set** | |
|---|---|
| Study Citation | Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review*, *38*(2), 266-283. |
| Research Study Contributors | NA |
| Type of Study | Reading passage field test study. |
| Sample Size | 28 students in grade 2 and 40 students in grade 3. |
| Description of Samples | Samples were drawn from two elementary schools in a southeastern region, characterized as having between 19-32% low-income students. Within the sample, 73% of students was |

| | |
|---|---|
| | white. |
| Outcomes Measured | R-CBM indices for three passage sets. |

In this study, researchers compared the magnitude of estimated standard errors for the slopes of individual growth estimates across three passage sets, with 20 passages each (FAIR-R, DIBELS, and aimsweb). Study participants included 28 second-grade and 40 third-grade students enrolled in one of two elementary schools in a southeastern region of the US. Students were tested twice weekly on each passage set over approximately 10 weeks. Individual student growth estimates on the aimsweb R-CBM were found to be more reliable than such estimates for the DIBELS, but less reliable than such estimates for the FAIP-R passage set. Specifically, the average standard errors of the growth slopes produced using the FAIP-R passage set were smaller than those produced using the aimsweb passage set (0.64 compared to 0.71), and the average standard errors of the slopes from the aimsweb passage set were smaller than those produced using the DIBELS passage set (0.71 compared to 0.91).

It should be noted that error rates for both DIBELS and aimsweb passage sets were large enough that the authors recommended not using them as the sole factor in identifying students for special education. In addition, sample sizes were quite small, so results may not generalize. Some of the study participants (and their peers at the same school) had previously participated in a field test of the FAIP-R passages. Moreover, students were provided an incentive each time they exceeded their own personal best reading performance; thus, it is unknown whether similar rates of improvement would be observed when students are not provided such incentives.

| **aimsweb Reliability of the Rate of Improvement Study** | |
|---|---|
| Study Citation | NCS Pearson, Inc. (2012). *aimsweb Technical Manual*. Bloomington, MN: NCS Pearson, Inc. |
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | M-CAP: Participants included roughly between 3,000 to 8,000 students in each of grades 2-8 whose progress had been monitored at least 10 times during the school year.<br><br>M-COMP: Participants included between 2,500 to 6,750 students in each of grades 2-8 whose progress had been monitored at least 10 times during the school year. |
| Description of Samples | All participants were drawn from the aimsweb database during the 2010-2011 school year. Demographic data on the samples were not reported, so it is unknown how representative the |

| | |
|---|---|
| | samples were of the target population in terms of gender, racial/ethnic background, and other characteristics. |
| Outcomes Measured | Student performance on the M-CAP and M-COMP progress monitoring measures were used to compute a weekly rate of improvement. |

In this study, students whose progress had been monitored at least ten times during the academic school-year were selected for the sample. Researchers calculated a weekly Rate of Improvement, expressed in raw M-CAP and M-COMP scores, for each student over the course of the school year. A split-half correlation coefficient, adjusted for test length using the Spearman-Brown Formula, was computed between ROI estimates calculated for odd- versus even-numbered administrations. This correlation represents the reliability of ROI estimates. The correlations were as follows:

- For M-CAP, correlations ranged from 0.75 to 0.79.
- For M-COMP, correlations ranged from 0.74 to 0.82.

| **aimsweb Alternate Form Reliability Study** | |
|---|---|
| Study Citation | NCS Pearson, Inc. (2012). *aimsweb Technical Manual*. Bloomington, MN: NCS Pearson, Inc. |
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | R-CBM: Participants included 1,000 students in each of grades 1-8.<br><br>Reading Maze: Participants included between 6,000 and 25,000 students in each of grades 2-8.<br><br>M-CAP: Participants included between 800 and 1,000 students in each of grades 2-8.<br><br>M-COMP: Participants included approximately 1,000 students in each of grades 2-8.<br><br>TEL: Grade K participants included between 665 and 1,463 students, depending on the testing period and measure; grade 1 participants included between 925 to 942 students, depending on the testing period and measure. |

| | |
|---|---|
| Description of Samples | R-CBM: All participants were randomly selected from the aimsweb database during the 2009-2010 school year. Based on demographic data for participating schools, the sample had similar demographic characteristics to the US population in terms of gender, SES, race, and region, although the sample had somewhat fewer African-American and Hispanic students than the general population and Midwestern schools were over-represented.<br><br>Reading Maze: Participants were drawn from the aimsweb database during the 2009-2010 school year. Although the samples are very large, sample demographics are not reported so it is not clear how well results may generalize to specific populations.<br><br>M-CAP: Participants came from a national field test sample that was similar to the US population with respect to gender, race/ethnicity, and geographic region.<br><br>M-COMP: Participants came from a national field test sample that was similar to the US population with respect to gender, race/ethnicity, and geographic region.<br><br>TEL: Participants were drawn from the aimsweb database during the 2007-2008 school year. Although the samples are very large, sample demographics are not reported so it is not clear how well results may generalize to specific populations. |
| Outcomes Measured | Scores from multiple passages/forms of the aimsweb R-CBM, Reading Maze, M-CAP, M-COMP, and TEL subtests. |

Researchers computed alternate forms reliability for several aimsweb subtests by administering multiple passages/forms of the measures and computing correlations between scores.Correlations were as follows:

- For R-CBM, mean alternate forms reliability based on a single probe for the benchmarking passage set ranged from 0.93 to 0.95 across grade levels and benchmarking periods, and mean alternate forms reliability for the progress monitoring passage set ranged from 0.80 to 0.90.
- For Reading Maze, alternate forms reliability ranged from 0.68-0.78 across grade levels; because of the long time period between administrations (four months), these are conservative estimates of Maze alternate-forms reliability, as they capture both measurement error and true growth.
- For M-CAP, alternate-forms reliability ranged from 0.80 to 0.88 across grade levels.
- For M-COMP, alternate forms reliability ranged from 0.82 to 0.90 across grade levels.
- For TEL, alternate forms reliability ranged from 0.59 to 0.82 for grade K and from 0.61 to 0.78 for grade one; because of the long time period between administrations (four

months), these are conservative estimates of TEL alternate-forms reliability, as they capture both measurement error and true growth.

| aimsweb Interrater Reliability Study | |
|---|---|
| Study Citation | NCS Pearson, Inc. (2012). *aimsweb Technical Manual*. Bloomington, MN: NCS Pearson, Inc. |
| Research Study Contributors | NA |
| Type of Study | Correlational |
| Sample Size | R-CBM: Participants included between 61 and 71 students from each of grades 2, 4, 6, and 8.<br><br>M-CAP: Participants included roughly 60 students in each of grades 2-8 |
| Description of Samples | R-CBM: Participants were sampled from five public schools in Minnesota and Texas. No demographic data on participants were available, so it is not clear how well results may generalize.<br><br>M-CAP: Participants were randomly sampled from a national field test sample that was similar to the US population with respect to gender, race/ethnicity, and geographic region. |
| Outcomes Measured | Scores from the aimsweb R-CBM and M-CAP subtests. |

The R-CBM study sample was obtained in March and April 2011 at Grades two, four, six, and eight from five public schools in Minnesota and Texas. Each administration was audio recorded. At each grade, half of the recordings were scored by two scorers working independently, while the other half were scored by a different pair of scorers working independently.

For the M-CAP, 60 cases were selected at random from the field test sample and independently scored by two individuals who were provided with the training materials from the M–CAP Administration and Scoring Guide.

For both subtests, interrater reliability coefficients were calculated using Shrout and Fleiss (1979) Formula 2, which takes into account differences in the level of scores assigned by different raters as well as differences in how they rank-order students.

- For the R-CBM, interrater reliability was uniformly 0.99 across all grade levels.

- For the M-CAP, interrater reliability was 0.99 at all grade levels except grade eight, where it was 0.97.

# Future Research Plans

## Overview of Future Research Plans

aimswebPlus research and development was designed intentionally to obtain evidence on item quality, test score reliability, validity and sensitivity to academic growth, and to ensure the assessments conform to professional standards and meet customer expectations. But, there's more to be done. Immediate plans for 2016 studies include inter-rater reliability studies to show evidence of scoring objectivity. The team is currently prioritizing other 2016 and longer-term research efforts which will likely include evidence:

- that scores from frequent monitoring of students are reliable, and,
- that decision rules built into the progress monitoring system are accurate and valid.

## Future Research Plans

| aimswebPlus Correlation with PARCC and Smarter Balanced | |
|---|---|
| Intended Start Date | May 2016 |
| Anticipated Length of Study | TBD |
| Type of Study | Correlational |
| Research Leads | John Bielinski, PhD & Cheryl Johnson |
| Intended sample size | TBD |
| Description of Sample | TBD |
| Outcomes to be Measured | Validity - relationships with external measures |

| aimswebPlus Correlation with SuccessMaker | |
|---|---|
| Intended Start Date | Underway |
| Anticipated Length of Study | 9 months |
| Type of Study | Correlational |
| Research Leads | John Bielinski, PhD & Cheryl Johnson |
| Intended Sample Size | TBD |

| | |
|---|---|
| Description of Sample | TBD |
| Outcomes to be Measured | Validity - relationships with external measures |

# References


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.

Andren, K. J. (2010). *An analysis of the concurrent and predictive validity of curriculum based measures (CBM), the Measures of Academic Progress (MAP), and the New England Common Assessment Program (NECAP) for reading* (Doctoral dissertation, The University of Southern Maine).

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review*, *38*(2), 2662-83.

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), 89-102.

Case, R. & Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development, 61*(1-2, Serial No. 246).

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.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, *33*(2), 234-248.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*(1), 36-45.

DiPerna, J. C., & Elliott, S. N. (2000). *Academic competence evaluation scales*. San Antonio, TX: The Psychological Corporation.

Elliot, J. , Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the dynamic indicators of basic early literacy skills-modified. *School Psychology Review, 30*(1), 33-49.

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34*(3), 140-153.

Fuchs, L. S., Fuchs, D., & Deno, S. L. (1982). Reliability and validity of curriculum-based informal reading inventories. *Reading Research Quarterly, 18*(1), 6-26.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

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.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review*, *31*(4), 477-497.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test* (9th ed.). San Antonio, TX: Harcourt Assessment.

Harcourt Brace Educational Measurement. (2002). *Stanford Achievement Test* (10th ed.). San Antonio, TX: Harcourt Assessment.

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, *34*(1), 27-44.

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.

McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review*, *37*(4), 550-566.

Merino, K., & Beckman, T. O. (2010). Using reading curriculum-based measurements as predictors for the measures of academic progress map standardized test in Nebraska. *International Journal of Psychology: A Biopsychosocial Approach*, *6*, 85-98.

MetaMetrics. (2011). *Linking the R-CBM and MAZE with the Lexile Framework for Reading: A study to link the reading curriculum-based measurement and reading maze with the Lexile Framework*. Durham, NC: MetaMetrics.

NCS Pearson, Inc. (2012). *aimsweb Technical Manual*. Bloomington, MN: NCS Pearson, Inc.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *24*(1), 19-35.

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.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Stenner, A. J. (1997). *The Lexile framework: A map to higher levels of achievement*. Durham, NC: MetaMetrics.

Woodcock, R.W., & Johnson, M. B. (1990). *Woodcock-Johnson psycho-educational battery-revised*. Allen, TX: DLM Teaching Resources.