mCLASS Dyslexia Screening Measures Technical Manual

Amplify.

Table of Contents

Introduction

To better meet the growing needs of educators to screen students for reading difficulties, including dyslexia, Amplify has created three additional early literacy measures. These measures are designed to provide additional information on students' skills in the areas of vocabulary, spelling (encoding), and rapid automatized naming (RAN). These additional screeners will provide indication of risk in these domains, which are important for reading success, and may provide additional information related to risk for reading difficulties, including dyslexia. The measures were created to complement the mCLASS:Reading3D assessment when screening students for reading difficulties. Educators may use the additional information collected from these measures to develop targeted interventions for students at risk.

New Measures Overview

Vocabulary

Background

The Vocabulary measure is an indicator of a student's general vocabulary knowledge. It assesses a student's depth of knowledge of a set of grade level high frequency and high utility words that are used across domains (i.e., Tier 2 words; Beck, McKeown, & Kucan, 2002) and content specific words and the ability to make meaning of words encountered in text. It incorporates a variety of tasks that vary across grade levels and allow for assessing multiple dimensions of vocabulary knowledge across multiple contexts. In the development of the measure, multiple task types were piloted with students. The tasks included in the final measure for each grade level are those that produced the most reliable results for that grade. The number of items increases in each grade level as students' ability to sustain attention increases with age.

The first task administered to all students is the two-question vocabulary task adapted from the work of Kearns & Biemiller (2010). In this task, students are asked two yes or no questions about a target word. This requires deeper knowledge of a word than many traditional tasks as the same word is included in two contexts or questions. Evidence of the reliability, validity and sensitivity to differences in students vocabulary skills has been documented (see Kearns & Biemiller, 2010). Scores on the two-question task were strongly correlated with an external measure of vocabulary, the Peabody Picture Vocabulary Test (PPVT) with coefficients ranging from .76 to .90.

In grades 1 through 3, students also complete traditional fill in the blank questions. They read a sentence with a missing word and select the word that best completes the sentence from a set of four words. Distractor or incorrect responses for each item include words that look or sound similar to the target word, words that mean almost the same thing as the target word but are not correct in the context of the sentence, or words that are related to the target word or sentence context but are not correct in the context of the full sentence.

Finally, grades 2 and 3, students complete items that require matching a word to its basic definition. The words included in this question type are words that are included in earlier portions of the assessment. The purpose is to determine whether a student has basic knowledge of the definitions of these same words.

Word Selection

Both high utility (i.e., Tier 2) and content specific words (starting in grade 1) are included within the assessment. Tier 2 words were selected from the vocabulary lists included within many published core reading programs (e.g., Core Knowledge Language Arts) as well as published lists of frequently occurring and high utility words that students should know to comprehend text (see Graves, 2016). The approximate age of acquisition of each word and meaning and prioritization for instruction of each word and meaning according to Biemiller's Words Worth Teaching (2009) was also specified. All words in grades K and 1 were rated as either known by most children at the end of grade 2 (E) or high-priority words known by 40 to 80 percent of children by the end of grade 2 (T2). All words in grades 2 and 3 were rated T2 on the list.

Content area words were randomly selected from widely available lists of Language Arts, Math, Science, and Social Studies vocabulary words (e.g., Marzano & Simms, 2013). Table 1 includes the item types and word types included on the Vocabulary measure for each grade.

Table 1. Vocabulary Item and Word Types for each Grade

Grade	Item Types	Word Type
K	20 (10 pairs) two-question vocabulary	Tier 2
1	20 (10 pairs) two-question vocabulary	Tier 2
	10 fill in the blank	Tier 2, Content specific
2	14 (7 pairs) two-question vocabular	Tier 2
	8 fill in the blank	Tier 2, Content specific
	15 basic definition	Words included in other tasks
3	14 (7 pairs) two-question vocabulary	Tier 2
	8 fill in the blank	Tier 2, Content specific
	15 basic definition	Words included in other tasks

Administration and Scoring

The Vocabulary measure is administered on a computer or tablet and is completed independently by the student. Each item and all answer options are spoken (by the computer) to the students who then select their answer choice. The Vocabulary measure is administered from the beginning of kindergarten through the end of grade 3. Each question presented is worth one point. A student's final score is the total number of questions answered correctly.

Spelling

Background

The Spelling measure is an indicator of a student's level of general spelling skills. It is designed based on the principles of General Outcome Measurement and Curriculum-Based Measurement (CBM; Deno, 1992). Assessments from this approach are designed to efficiently screen for students who are at-risk for difficulty; they are brief assessments of critical skills that are sensitive to student learning and growth overall. CBM measures do not assess all skills within a domain but provide a snapshot of a student's skills in a given area using tasks that are instructionally useful and can be reliably administered (Deno, 2003).

CBM Spelling has been researched since the 1980s, and there is a body of evidence for the reliability (range: 0.83 - 0.99) and validity (range: 0.81 - 0.98) of results produced (e.g., Deno, Mirkin, Lowry, & Kuehnle, 1980; Marston, 1992; Shinn, 1981; Tindal, Germann, et al., 1988, as cited in Hosp, Hosp, & Howell, 2007) as well as its sensitivity to growth in students' skills (Fuchs, Fuchs, Hamlett, Walz, & Germann, 2003).

In CBM Spelling measures, a list of grade appropriate words is dictated to the students at a predetermined pace. The lists contain 12 words in the early grades (K to 2) and 17 words in grade 3 and beyond. At each grade level, a form includes words that address skills taught over the entire grade level; all forms within a grade level are designed to be of equivalent difficulty so a student's growth over the school year can be tracked. The student must write each word that is dictated. CBM Spelling measures produce two scores: Correct Letter Sequences (CLS) and Words Spelled Correctly (WSC). CLS is a more sensitive measure; a correct letter sequence is counted for each correct letter to letter sequence, including the beginning space to the first letter, a letter to punctuation (e.g., an apostrophe in a contraction), punctuation to a letter, and the end letter to a space. WSC is a count of the number of words a student spells correctly (Hosp, Hosp, & Howell, 2007). As an example, if the word is bake and the student writes bak, he would receive 3 points for CLS and 0 points for WSC. See Figure 1 for an additional sample of the CLS rules for the word chin.

Figure 1. Example Scoring for the Target Word: chin

chin	"chin"	"shin"	"cin"
1. _c			
2. ch			
3. hi			
4. in			
5. n_			
Total CLS	5	3	3

Word Selection

Items on the Spelling measure represent a random sample of grade-specific words drawn from a pool of words covering the phoneme-grapheme correspondences that students at each grade level are expected to learn as well as common high frequency words drawn from the Dolch lists. Common scopes and sequences of phonics instruction (e.g., that of Core Knowledge Language Arts) as well as the expectations included in the Common Core State Standards informed the word types included at each grade level. These are listed in table 2. Each form for a grade level includes the same number of words of each type.

Table 2. Word types included in the Spelling Measure by Grade

Grade	Word Types
K	- 2 VC (vowel consonant)
	- 6 CVC
	- 4 high frequency words (pre-primer and primer Dolch)
1	- 1 CVC
	- 1 CVCe (silent e making the vowel long)
	- 3 one-syllable with one 2 consonant cluster
	- 3 one-syllable with consonant digraph
	- 2 one-syllable with r- or l- controlled vowel
	- 2 high frequency (grade 1 Dolch)

2	
	- 2 one-syllable with vowel digraph, dipthong, or variant vowel
	- 2 two-syllable with long vowel
	- 1 two-syllable compound
	- 1 two-syllable with r-controlled vowel
	- 1 two-syllable with consonant digraph
	- 1 -ing ending with a change in spelling of the root word
	- 1 two-syllable with prefix or suffix
	- 1 possessive or contraction
	- 2 high frequency (grade 2 Dolch)
3	
	- 1 one-syllable with -ed ending
	- 2 two-syllable with open syllable or y as a vowel
	- 2 two-syllable with digraph
	- 2 two-syllable with dipthong
	- 2 multi-syllable compound
	- 2 three-syllable
	- 2 multi-syllable with prefix
	- 2 multi-syllable with suffix
	- 2 high frequency (grade 3 Dolch)

Administration and Scoring

The Spelling measure incorporates the key features of CBM Spelling measure design and administration. It is administered on a computer or tablet so typical procedures for administering the measure were modified to fit the software environment. The target word is spoken (by the computer) and the student uses letter tiles to spell the word. Because there is variability in how long it takes students to manipulate the online letter tiles, the time between words is much longer than the time typically provided in a traditional paper and pencil CBM Spelling assessment. Both CLS and WSC scores are calculated. The Spelling measure is administered from the middle of kindergarten through Grade 3.

RAN

Background

The Rapid Automatized Naming (RAN) measure indicates how quickly students can name aloud numeric symbols. While there is not strong agreement in the field of exactly which cognitive processes RAN is measuring, a large body of evidence has documented RAN as one of the best predictors of overall reading skill, including word reading, text reading, reading fluency, and comprehension and an area difficulty for students with reading disabilities (Araujo, Reis, Petersson, & Faisca, 2015). Deficits in rapid automatized naming have also been shown to be a robust indicator of risk for dyslexia in children (Gaab, 2017).

Because of the strong predictive relationship that RAN displays with tasks that measure various reading skills, researchers hypothesize that completion of RAN tasks requires the coordination of multiple processes. "The seemingly simple task of naming a series of familiar items as quickly as possible appears to invoke a microcosm of the later developing, more elaborated reading circuit" (Norton & Wolf, 2012, p. 427). The full circuit requires coordination of attention, working memory, visual processing, phonological processing, etc.; these processes are also required for reading.

RAN tasks typically include five stimuli repeated 10 times for a total of 50 items. RAN tasks are administered using a variety of stimuli: letters, numbers, colors and objects. Colors and objects are sometimes used for younger children (ages 6 and below) who may not know the names of letters or numbers while letter and numbers (alpha-numeric items) are used for older children through adulthood (Norton & Wolf, 2012). In a meta-analysis of Rapid Automatized Naming, results showed that RAN tasks that use numbers or letters have the highest correlation with overall reading skill than tasks that use colors or pictures, with the number-naming task displaying the highest correlation across grades (Araujo, Reis, Magnus Petersson, & Faisca, 2014).

Concurrent correlations of RAN tasks with reading outcomes are typically in the moderate to strong range (around 0.3 to 0.6) (e.g., Araujo, Reis, Magnus Petersson, & Faisca, 2014; PRO-ED, 2005).

Test Construction

Amplify piloted three versions of the RAN assessment: colors, objects, and numbers. In order to streamline testing, the goal was to create one measure of RAN that (a) provides results that correspond to those of existing gold standard measures of rapid automatized naming; in this case, the Pro-Ed RAN measure and (b) relates to overall reading performance in ways that are similar to gold standard measures of rapid automatized naming.

To determine which tasks best corresponded to the results of an existing gold standard measure of rapid automatized naming, correlations of each task with the corresponding task on the Pro-Ed RAN assessment were conducted. Across all grades, the correlations of the numbers tasks were most often highest.

To determine the relationships of the RAN tasks with overall reading performance, multiple regression with stepwise selection was used to test which set of RAN measures best predicted students' performance on the DIBELS Next Composite score (Good et al., 2013) and the STAR Early Literacy (grades K and 1) Scaled Score (Renaissance, 2014) and STAR Reading (grades 2 and 3) Scaled Score (Renaissance, 2015). The purpose of stepwise selection was to reduce the set of predictor variables (i.e., RAN tasks) to those that are essential and account for as much of the variance as is explained by the total set. The nature of current research questions along with theoretical background and circumstances of the study determined the selection of predictor variables. The results indicate that the RAN Number task explained most of the variance of the overall measures of reading skills. In addition, previous research has documented the utility of the numbers tasks (Araujo, Reis, Magnus Petersson, & Faisca, 2014). As a result, the number task was selected for the RAN assessment.

Administration and Scoring
The RAN measure is administered on a computer or tablet, with a shared teacher-student interface. Students are presented with 50 stimulus items (five rows of ten) of five randomly alternated numbers. The student must orally identify the stimulus items as quickly as possible. The final score is the time it takes a student to name all of the items. RAN can be administered from the beginning of kindergarten to the end of grade 3.

Research Overview

This technical manual describes evidence for the reliability and validity of the Vocabulary, Spelling, and RAN measures based on a study conducted during the 2017-2018 school year (Beginning of Year, BOY; Middle of Year, MOY; and End of Year, EOY).

Purpose: The purpose of this study was to examine the reliability and validity of the Vocabulary, Spelling, and RAN measures, to inform improvements and changes to each measure based on item analysis, and to establish cut points for each measure by grade and time of year (TOY).

Recruitment: Participants were recruited from existing mCLASS:DIBELS Next customers. The following criteria were used to screen for eligible participants: a) school must demonstrate a range of student reading proficiency levels, b) school must demonstrate a variety of demographic characteristics, c) school must have students in any or all grades kindergarten through 3, and d) schools must contribute a minimum of 15 students per grade.

Participants: The study was conducted during the 2017-2018 academic year, with performance data collected around three benchmark periods (BOY, MOY, and EOY). In total, 584 students in grades kindergarten through 3 were assessed in three schools representing two districts.

The Vocabulary, Spelling, and RAN measures were administered to students by reading coaches and Amplify data collectors in the school computer lab. All external measures were collected by a combination of reading coaches and Amplify data collectors.

Demographic Information: Participants in the field study were educators and students from following geographic regions: Northeast and Midwest (US Census Bureau, n.d.). Demographic information including gender, ethnicity, and Free and Reduced Price Lunch Eligibility (FRL; an indicator of socioeconomic status) is provided in Table 3.

Table 3: Sample Size and Demographics by Grade

	All	Kindergarten	Grade 1	Grade 2	Grade 3
Sample Size (n)					
Districts	2	2	2	2	2
Schools	3	3	3	3	3
Students	584	146 (25%)	165 (28%)	129 (22%)	144 (25%)
Gender (n)					
Female	302 (52%)	76 (52%)	87 (53%)	58 (45%)	81 (56%)
Male	282 (48%)	70 (48%)	78 (47%)	71 (55%)	63 (44%)
Ethnicity (n)					
White	548 (94%)	139 (96%)	153 (93%)	122 (96%)	134 (94%)
Black	11 (2%)	2 (1%)	3 (2%)	2 (1%)	3 (2%)
Asian	3 (0%)	0	1 (0%)	1 (1%)	2 (1%)
Multiracial	4 (1%)	0	2 (1%)	2 (1%)	2 (1%)
Ethnicity Not Specified	18 (3%)	5 (3%)	6 (4%)	2 (1%)	3 (2%)
Other Demographics (n)					
FRL Eligible	45 (30%)	127 (87%)	138 (84%)	112 (87%)	120 (83%
FRL Not Eligible	23%	19 (13%)	27 (16%)	17 (13%)	24 (17%)

Research Procedures

All educators or data collectors administering the assessments attended a half-day, web-based training on standardized administration and scoring procedures for the Vocabulary, Spelling, and RAN measures and all external measures. A reliability check was administered to attendees to confirm their ability to administer and score the assessments according to standardized procedures prior to assessing students. Assessors administered the Vocabulary, Spelling and RAN measures to all students following standard procedures in a small group setting (i.e., three to five students) in kindergarten and Grade 1 and in larger groups (i.e., eight to ten students) in grades 2 and 3. The Vocabulary, Spelling, and RAN measures were administered via paper and pencil at BOY and online at MOY and EOY within 2 weeks of the usual benchmark testing period and within 2 weeks of each other.

Students were administered multiple external measures to establish evidence of concurrent and predictive validity—that is, the degree to which the Vocabulary, Spelling, and RAN measures compare with a measure that has been previously validated. Students were administered mCLASS:DIBELS Next, one of the external measures and the tool all participants already use for universal screening purposes, following standard procedures in a one-on-one setting during their usual benchmark testing period. mCLASS:DIBELS Next was selected as an outcome measure due to its established validity and reliability as a test of early literacy skills, which are related to the Vocabulary, Spelling, and RAN measures. An additional external measure selected for all students in kindergarten and Grade 1 was STAR Early Literacy (Renaissance, 2014), and STAR Reading (Renaissance, 2015) served as the external measure for all students in Grade 2 and Grade 3 at BOY, MOY, and EOY. A subset of students were administered the PRO-ED RAN (PRO-ED, 2005) measure as a measure of concurrent validity for mCLASS:RAN at MOY and EOY. All external measures were administered within two weeks of typical benchmarking windows.

External Measure Descriptions
mCLASS:DIBELS Next: DIBELS Next is a nationally recognized observational assessment of reading foundational skills including letter knowledge, phonological awareness, alphabetic principle, oral reading fluency, and comprehension. Skills assessed vary by grade level and time of year. Teachers score students' oral responses to determine risk level for reading failure, and scores for each individual measure as well as a composite score are reported.

Together, the measures administered at each benchmark period comprise a DIBELS Composite Score. The DIBELS Composite Score is a combination of multiple DIBELS Next scores and provides the best overall estimate of the student's reading proficiency and risk level (Good et al., 2013). The DIBELS Next Technical Manual

reports strong reliability support for the measures and overall for the Composite Score. Specific evidence for the reliability and validity of one submeasure used in the analysis for mCLASS:Spelling and the Composite Score is presented below.

Correct Letter Sounds (CLS): CLS is a score provided by DIBELS Next Nonsense Word Fluency (NWF). NWF assesses a students knowledge of letter-sound correspondence and ability to blend sounds into words. The student reads a page of nonsense words [e.g., sim, lut] and can read the words sound-by-sound [/s/ /i/ /m/] or as a whole word [sim]. The CLS score represents the number of letter sounds produced correctly in one minute. The authors report NWF-CLS alternate form reliability ranges from 0.71 to 0.94, inter-rater reliability ranges from 0.99 to 1.00, and test-retest ranges from 0.76 to 0.90 (Good et al., 2013). Predictive validity, as measured as the correlation with the Group Reading and Diagnostic Evaluation (GRADE) at EOY, ranges from 0.43 to 0.56.

DIBELS Composite Score: The authors report alternate form reliability ranges from 0.66 to 0.97, inter-rater reliability ranges from 0.81 to 0.94, and test-retest reliability ranges from 0.97 to 0.99 (Good et al., 2013). Predictive validity, as measured as the correlation with GRADE EOY, ranges from 0.50 to 0.80.

STAR Reading (SR): SR is a computer-adaptive assessment that tests students in five content domains: Word Knowledge and Skills, Comprehension Strategies and Constructing Meaning, Analyzing Literary Text, Understanding Author's Craft, and Analyzing Argument and Evaluating Text (Renaissance Learning, 2015). Students in grades 2 and 3 are administered 20 vocabulary-in-context (i.e., cloze) items and five authentic text passages with multiple-choice literal or inferential questions. Students respond to these questions using a mouse or keyboard. SR takes approximately 10 to 15 minutes for students to complete. It is intended to provide interim data on students' reading skills so educators can set goals, respond quickly to student needs, monitor progress, and maximize growth. Overall scale scores are reported. The SR technical manual reports reliability coefficient ranges for grades 1 through 5 as follows: split-half reliability of 0.88 to 0.89; test-retest reliability of 0.82 to 0.89; and "generic" reliability (i.e., calculated from the conditional error variance of IRT ability estimates) of 0.89 to 0.91 (Renaissance Learning, 2011).Concurrent validity coefficients for grades 1 through 4 ranged from 0.71 to 0.87 with DIBELS Oral Reading Fluency. Predictive validity was also examined using test scores from a variety of other reading measures (e.g., Colorado Student Assessment Program, Illinois Standards Achievement Test). For grades 1 through 6 the validity coefficients ranged from 0.68 to 0.82 for STAR Reading scores predicting later performance on these measures (Renaissance Learning, 2011).

STAR Early Literacy (SEL): STAR Early Literacy is a computer-adaptive assessment of student reading skills in seven domains: General Readiness, Graphophonemic

Knowledge, Phonemic Awareness, Phonics, Comprehension, Structural Analysis, and Vocabulary (Renaissance Learning, 2014). Items within the comprehension and structural analysis domains are omitted in kindergarten administrations. Each administration consists of 25 items presented in multiple choice format (three answer choices per item). Each item includes a graphic display and is dictated by audio recordings. SEL takes approximately 10 to 15 minutes for students in kindergarten and Grade 1 to complete. Similar to STAR Reading, it is intended as an interim assessment to monitor student progress and instructional needs.

SEL's scaled score has generic reliability ranges from 0.78 to 0.86; split-half reliability ranges from 0.75 to 0.85; and alternate form reliability ranges from 0.63 to 0.78. The concurrent validity ranges from 0.50 to 0.88 (measured by the correlations with teachers' ratings of students' skills, Brigance scale, Developing Skills Checklist, Metropolitan Early Childhood Assessment, Texas Primary Reading Inventory, and Test of Phonological Awareness); the concurrent validity ranges from 0.42 to 0.73 to predict STAR Reading scores Renaissance Learning, 2014).

SEL total scaled scores were used in the present analyses rather than scores from the seven subscales within SEL because students may only see a limited number of items in some domains based on their item responses. Thus, scaled scores are considered the strongest estimate of a student's overall reading skills at a particular time (Renaissance Learning, 2014).

PRO-ED RAN: The PRO-ED Rapid Automatized Naming (RAN) Number measure indicates how quickly students can name numeric symbols aloud. This measure is administered via paper and pencil, whereby a student says aloud the items presented on the stimulus card while the assessor records any errors and the student's total time.

The PRO-ED technical manual reports reliability coefficient ranges for kindergarten through Grade 5 as follows: test-retest reliability of 0.81 to 0.91; and interscorer reliability of 0.98 to 0.99 (PRO-ED, 2005). Criterion-prediction validity coefficients between PRO-ED RAN Number and Comprehensive Test of Phonological Processing (CTOPP) Rapid Digit Naming was .72 and between PRO-ED RAN Letters and CTOPP Rapid Letter Naming was .71. Concurrent validity was also examined using test scores from a variety of other reading measures (e.g., Woodcock Johnson Psycho-Educational Battery-Revised: Letter-Word Identification and Passage Comprehension). The validity coefficients for PRO-ED RAN/RAS and Tests of Word Identification and Reading Comprehension ranged from 0.25 to 0.69 (PRO-ED, 2005).

Descriptive Statistics
Table 4 shows percentages of students at each performance level on

mCLASS:DIBELS Next. Percentages of students at each benchmark category are similar to those in a national sample of students; however, the differences in the research sample represent a more equal distribution of students in each benchmark category to ensure students with all skills are represented.

Table 4: Percentages of students at each benchmark performance level on mCLASS:DIBELS Next by Grade

	Beginning of Year					Middle of Year					End of Year				
	Well Below	Below Mean	At	Above	Overall CS	Well Below	Below Mean	At	Above	Overall CS	Well Below	Below Mean	At	Above	Overall CS
Kindergarten															
Study Sample	40 (31%)	32 (25%)	19 (15%)	37 (29%)	128	43 (34%)	26 (21%)	20 (16%)	37 (29%)	126	28 (21%)	41 (31%)	37 (28%)	25 (20%)	131
National sample	28%	18%	15%	39%	356891	20%	19%	23%	39%	355658	13%	15%	29%	43%	351634
Grade 1															
Study Sample	43 (28%)	25 (16%)	29 (23%)	56 (37%)	158	45 (28%)	24 (15%)	34 (21%)	56 (35%)	159	55 (36%)	32 (21%)	24 (16%)	42 (27%)	153
National sample	28%	15%	16%	41%	356828	26%	11%	18%	45%	352890	23%	12%	20%	44%	347706
Grade 2															
Study Sample	24 (21%)	18 (16%)	31 (27%)	41 (36%)	114	25 (22%)	16 (14%)	33 (29%)	40 (35%)	114	28 (24%)	29 (24%)	26 (21%)	38 (32%)	121
National sample	21%	9%	26%	44%	343804	21%	8%	23%	48%	339568	19%	12%	20%	50%	335438
Grade 3															
Study Sample	45 (30%)	11 (7%)	33 (22%)	61 (41%)	150	39 (30%)	10 (8%)	32 (25%)	49 (38%)	130	46 (33%)	21 (15%)	38 (27%)	35 251%)	140
National sample	24%	9%	23%	44%	285721	23%	12%	23%	44%	282669	21%	11%	24%	44%	280266

Table 5 shows STAR Scaled Scores by grade; the grade-level benchmark for SEL in kindergarten and Grade 1 and SR in grades 2 and 3 are also listed for reference (Renaissance Learning, 2011 & 2015 respectively). Student mean performance increases across benchmark periods within a grade. In most cases, mean student performance was above grade-level benchmarks.

Table 5: Descriptive Statistics for STAR Scale Score by Grade

Grade	Beginning of Year				Middle of Year				End of Year			
	N	Mean	SD	Grade-Level Benchmark	N	Mean	SD	Grade-Level Benchmark	N	Mean	SD	Grade-Level Benchmark
K	136	538.47	97.77	499	103	577.26	89.14	570	130	632.38	99.01	644
1	153	658.23	104.63	612	147	705.55	96.93	681	156	740.28	100.57	743
2	121	221.93	107.90	182	113	256.96	125.92	230	118	288.03	124.89	279
3	133	329.38	142.07	323	133	353.73	145.75	360	141	362.59	152.17	392

Table 6 shows student performance on the Vocabulary. In the kindergarten assessment, students complete one item type (i.e., two-question vocabulary) for a total of 20 questions in each benchmark assessment. In the Grade 1 assessment, students complete two item types (i.e., two-question vocabulary and fill in the blank) for a total of 30 questions in each benchmark assessment. In grades 2 and 3, students complete three item types (i.e., two-question vocabulary, fill in the blank, and basic definition) for a total of 37 questions in each benchmark assessment.

For all grades, mean performance on Vocabulary from BOY to MOY and MOY to EOY remains stable; slight differences in performance between benchmarks are likely due to differences in the words included in each benchmark assessment.

Table 6: Descriptive Statistics for the Vocabulary Measure by Grade

Grade	Beginning of Year			Middle of Year			End of Year		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
K	115	13.91	2.68	115	14.92	2.95	134	15.36	2.47
1	128	19.21	3.74	137	20.86	4.42	151	18.24	3.77
2	106	22.70	6.38	121	25.77	5.65	117	27.24	5.79
3	111	28.58	5.60	119	25.71	5.52	141	27.01	5.85

Table 7 shows student performance on the Spelling measure by grade level. The final score is the number of correct letter sequences (CLS). The first kindergarten administration of the Spelling measure was at MOY. The student mean performance on the Spelling measure increases from BOY to MOY for all grades except grade 3. It is important to note that the format of the measure changed from paper and pencil at BOY to online at MOY. For all grades, mean performance on the Spelling measure increases from MOY to EOY.

Table 7: Descriptive Statistics for the Spelling Measure by Grade

	Beginning of Year			Middle of Year			End of Year		
Grade	N	Mean	Grade	N	Mean	Grade	N	Mean	Grade
K	n/a	n/a	n/a	138	13.21	11.62	140	23.32	11.86
1	158	21.06	17.83	159	25.75	13.48	153	41.17	13.21
2	119	34.76	14.16	119	38.74	13.57	117	61.19	15.14
3	134	58.25	18.83	140	60.18	24.65	137	97.80	25.92

Table 8 shows student performance on mCLASS:RAN. RAN scores represent the total time in seconds a student takes to complete the measure. As such, the lower the mean the more quickly the student completed the measure. In most cases, student mean performance in seconds is consistent (within 10 seconds) between BOY to MOY and MOY to EOY. The only exception is MOY to EOY for Grade K. Grade K RAN scores from BOY and MOY to EOY decrease indicating that at EOY Grade K students completed the measure more quickly.

Table 8: Descriptive Statistics for the RAN Measure by Grade

	Beginning of Year					Middle of Year					End of Year				
Grade	N	Mean	Min	Max	SD	N	Mean	Min	Max	SD	N	Mean	Min	Max	SD
K	139	88.75	32	208	37.78	132	75.14	31.8	214.6	36.38	129	65.78	30.80	176.60	27.06
1	154	48.19	25	173	17.25	156	51.04	28.0	96.8	13.94	152	43.51	27.30	87.70	11.68
2	117	36.80	22	68	9.12	118	39.93	23.4	94.5	10.20	118	36.88	22.20	87.50	9.78
3	142	30.01	17	66	7.49	133	35.53	22.2	71.7	9.32	136	34.99	20.60	106.10	13.84

Reliability

Reliability is generally described as the consistency of a measuring instrument; reliability statistics present information about the precision of an instrument, expressed as a ratio. A test with perfect score precision has a reliability coefficient equal to 1, meaning that 100 percent of the variation among persons' scores is attributable to variation in the trait or skill the test measures, and none of the variation is attributable to error. Perfect reliability is unattainable in educational measurement; a test with a reliability coefficient of 0.90 is more likely. On such a test, 90 percent of the variation among students' scores is attributable to the trait or skill being measured, and 10 percent is attributable to errors of measurement. If the trait or skill were measured a second time, students' scores would fluctuate to some degree; that is, scores on the second test would not be perfectly consistent with the same students' initial scores.

Reliability is an essential characteristic of screening assessments that are used for instructional decision-making; if results are spurious and unreliable, inappropriate decisions might be made. Salvia, Ysseldyke, & Bolt's (2013) standards for reliability were used to evaluate the reliability data for the Vocabulary, Spelling, and RAN measures. According to these standards, a minimum reliability of 0.60 is ideal for making educational decisions about groups of students, a minimum of 0.70 suggests adequate reliability generally, a minimum of 0.80 is ideal for screening decisions, and a minimum of 0.90 is required for important educational decisions concerning an individual student.

This section provides details on two types of reliability evidence for the Vocabulary, Spelling, and RAN measures: internal consistency and test-retest reliability.
- Internal consistency reliability refers to a person's degree of confidence in the precision of scores from a single measurement.

- Test-retest reliability refers to the degree of agreement between two administrations of the same measure to the same students under the same conditions over a period of time.

Internal Consistency

Internal consistency reliability estimates (Cronbach's alpha) are shown in Table 9 for Vocabulary and Table 10 for Spelling. Overall, for the Vocabulary measure, internal consistency reliability ranged from 0.65 to 0.84 in all grades except Kindergarten representing consistency above 0.60, which is characterized as moderate to strong reliability (Salvia, Ysseldyke, & Bolt, 2013). Internal consistency reliability for kindergarten ranged from 0.51 at BOY to 0.58 at EOY. Lower estimates of reliability in grades K and 1 may be reflective of the variable nature of students' vocabulary knowledge in these grades.

Table 9: Internal Consistency of the Vocabulary Measure

Grade	Number of Items per Form	Beginning of Year		Middle of Year		End of Year	
		N	Alpha	N	Alpha	N	Alpha
K	20	115	0.51	131	0.53	134	0.58
1	30	128	0.65	153	0.71	151	0.72
2	37	106	0.82	114	0.81	117	0.84
3	37	111	0.83	119	0.80	141	0.82

For the Spelling measure, internal consistency reliability ranged from 0.76 to 0.93, demonstrating strong internal consistency.

Table 10: Internal Consistency of the Spelling Measure

Grade	Number of Items per Form	Beginning of Year		Middle of Year		End of Year	
		N	Alpha	N	Alpha	N	Alpha
K	12	n/a	n/a	138	0.91	140	0.89
1	12	153	0.91	159	0.88	153	0.89
2	12	123	0.86	124	0.76	117	0.84
3	17	132	0.93	140	0.88	137	0.91

These results demonstrate that the Vocabulary and Spelling measures are reliable assessments for making educational decisions.

Test-Retest

Test-retest reliability is more appropriate for the RAN measure. The RAN task is based on speed with repeated presentation of the same items. RAN was administered for a second time to a subset of students approximately two weeks after the first administration to examine test-retest reliability. A reliability coefficient of 0.88 demonstrates a high degree of test-retest reliability for RAN (see Table 11).

Table 11: Test-Retest Reliability of RAN

	Middle of Year
Grade	Test-Retest
K-3 overall	0.88

Validity

The validity of a test is the degree to which it assesses what it claims to measure. Formally, validity is defined as the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests (American Educational Research Association, Psychological Association, & National Council on Measurement in Education, 1999). In other words, validity represents how confident we are that interpretations of test scores accurately represent what we believe they do (e.g., high scores on a comprehension assessment actually represent high comprehension skill). In this sense, validity is a way to describe a test's accuracy or utility.

Validity is not "proven" but rather evidence is collected to strengthen the assertion that a test accurately measures the desired construct(s). Validity was traditionally considered a property assessments themselves possessed; it was categorized as content, construct, and criterion validity. The current view, however, considers a more unified treatment under which validity evidence is collected to support test score interpretations for their intended or unintended use (Kane, 2001; Messick, 1989) and may be captured under a more general heading of evidence for construct validity. Determining the validity of a test involves the use of data and other information, both internal and external to the test instrument itself.

To facilitate discussion and demonstration, evidence for the construct validity of the Vocabulary, Spelling, and RAN is presented via concurrent and predictive correlations with criterion measures. Criterion-related validity is the extent to which student performance on the assessment procedure being validated can estimate student performance on a criterion measure (Salvia, Ysseldyke, & Bolt, 2013). Criterion-related validity includes concurrent and predictive validity. Evidence for the concurrent or predictive validity of an assessment refers to the degree to which current outcomes are associated with outcomes on an external, conceptually related, instrument administered near-concurrently (concurrent validity evidence) or subsequently (predictive validity evidence).

Concurrent Validity

Evidence of concurrent validity is often presented as a correlation between the assessment and an external criterion measure. The final scores obtained using

the Vocabulary, Spelling, and RAN measures should correlate highly with other accepted procedures and measures that assess overall reading achievement as well as the specific skills assessed by each measure and other related skills. The degree of correlation between two conceptually related, concurrently administered tests suggests the tests measure the same underlying psychological constructs or processes.

Concurrent validity results for the Vocabulary, Spelling, and RAN measures with respect to mCLASS:DIBELS Next and the STAR Scaled Score, measures of overall reading achievement, are presented in Tables 11 through 14. For vocabulary and spelling, correlations with the mCLASS:DIBELS Next Composite Score and the STAR Scaled Score are presented. The spelling measure was also correlated with Nonsense Word Fluency - Correct Letter Sounds (NWF-CLS), a measure of knowledge of letter-sound correspondence, for grades and times of year when NWF is administered as part of typical benchmarking assessment. For RAN, correlations are presented between RAN and the mCLASS:DIBELS Next Composite Score and PRO-ED RAN Number. All students who completed each pair of measures are included in the analyses. Correlations between RAN and PRO-ED RAN Number were not available for BOY as PRO-ED RAN was not administered.

Vocabulary: The relationship between the Vocabulary measure and the DIBELS Next Composite Score was examined at BOY, MOY, and EOY. The strongest correlation was demonstrated between vocabulary and Grade 3 DIBELS Next Composite Scores (0.63) at EOY with correlations in kindergarten, Grade 1, Grade 2 at or above 0.50 (0.58, 0.50, and 0.50 respectively). The strongest correlation between the Vocabulary measure and STAR Scaled Scores was demonstrated in Grade 2 (0.70) at EOY with correlations in kindergarten, Grade 1, and Grade 3 at or above 0.46 (0.46, 0.52, and 0.67 respectively). It is important to note that the Vocabulary measure is an assessment of vocabulary specifically, while the DIBELS Next Composite and STAR Scaled Scores are measures of overall reading skill. Thus, these correlations are moderate as expected.

Table 12: Concurrent Validity of the Vocabulary Measure with mCLASS:DIBELS Next and STAR Scaled Scores

Grade	Beginning of Year		Middle of Year		End of Year	
	mCLASS DIBELS Next Composite Score	STAR Scaled Score	mCLASS DIBELS Next Composite Score	STAR Scaled Score	mCLASS DIBELS Next Composite Score	STAR Scaled Score
K	0.54	0.58	0.43	0.50	0.58	0.46
1	0.44	0.40	0.47	0.53	0.50	0.52
2	0.51	0.59	0.36	0.55	0.50	0.70
3	0.56	0.54	0.56	0.64	0.63	0.67

Spelling: Strong, positive relationships were found at BOY between spelling and NWF-CLS in grades 1 and 2; between the Spelling measure and DIBELS Next Composite Scores in grades 1, 2, and 3; and between the Spelling measure and STAR Scaled Scores in grades 1, 2, and 3. Validity coefficients ranged from 0.60 to 0.73. At MOY, strong correlations (at or above 0.60) were found between spelling and NWF-CLS in Grade K and Grade 1; moderate to strong correlations (at or above .57) were determined between the Spelling measure and DIBELS Next Composite Scores in kindergarten, Grade 1, Grade 2, and Grade 3; and between spelling and STAR Scaled Scores in kindergarten, Grade 1, and Grade 3. At EOY, moderate to strong correlations (at or above .48) were found between the Spelling measure and NWF-CLS in Grade K and Grade 1; spelling and DIBELS Next Composite Scores in kindergarten, Grade 1, Grade 2, and Grade 3; and between the Spelling measure and STAR Scaled Scores in kindergarten, Grade 1, Grade 2, and Grade 3. All correlations are as predicted; Spelling measures skills that are highly related, but not the same as, the skills assessed by DIBELS Next Composite Score, NWF-CLS, and STAR.

Table 13: Concurrent Validity of the Spelling Measure with mCLASS:DIBELS Next and STAR Scaled Scores

Grade	Beginning of Year			Middle of Year			End of Year		
	NWF-CLS	mCLASS DIBELS Next Composite Score	STAR Scaled Score	NWF-CLS	mCLASS DIBELS Next Composite Score	STAR Scaled Score	NWF-CLS	mCLASS DIBELS Next Composite Score	STAR Scaled Score
K	n/a	n/a	n/a	0.66	0.73	0.60	0.49	0.59	0.48
1	0.61	0.68	0.73	0.61	0.68	0.58	0.63	0.69	0.56
2	0.60	0.70	0.60	n/a	0.57	0.48	n/a	0.60	0.54
3	n/a	0.73	0.71	n/a	0.70	0.65	n/a	0.68	0.67

RAN: Concurrent validity of RAN with PRO-ED RAN was examined at MOY and EOY. The relationship between RAN with PRO-ED RAN overall (for all students in kindergarten through Grade 3) was strong at 0.84 MOY and 0.79 at EOY.

Table 14: Concurrent Validity of the RAN Measure with PRO-ED RAN Number

	Middle of Year	End of Year
Grade	PRO-ED RAN Number	PRO-ED RAN Number
K-3 overall	0.84	0.79

Concurrent validity of the RAN measure with DIBELS Next Composite Scores was moderate to strong at BOY, MOY, and EOY, which may be explained by the differences in the two assessments. The DIBELS Next Composite Score is a measure of overall reading whereas mCLASS:RAN measures only one aspect of reading. These findings are similar to those described above for concurrent correlations of PRO-ED RAN with measures of overall reading. The validity coefficients for PRO-ED RAN/RAS and Tests of Word Identification and Reading Comprehension ranged from 0.25 to 0.69 (PRO-ED, 2005).

Table 15: Concurrent Validity of the RAN Measure with mCLASS:DIBELS Next Composite Scores

	Beginning of Year	Middle of Year	End of Year
Grade	mCLASS DIBELS Next Composite Score	mCLASS DIBELS Next Composite Score	mCLASS DIBELS Next Composite Score
K-3 overall	0.61	0.56	0.51

Predictive Validity

Predictive validity provides an estimate of the extent to which student performance on the Vocabulary, Spelling and RAN measures predict scores on criterion measures administered at a later point in time, operationally defined as more than two months after the initial administration of the Vocabulary, Spelling and RAN. Estimated as the linear relationship between student performance on Vocabulary, Spelling, and RAN and the criterion measures, such predictive correlations are attenuated by time because students gain skills in the interim between testing occasions, and also by differences in the content specifications of the assessments.

Predictive validity results for each measure are provided in Tables 15 through 18. The performance of students on the Vocabulary and Spelling measures at the beginning of the year were used to predict performance on mCLASS:DIBELS Next Composite Scores, STAR Scaled Scores, and where applicable NWF-CLS at the end of the year.

The performance of students on RAN at the beginning of the year was used to predict performance on mCLASS:DIBELS Next Composite Scores and PRO-ED RAN Number at the end of the year.

Vocabulary: The predictive validity of the Vocabulary measure with mCLASS:DIBELS Next Composite Scores and STAR Scaled Scores was moderate in kindergarten, Grade 1, Grade 2, and Grade 3. As discussed previously, because the external measures assess overall reading skills and not vocabulary specifically, the magnitude of these correlations is as expected.

Table 16: Predictive Validity of the Vocabulary Measure with mCLASS:DIBELS Next and STAR Scaled Scores

Grade	mCLASS DIBELS Next Composite Score	STAR Scaled Score
K	0.48	0.50
1	0.44	0.55
2	0.51	0.59
3	0.57	0.58

RAN: The The predictive validity of the RAN measure with mCLASS:DIBELS Next in kindergarten, Grade 1, Grade 2, and Grade 3 was moderate. The predictive validity of the RAN measure with Pro-ED RAN Number was strong in kindergarten, Grade 1, Grade 2, and Grade 3. These correlations align with prior research on relationships between RAN measures and measures of overall reading skill (Araujo, Reis, Petersson, & Faisca, 2015).

Table 17: edictive Validity of the RAN Measure with mCLASS:DIBELS Next and PRO-ED RAN Number

Grade	mCLASS DIBELS Next Composite Score	PRO-ED RAN Number
K	0.59	0.60
1	0.51	0.64
2	0.45	0.64
3	0.41	0.70

Cut Points

Cut Points for Vocabulary and Spelling

STAR administered to students in the end of year 2018 was applied as a criterion measure for determining Vocabulary and Spelling cut points for BOY, MOY and EOY 2018. The procedure for determining the cut points is as follows:

1. Students' scale scores on STAR were categorized into three groups based on the STAR cut points: less than 25th percentile (25%), 25th-40th percentile (25%-40%), and greater than 40th percentile (40%). These classifications help educators identify students who may need intervention to accelerate growth and move toward proficiency of early literacy skills.

2. The percentage of students in each STAR group out of the total number of students in each analysis as well as mean and median performance on the measures for each STAR group were calculated.

3. The range of tentative cut points was determined by using the contrasting group analysis (Cizek & Bunch, 2007). Under this method, the mean and median Vocabulary and Spelling performance is calculated for each STAR group. The midpoint of these mean and median values are next calculated to provide tentative thresholds between the STAR groups: <25% and 25–40% which correspond to tentative cut points for the Well Below Benchmark/ Below Benchmark range, and between STAR groups of 25–40% and ≥40 %, used to set the tentative cut points for the Below Benchmark/At or ABove Benchmark range.

4. Youden Index Method and diagnostic test accuracy measures (Sensitivity/ Specificity, Predictive Values and Diagnostic Likelihood Ratios) were also implemented for selecting optimal cutoffs (Lopez-Raton et al., 2014). Area Under the Curve (AUC) Statistic for each measure and time of year for administration (beginning, middle, end) was also calculated.

5. Percentages of students in each STAR group were also considered. Attempts were made to keep the percentages of students in each group consistent from predictor to criterion.

6. We specified the cut points Well Below and Below Benchmark as an overall evaluative judgment of the above considerations.

7. There were other considerations specific to each measure including: Preserving the three skill ratings for each measure by ensuring non-overlapping cut points; Setting cut points below the maximum score of each measure to allow for minor

student errors; Considering theoretical knowledge about reading development when choosing between multiple possible cut points.

8. Cut points for each measure by grade, time of year, and Performance Descriptors are presented in Tables 18 through 20. Cut points determine whether a student's skills are At or Above Benchmark, Below Benchmark, or Well Below Benchmark. The Benchmark goal is the minimum score a student must reach to be on track for that grade and time of year. Scores Below or Well Below benchmark indicate a likely need for additional instructional support.

Cut points for the Vocabulary, Spelling and RAN measures are provided in Tables 18 through 20.

Table 18: Vocabulary Measure Cut Points

Grade	Time of Year	Student Score	Performance Descriptor
K	BOY	0–12 13-14 15+	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–14 15–16 17+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-15 16-17 18+	Well Below Benchmark Below Benchmark At or Above Benchmark
1	BOY	0–18 19–22 23+	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–19 20–23 24+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-20 21-24 25+	Well Below Benchmark Below Benchmark At or Above Benchmark
2	BOY	0–19 20–22 23+	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–22 23–25 26+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-24 25-28 29+	Well Below Benchmark Below Benchmark At or Above Benchmark

3	BOY	0–22 23–25 26+	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–24 25–27 28+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-26 27-30 31+	Well Below Benchmark Below Benchmark At or Above Benchmark

Table 19: Spelling Measure Cut Points

Grade	Time of Year	Student Score	Performance Descriptor
K	BOY	Not Applicable	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–6 7–10 11+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-17 18-26 27+	Well Below Benchmark Below Benchmark At or Above Benchmark
1	BOY	0–17 18–22 23+	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–19 20–24 25+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-34 35-41 42+	Well Below Benchmark Below Benchmark At or Above Benchmark
2	BOY	0–32 33–38 39+	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–36 37–41 42+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-49 50-62 63+	Well Below Benchmark Below Benchmark At or Above Benchmark

3	BOY	0–52 53–57 58+	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	0–54 55–58 59+	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	0-85 86-102 103+	Well Below Benchmark Below Benchmark At or Above Benchmark

Cut Points for RAN

1. RAN BOY data collected through the 17-18 school year and the BOY data collected through 18-19 school year are combined to generate cut off scores.

2. The combined data was first screened for outliers, students who read too fast or too slow were identified and then excluded from the analysis.

3. Student reading time for each grade (Grade K to Grade 3) was identified to follow a normal distribution, then standard scores (i.e., z-scores) were calculated (a) to help us find the probability of a score occurring within normal distribution and (b) enables us to compare two scores that are from different normal distributions.

4. For each grade, students' scale scores on RAN were categorized into three groups based on the rule of thumb of ProED RAN: less than 20th percentile (20%), 20th-40th percentile (20%-40%), and greater than 40th percentile (40%). These classifications help educators identify students who may need intervention to accelerate growth and move toward proficiency of rapid automatic naming skills.

5. The cut-off score(z-score) separating 60% of highest scores (i.e., lowest reading time) from the lower score (i.e., higher reading time) was set to be the tentative cut points for the Below Benchmark/At or Above Benchmark range. And the cut-off score(z-score) separating 20% of lowest scores (i.e., highest reading time) from the higher score (i.e., lower reading time) was used to be the tentative cut points for the Well Below Benchmark/ Below Benchmark range

6. There were other considerations specific to each measure including: Preserving the three skill ratings for each measure by ensuring non-overlapping cut points; Setting cut points below the maximum score of each measure to allow for minor student errors; Considering theoretical knowledge about reading development when choosing between multiple possible cut points.

Table 20: RAN Measure Cut Points

Grade	Time of Year	Student Score	Performance Descriptor
K	BOY	>124 101–124 <101	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	>124 101–124 <101	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	>124 101–124 <101	Well Below Benchmark Below Benchmark At or Above Benchmark
1	BOY	>76 62–76 <62	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	>76 62–76 <62	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	>76 62–76 <62	Well Below Benchmark Below Benchmark At or Above Benchmark

2	BOY	>57 48–57 <48	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	>57 48–57 <48	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	>57 48–57 <48	Well Below Benchmark Below Benchmark At or Above Benchmark
3	BOY	>49 41–49 <41	Well Below Benchmark Below Benchmark At or Above Benchmark
	MOY	>49 41–49 <41	Well Below Benchmark Below Benchmark At or Above Benchmark
	EOY	>49 41–49 <41	Well Below Benchmark Below Benchmark At or Above Benchmark

References

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

Araujo, S., Reis, A., Petersson, K. M., & Faisca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. **Journal of Educational Psychology, 107(3),** 868-883.

Beck, I., McKeown, M. G., & Kucan, L. (2013). **Bring Words to Life,** 2nd edition. New York: Guilford Press.

Biemiller, A. (2009). **Words Worth Teaching.** New York: SRA.

Cizek, G. J., & Bunch, M. B. (2007). **Standard setting: A guide to establishing and evaluating performance standards on tests.** SAGE Publications Ltd.

Deno, S. (1992). The nature and development of Curriculum-Based Measurement. **Preventing School Failure,** 36(2), 5-10.

Deno, S. (2003). Curriculum-Based Measures: Development and perspectives. **Assessment for Effective Intervention,** 28(3&4), 3-12.

Dynamic Measurement Group. (2010). Dynamic Indicators of Basic Early Literacy Skills (7th ed.). Eugene, OR: Author.

Dynamic Measurement Group. (2016). DIBELS Benchmark Goals and Composite Score report. Eugene, OR: Author.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G (1993). Formative evaluation of academic progress: How much growth can we expect? **School Psychology Review,** 22, 27-49.

Gaab, N. (2017, February). **It's a Myth That Young Children Cannot Be Screened for Dyslexia!** Baltimore, MD: International Dyslexia Association.

Good, R.H., Kaminski, R.A., Dewey, E.N., Wallin, J., Powell-Smith, K.A., & Latimer, R.J. (2013). **DIBELS Next Technical Manual.** Eugene, OR: Dynamic Measurement Group.

Graves, M. F., Sales, G. C., & Ruda, M. (2008). **The first 4,000 words.** Minneapolis, MN: Seward.

Graves, M. (2016). **The Vocabulary Book, 2nd edition.** New York: Teachers College Press.

Hosp, M.K., Hosp, J.L., & Howell, K.W. (2007). The ABCs of **CBM: A Practical Guide to Curriculum-Based Measurement.** New York, NY: Guilford.

Kane, M. (2001). Current Concerns In Validity Theory. **Journal of Educational Measurement,** 38(4), 319–342.

Kearns, G., & Biemiller, A. (2010/2011). Two-questions vocabulary assessment: Developing a new method for group testing in kindergarten through second grade. **Journal of Education,** 190, 31-41.

Lopez-Raton, M., Rodriguez-Alvarez, M.X, Cadarso-Suarez, C. and Gude-Sampedro, F. (2014).

Optimal Cutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software 61(8), 1–36. URL http://www.jstatsoft.org/v61/i08/.

Marzano, R.J., & Simms, J.A. (2013). Vocabulary for the Common Core. Centennial, CO: Marzano Research.

Norton, E.S., & Wolf, M. (2012). Rapid Automatized Naming (RAN) and Reading Fluency: Implications for Understanding and Treatment of Reading Disabilities. The Annual Review of Psychology, 63, 427-452.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 13–103). New York: Macmillan.

PRO-ED. (2005). RAN/RAS: Rapid Automatized Naming and Rapid Alternating Stimulus Tests. Austin, TX: Author.

Renaissance Learning. (2011). STAR Reading. Wisconsin Rapids, WI: Author.

Renaissance Learning. (2014). STAR Early Literacy technical manual. Wisconsin Rapids, WI: Author. Available by request to research@renaissance.com

Renaissance Learning. (2015). STAR Reading technical manual. Wisconsin Rapids, WI: Author. Available by request to research@renaissance.com

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2013). Assessment in special and inclusive education. Belmont, CA: Wadsworth/Cengage Learning.

Woodcock, R.W., & Johnson, M.B. (1989). Woodcock-Johnson Psycho-Educational Battery-Revised. Itasca, IL: Riverside.

For more information visit amplify.com

Corporate:
55 Washington Street
Suite 900
Brooklyn, NY 11201-1071
(212) 796-2200

Sales Inquiries:
(866) 212-8688 • amplify.com

Amplify.